Genomic evaluation considering the mosaic genome of the crossbreed pig

Claudia A. Sevillano

Propositions

1. Information from commercial crossbred pigs is crucial for selecting purebreds for crossbred performance.
   (this thesis)

2. Practical implementation of genomic prediction for crossbred performance will not use breed of origin of alleles.
   (this thesis)

3. Humans are predisposed to stereotypes because human brains process information in a Bayesian style.

4. Low plasticity of dietary preferences proves that we are not a migrant species.

5. An academic CV with expertise in other fields of science has a greater value than expertise abroad in one's own field.

6. The lack of solidarity and social citizenship in a country is often rooted in colonialism.

Propositions belonging to the thesis entitled

"Genomic evaluation considering the mosaic genome of the crossbred pig"

Claudia Alejandra Sevillano

Wageningen, 21 December 2018

Genomic evaluation considering the mosaic genome of the crossbred pig

Claudia A. Sevillano

Thesis committee

Promotor
Prof. Dr H. Bovenhuis
Professor of Animal Breeding and Genetics
Wageningen University & Research

Co-promotor
Dr M.P.L. Calus
Senior researcher, Animal Breeding and Genomics Centre
Wageningen University & Research

Dr R. Bergsma
Senior researcher
Topigs Norsvin Research Center, Beuningen

Other members
Prof. Dr F.A. van Eeuwijk, Wageningen University & Research
Dr H. Gilbert, INRA, Toulouse, France
Dr N. Ibañez-Escriche, Universitat Politècnica de València, Spain
Dr M.C.A.M. Bink, Hendrix Genetics, Boxmeer, the Netherlands

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Genomic evaluation considering the mosaic genome of the crossbred pig

Claudia A. Sevillano

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at 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
to be defended in public
on Friday December 21, 2018
at 4 p.m. in the Aula.

Sevillano, C.A.
Genomic evaluation considering the mosaic genome of the crossbred pig,
184 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018)
With references, with summaries in English
ISBN: 978-94-6343-381-5
DOI: 10.18174/464062

Abstract

Sevillano, C.A. (2018). Genomic evaluation considering the mosaic genome of the crossbred pig. PhD thesis, Wageningen University, the Netherlands

In pigs, the breeding goal is to improve performance of crossbred (CB) animals in commercial farms. The best purebred (PB) animals to produce CB animals can be selected based on their genomic estimated breeding value (GEBV) for CB performance. GEBVs are the result of combining estimated effects of single nucleotide polymorphisms (SNPs) with the animal's genotype. Using CB genomic information allows to estimate SNP allele effects accounting for the CB genetic background. The genome of CB animals is a mosaic of genomic regions inherited from the different parental breeds, therefore, this thesis aimed to investigate whether SNP alleles have different effects depending from which parental breed the allele was inherited and study the impact on GEBV of PB animals for CB performance when breed-specific allele effects were taken into consideration. Firstly, I showed that around 94 % of alleles of three-way CB pigs can be assigned a breed of origin. Knowing this, allowed me to implement a model that accounts for breed-specific effects of all SNP alleles. Using results of this model, I showed that estimated effects and explained variance of SNPs strongly associated with CB performance are different depending upon from which parental breed they were inherited, however, the majority of the genomic regions are not or only weakly associated with CB performance. Therefore, I implemented a new model that allows to estimate breed-specific effects only for alleles of SNPs strongly associated with CB performance, and for the rest of the SNPs assumes that allele effects are the same across breeds. Differences of prediction accuracies between models were generally small. When the estimated genetic correlation between the performance of PB and CB animals per breed of origin differed largely across models, it was better to use models that make a distinction of alleles according to their breed of origin and whether or not they were associated to the trait.

Contents

5 Abstract

9 Chapter 1 – General introduction

23 Chapter 2 – Empirical determination of breed of origin of alleles in three-breed cross pigs

43 Chapter 3 – Genomic evaluation for a three-way crossbreeding system considering breed of origin of alleles

71 Chapter 4 – Effects of alleles in crossbred pigs estimated for genomic prediction depend on their breed of origin

99 Chapter 5 – Genomic evaluation for a crossbreeding system implementing breed of origin for targeted markers

119 Chapter 6 – General discussion

139 Supplementary material

147 References

159 Summary

165 Training and Supervision plan

169 Curriculum vitae

177 Acknowledgements

Chapter 1

General introduction

1.1 Introduction

Current pig breeds originated from the Eurasian wild boar (Sus scrofa), which began to be domesticated about 9,000 years ago (Ervynck et al., 2001; Cucchi et al., 2011). The domestication process occurred independently in multiple locations across Eurasia (Larson et al., 2005; Groenen et al., 2012). Subsequently, during the 18th and early 19th centuries, the Asian pigs were introduced into Europe and were hybridized (Kijas and Andersson, 2001). As explained by White (2011): "The coming of Chinese swine stock, with its enhanced capacity for rapid weight gain, played a critical role in transforming Western pigs from peasant subsistence to industrial meat production." Early in the 20th century, herd books were set up for the various breeds in most European countries. This work was initiated by national governments and implemented regionally. All herd books focused on purebred selection and breeding for efficient pork production. In 1964, Smith demonstrated the benefits of crossbreeding programs in pigs (Smith, 1964). Since then, crossbreeding has been widely practiced by pig breeders. In crossbreeding programs, breeds were developed into specialized sire and dam lines. In parallel, breeding farms also became specialized for pure breeding, multiplication and/or crossbreeding, and commercial farms for piglet production and/or fattening. The commercial farms bought replacement gilts and boars from the breeding farms instead of breeding these replacement pigs themselves (Visscher et al., 2000). This is how pig-breeding programs (and companies), as we know them today, started. Currently, any pig-breeding program involves a pyramid breeding scheme, where selection is within purebreds at breeding farms to optimize the performance of crossbreds in commercial farms (Figure 1.1).

The only difference between the breeding programs of the late 60s and the breeding programs nowadays, is the type of information collected and how it is used to select the best pigs. In the early days, pigs were selected based only on their physical characteristics, (e.g., number of teats, body weight) or on their performance, (e.g., litter size, average daily gain). These observable traits, known as phenotypes, were recorded, and the pigs with the best phenotypes were selected to become parents of the next generation. This is phenotypic selection or mass selection, and it is the simplest form of selection. Soon after, it was realized that if a trait is recorded for various relatives, these data could be combined to better predict the potential of a pig as a parent for that trait. This is known as a selection index. Nowadays, the potential of a pig is predicted using phenotypes and family relationships via mixed-model equations. Mixed-model equations, in

comparison to the selection index, allow fixed effects in the data to be accounted for in an optimum way. The predictions from the mixed-model equations are known as estimated breeding values. Accounting only for relationships through the sire led to a “sire model”, and accounting for all relationships among all animals led to an “animal model”. Family relationships used in mixed-model equations were firstly established based on pedigree but nowadays are based also on genomic information.

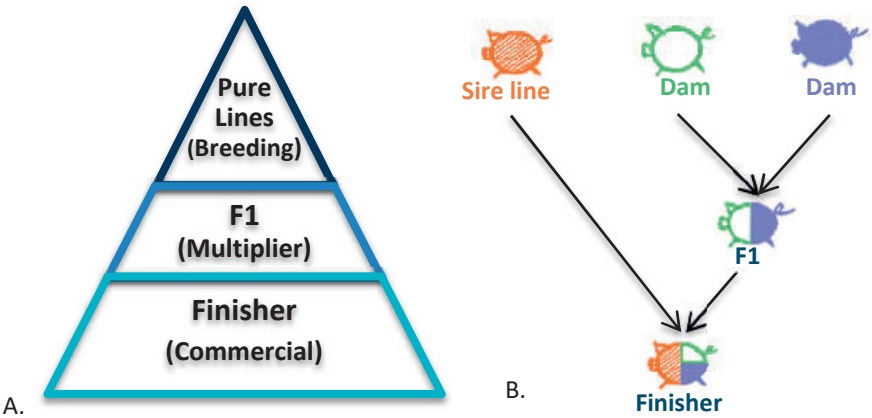

Figure 1.1 Pyramid breeding system (**A**) and three-way terminal crossbreeding scheme (**B**).

Following these developments in selection strategies, a future improvement will be selection of purebred pigs based on their potential to produce crossbred offspring with superior performance, i.e., their estimated breeding value for crossbred performance. As already shown in some studies, there is a benefit of using crossbred phenotypes for estimating breeding values of purebred individuals for crossbred performance, when using pedigree (Wei and Van der Steen, 1991) or genomic information (Hidalgo et al., 2015a; Veroneze et al., 2015; Sewell et al., 2018). Importantly, in crossbred animals, the genome is a mosaic of genomic regions inherited from the different parental breeds. As a result, depending on which breed a genomic region was inherited from, it might have different effects. Therefore, a given genomic region could have a different effect on crossbred performance, depending on the breed it was inherited from. In previous studies using real crossbred genomic information, this was largely unaccounted for. Thus, a new model, as proposed by Ibáñẽz-Escriche et al. (2009) and Christensen et al. (2014), that accounts for SNP allele differences according to the breed of origin,

might be beneficial if we wish to start using crossbred genomic information for estimation of breeding values of purebred pigs for crossbred performance. In this thesis, I evaluated the outcome of using genomic information from commercial crossbred individuals for estimating breeding values of purebred individuals for crossbred performance. More precisely, I evaluated the outcome of accounting for the effects of genomic regions which are dependent on the breed of origin.

Text box 1 provides a glossary of some of the technical terms that geneticists and animal breeders use, and that I will use throughout this thesis.

Box 1. GLOSSARY
Quantitative trait A particular phenotypic trait. They are continuous, usually affected by many genes (polygenic) and by environmental factors. In contrast, qualitative traits, are expressed in different categories, typically monogenic and little affected by environmental factors. Quantitative traits are the most interesting for animal breeders.
Locus (plural loci) A position in the DNA, such as the position of a gene or a marker.
QTL (Quantitative trait locus) A locus variation which is statistically associated with the variation of a quantitative trait. A QTL is not necessarily a causal mutation or a gene.
SNP (Single-nucleotide polymorphism) A variation in a single nucleotide that occurs at a base position in the genome. SNPs are markers, commonly bi-allelic, used in animal genetics.
LD (Linkage disequilibrium) The co-occurrence, more or less often than expected in a random association, of two alleles in different loci in individuals from a population.
Dominance effect A non-additive effect due to the interaction between alleles at the same locus.
Epistasis effect A non-additive effect due to the interaction between genotypes at different loci.
Allele substitution effect (α) The effect that the presence of a copy of that allele has on the phenotype. This depends on the additive effect (a), dominance effect (d) and allele frequencies (p and q for bi-allelic loci) in the observed population. $$\boldsymbol{\alpha = a + d(q - p)}$$

Breeding value The sum of the effects of the alleles carried by an individual.
Additive genetic variance The main cause of resemblance between relatives. It is the variation in true breeding values across all the individuals in a population. $$V_A = 2pq\alpha^2$$
Heterosis Also known as hybrid vigor, this is the deviation of any trait in a crossbred offspring compared to the parental average. One of the main causes of heterosis is dominance.

1.2 Crossbreeding

The practice of crossing different breeds or lines to generate descendants is known as crossbreeding, and this leads to an improvement of the performance of the crossbred offspring compared to the average performance of their parents. The improvement is obtained from heterosis and breed complementarity (Visscher et al., 2000). One of the main causes of heterosis is dominance (Falconer et al., 1996). This is because crossbred superiority is attributed to the advantage of heterozygotes over the mean of the two homozygotes. Crossbreeding is common practice for pigs and poultry, and in cattle, the use of crosses or composite breeds also makes a contribution to the beef and dairy industries. Large-scale commercial pig producers routinely use a terminal crossbreeding system. In this system, F1 females from two maternal breeds are mated to purebred or F1 boars. All pigs from this cross, i.e., commercial crossbreds, are sold as market pigs, thus making this a terminal crossbreeding system. A terminal crossbreeding system typically involves a pyramid breeding program (Figure 1.1), in which selection is within purebred individuals at the nucleus level to optimize the performance of commercial crossbred individuals in commercial farms. However, the genetic progress obtained at the nucleus level is not fully transferred to the commercial level. For traits presenting low (<0.8) genetic correlation between the performance of purebred and commercial crossbred (r_{pc}) individuals, the use of information from commercial crossbred individuals has the potential to maximize genetic progress at the commercial level (Wei and van der Werf, 1994; Bijma and Van Arendonk, 1998). In their review paper, Wientjes and Calus (2017) reported the r_{pc} for 39 different traits measured in 27 studies using either two-way or three-way crossbred pigs (Table 1.1).

Table 1.1 Genetic correlations between the performance of purebred and crossbred (r_{pc}) individuals per trait category (Wientjes and Calus, 2017).

Trait category	Traits (number of r_{pc} estimates per trait)	Total number of r_{pc} estimates	Avg. r_{pc}
Growth	Average daily gain (28), body weight (13), age at test weight (2)	43	0.66
Meat amount	Back fat (30), lean meat content (16), muscle depth (3), carcass length (2), meat content (2), muscle area (2), meat:fat ratio (2), ham content (2), body length (1), belly meat content (1), lipid deposition (1), protein deposition (1)	63	0.69
Meat quality	Meat pH (11), conductivity (5), meat clarity (2), meat quality score (1), drip loss (1), intramuscular fat (1)	21	0.67
Feed	Feed conversion ratio (4), feed efficiency (2), feed intake (1), residual energy intake (1) residual feed intake (1)	9	0.67

There are two major causes that lead to r_{pc} being lower than unity. Firstly, additive genetic variance between purebred and commercial crossbred individuals can be different (Dekkers, 2007). As previously mentioned, crossbreeding maximizes heterosis, and one of the main causes of heterosis is dominance (Falconer et al., 1996). Therefore, differences in additive genetic variance between purebred and commercial crossbred individuals can be due to differences in allele frequencies that affect the contribution of dominance effects to the additive genetic variance (see glossary for an explanation of how additive genetic variance is derived). Secondly, purebred animals are kept under nucleus conditions including superior management and biosecurity, in contrast to the range of conditions and disease challenges experienced by commercial crossbred animals. These differences in environment can cause (partly) different genes to be important for the trait when measured in nucleus or commercial environments, i.e., genotype-by-environment interaction (GxE). Moreover, global consolidation of pig-breeding companies has resulted in commercial farms all around the world with specific local circumstances (such as tropical climate or alternative feed). These differences can also result in GxE across commercial environments. Therefore, to take into account genetic

differences between purebred and commercial crossbred individuals, and the local circumstances in which crossbred animals need to perform, selection strategies should include information from crossbred animals in commercial farms (Bijma and Van Arendonk, 1998).

This thesis is part of the NWO-project entitled "LocalPork". Text box 2 provides an overview of the project and of the three theses involved in the project.

Box 2. LocalPork Project

Efficient local production of pork in Brazil is essential for meeting the increasing internal demand for animal protein, and to make the Brazilian pork sector competitive and sustainable in the future. Two important threats are: 1) Brazilian pork production relies on feeding corn and soybeans, which are becoming more expensive due to the large demand for alternative uses and increasing transport distances and 2) consolidation of pig-breeding businesses has resulted in global breeding programs that do not necessarily select the best pigs for specific local circumstances (such as tropical climate or alternative feed). This multidisciplinary project aimed to quantify these threats, develop and evaluate alternatives involving feeding by-products from more locally produced sources and develop breeding strategies that allow global breeding plans to serve specific local breeding goals.

Thesis: "Enhancing the environmental and economic sustainability of pig farming"

Brazilian pig production relies heavily on high-quality feed ingredients (corn and soybean) and exotic pig breeds that are not bred for local production circumstances. This has caused economic and environmental problems. Economic problems follow from the increasing competition for corn and soybeans between the pig industry and other sectors, which ultimately results in rising feed costs and shrinking farm profits. The problems are exacerbated by feed and pork price volatility, which brings uncertainty that affects investment, production and other business decisions of farmers. Environmental problems follow from the strong dependence on scarce resources, (e.g., cropland, fossil fuel and water) and the release of pollutants into the air, water and soil. This thesis assessed the contributions of locally produced alternative feed sources and the genetic improvement of pigs to enhancing the environmental and economic sustainability of the Brazilian pig production system.

Thesis: "Genotype by environment interaction for feed efficiency in growing-finishing pigs in Brazil versus the Netherlands"

Improving the feed efficiency of crossbred pigs in commercial environments is a priority in pig-breeding programs. Selection for feed efficiency, as for other traits, traditionally relies on measuring the performance of purebred pigs in the nucleus environment, although the aim is to improve crossbred performance in the commercial environment. Thus, the differences between these two environments may give rise to genotype-by-environment interaction (GxE). When comparing nucleus farms in the Netherlands and commercial farms in Brazil, several environmental factors could be responsible for GxE, e.g., the management and hygiene status of the farms, the ingredients of the diets and the climate conditions. This thesis investigated the genetic correlation between purebred and crossbred performance and the GxE interactions for feed efficiency traits in pigs raised in different conditions and fed different diets.

Thesis: "Genomic evaluation considering the mosaic genome of the crossbred pig" (this thesis)

Feed efficiency plays an important role in the breeding goals of today's pig industry and it is one of the most important traits for efficient local production. However, if traits included in the breeding goal have genetic correlations between purebred and crossbred performance which differ from unity, selection response at the nucleus level (purebred animals) will not be fully expressed in the rate of genetic change at the commercial level (crossbred animals). The success of breeding programs in the near future will rely on the use of phenotypes and genotypes taken from crossbred animals at local commercial levels and the use of newly developed genomic models for handling this new type of information. This thesis investigated and developed new methodologies for using crossbred genomic information to increase the genetic change at the commercial level.

1.3 Selection in pigs

The aim of animal breeding is to select the best animals to be parents of the next generation, and thereby to improve the performance of future generations, i.e., genetic progress. The best animals are selected based on estimated breeding values. Traditionally, these estimated breeding values were calculated based on phenotypes and pedigrees, most often using best linear unbiased prediction (BLUP) via mixed-model equations (Henderson, 1975). With the possibility of using DNA information, new selection strategies were developed, firstly marker-assisted selection (Fernando and Grossman, 1989) and later genomic selection (Meuwissen

et al., 2001). Marker-assisted selection uses information on markers that are associated with a specific QTL to estimate breeding values for that trait (Dekkers, 2004). Marker-assisted selection was performed using microsatellite markers. The problem with this strategy is the difficulty in identifying markers affecting the trait and the low availability of microsatellite markers. In addition, the link between a microsatellite marker and a QTL is family-dependent, and therefore the association had to be established for every family in which the marker was to be used for selection (Meuwissen et al., 2001). Genomic selection uses a large number of genome-wide markers simultaneously to estimate genomic breeding values. In contrast to marker-assisted selection, in genomic selection all available markers are used without prior testing of their association with the trait. The most common markers used for genomic selection are single-nucleotide polymorphism (SNP) markers, because they are abundant and easy to genotype, therefore they are expected to be in linkage disequilibrium with the QTL. Therefore, SNPs can be used to explain the QTL variation. The first SNP chip for pigs with about 60,000 SNPs became available in 2009 (Ramos et al., 2009). The most common method for genomic selection is the genomic best linear unbiased prediction (GBLUP) VanRaden (2008) method, in which the pedigree-based relationship matrix is replaced by a genomic relationship matrix in the mixed-model equation. The benefit of genomic selection compared to pedigree-based selection for purebred pig populations was shown by Forni et al. (2010), Ibáñez-Escriche et al. (2014), Knol et al. (2016) and Lopes et al. (2018) using real data. In pig breeding, an extension of the GBLUP, the so-called 'single-step' method, is the most frequently used method for genomic selection (Aguilar et al., 2010; Christensen and Lund, 2010). The single-step method modifies the pedigree-based relationship matrix to include the genomic information. In this way, the evaluation is able to deal with both genotyped and non-genotyped animals.

1.4 Selection for crossbred performance

Additive genetic variance and local environments for purebred and commercial crossbred individuals can be different, resulting in a genetic correlation between performance of purebred and crossbred (r_{pc}) individuals different from unity. The more r_{pc} deviates from unity, the more important information on commercial crossbred individuals is for selecting purebred individuals for crossbred performance. A selection method that includes phenotypic information from crossbred individuals for selection of purebred individuals, denoted combined crossbred and purebred selection (CCPS), was evaluated by Wei and van der Werf

(1994). CCPS was shown to be superior to purebred selection for r_{pc} values lower than 0.8. CCPS improves accuracy but requires larger purebred populations to avoid an excess of inbreeding, leading to a decrease in selection intensity. The inbreeding problem for CCPS arises from the fact that estimated breeding values of purebred selection candidates are largely determined by the performance of crossbred half-sibs, especially for traits with low r_{pc}. This is because CCPS only uses pedigree (not genomics) for selection, therefore there is no information on the Mendelian sampling term of the purebred selection candidates to allow differentiation between sibs. This induces high intraclass correlation between estimated breeding values of full- and half-sibs and a tendency to between-family selection (Bijma et al., 2001). CCPS has not been extensively implemented in breeding programs, mainly due to the cost of collection of crossbred phenotypes in each generation (Kinghorn et al., 2010). The strategy of breeding organizations that implemented CCPS was to contract a limited number of test farms and slaughterhouses where crossbred information was collected. Pig-breeding companies include the phenotypic crossbred information in their routine evaluation through single-step genomic evaluation. Going one step further and including genomic crossbred information in the genetic evaluation is more accurate and will overcome the limitations of inbreeding (Dekkers, 2007). It will also overcome the need for extensive pedigree recording of commercial crossbred individuals, although pedigree recording, even for commercial crossbred individuals, is a routine task in most pig-breeding companies. Breeding companies recently started investing in genotyping commercial crossbred animals. How this information can be handled and its relevance for calculation of estimated breeding values is part of the research described in this thesis.

1.5 Genomic selection for crossbred performance

When information on commercial crossbred individuals is included in genomic selection models, estimated SNP allele effects for crossbred performance may be specific to the parental breed of origin. At least three possible components are known that influence SNP allele effects and can give rise to differences between populations; in this case, parental breeds. The first component is the extent of linkage disequilibrium between SNPs and the QTL. An inconsistent linkage disequilibrium phase of the SNP and QTL alleles between populations can explain why an SNP associated with an important effect in one population may not be effective for selection in a second population (De Roos et al., 2009). The persistence of phase between crossbreds and their parental breeds was higher

than 0.9 for distances of 50 kb for two-way crossbreds (Grossi et al., 2017) and three-way crossbreds (Veroneze et al., 2014). Across purebred populations, persistence of phase was around 0.8 when marker distances were smaller than 50 kb (Badke et al., 2012; Veroneze et al., 2014; Grossi et al., 2017). Therefore, the currently used SNP chip with around 60,000 SNPs equally distributed across the pig genome, having an average spacing of 47 kb, should allow for genomic selection with crossbred information.

The second possible component is the difference in allele frequencies of both the QTL and SNPs across populations. The QTL might segregate differently by population and may even not segregate at all in one population. These differences in allele frequencies by population result in differences in genetic variance explained by a QTL and by the SNP in LD with that QTL.

The third possible component is the presence of epistatic interactions, which implies that the effect of a QTL allele depends on the genotype and allele frequency of a locus with which the QTL interacts. If the allele frequency of the locus with which the QTL interacts varies among populations, the effect of the QTL allele can be significant in one population but not in another, or can even be of the opposite sign in two different populations (Mackay, 2014).

As SNP allele effects in crossbred animals may be specific to the parental breed of origin, genomic selection using crossbred genotype information could benefit from models that estimate different SNP allele effects for crossbred performance depending upon the breed of origin, as suggested by Dekkers (2007). Thus, implementing a model that uses genomic information both from purebred and commercial crossbred individuals and that takes into account the breed of origin of alleles, might improve estimation of breeding values of purebred individuals for crossbred performance. With simulated data, this model performed better than a model where SNP allele effects are assumed to be the same across breeds, when the number of markers was small and when breeds were distantly related or unrelated (Ibáñẽz-Escriche et al., 2009). With real data, predictions will be affected by the history of the breeds, non-additive effects (if present) and the errors that might be included when tracing the origin of the alleles. The performance of this model has not been investigated with real data from three-way crossbred commercial pigs.

1.5 Outline of the thesis

The objective of this thesis was to investigate new methodologies for using crossbred genomic information to increase the genetic progress at the commercial level. The hypothesis was that genomic selection using crossbred genotype information could benefit from models that estimate different SNP allele effects for crossbred performance, depending upon the breed of origin. To investigate this, I firstly needed to evaluate an approach that assigns the breed of origin of alleles in real data on three-way crossbred pigs, and this is discussed in **chapter 2**. I obtained genomic information from three populations, i.e., purebred, F1 and three-way crossbred populations. Following the approach for assigning the breed of origin to alleles, I phased the genotypes of all the pigs and determined the unique haplotypes among the purebreds. Finally, I assigned the breed of origin to each allele carried on the haplotypes of crossbred pigs, based on the knowledge of the breed of origin of the haplotypes, on the zygosity, (i.e., homozygosity or heterozygosity) of the locus and on the breed composition of the crossbred. I measured the percentage of assignments when the phasing was performed using or ignoring pedigree information. This chapter demonstrates that around 94% of the alleles of three-way crossbred pigs were assigned a breed of origin. Therefore, it was possible to implement a model that estimated SNP allele effects for crossbred performance based on their breed of origin (BOA model).

In **chapter 3** I evaluated the performance of the BOA model. I assigned the breed of origin to each allele carried on three-way crossbred pigs using the approach tested in **chapter 2**. Subsequently, I compared the results from the BOA model to those from models in which SNP allele effects for crossbred performance were assumed to be the same across breeds (G model), using either breed-specific allele frequencies or allele frequencies averaged across breeds. In this chapter, I used phenotypes from purebred and three-way crossbred pigs on average daily gain, back-fat thickness and loin depth. The BOA model predicted the crossbred performance better for average daily gain, but only for one of the maternal purebred lines, which showed the lowest genetic correlation between performance of purebred and crossbred (r_{pc}).

In **chapter 4** I investigated whether the allele effect of a given SNP differs depending on the environmental or genetic background where it is expressed. In **chapter 3**, the applied BOA model assumed that allele effects of all SNPs were breed-specific. This assumption might not hold for all the SNPs, and possible

reasons are discussed in **chapter 4**. In this chapter, I used phenotypes from purebred and three-way crossbred pigs on average daily gain, back-fat thickness and residual feed intake. This chapter demonstrates that the effect of haplotypes strongly associated with crossbred performance are different, depending upon the population from which they originate. This chapter also shows, however, that the majority of the genomic regions are not, or are only weakly, associated with crossbred performance.

With this knowledge in mind, in **chapter 5** I developed the SEL-BOA model, that accounts for breed-specific allele effects only for SNPs strongly associated with crossbred performance. For the rest of the SNPs the SEL-BOA models assumes the same effects across breeds. I selected the SNPs strongly associated with crossbred performance based on the results from **chapter 4**, so that they explain together 5% or 10% of the total crossbred genetic variance for average daily gain in each breed of origin. I compared the results from the SEL-BOA model to those from the G and BOA model. Differences of prediction accuracies between models were small. The BOA model predicted better crossbred performance than the G model when estimated crossbred genetic variances and r_{pc} differ largely between the G and the BOA models. Superiority of the SEL-BOA model compared to the BOA model was only observed for scenario 10% and when r_{pc} for the selected SNP and non-selected SNP differed strongly from the r_{pc} estimated by the G or BOA model.

In the last chapter, **chapter 6**, I present the main contributions of this thesis to the understanding of the mosaic genome of crossbred pigs and the results of using this knowledge in models for predicting crossbred performance. Further, I discuss aspects of the implementation of these prediction models in breeding programs, as well as potential avenues for research on alternative prediction models.

Chapter 2

Empirical determination of breed-of-origin of alleles in three-breed cross pigs

Claudia A. Sevillano[1,2], Jeremie Vandenplas[1], John W.M. Bastiaansen[1], Mario P.L. Calus[1]

[1] Wageningen University & Research Animal Breeding and Genomics, 6700AH, Wageningen, the Netherlands; [2] Topigs Norsvin Research Center, 6640AA, Beuningen, the Netherlands

Genetics Selection Evolution (2016) 48:55

Abstract

Although breeding programs for pigs and poultry aim at improving crossbred performance, they mainly use training populations that consist of purebred animals. For some traits, e.g. residual feed intake, the genetic correlation between purebred and crossbred performance is low and thus including crossbred animals in the training population is required. With crossbred animals, the effects of single nucleotide polymorphisms (SNPs) may be breed-specific because linkage disequilibrium (LD) patterns between a SNP and a quantitative trait locus (QTL), and allele frequencies and allele substitution effects of a QTL may differ between breeds. To estimate the breed-specific effects of alleles in a crossbred population, the breed of origin of alleles in crossbred animals must be known. This study was aimed at investigating the performance of an approach that assigns breed of origin of alleles in real data of three-breed cross pigs. Genotypic data were available for 14,187 purebred, 1354 F1, and 1723 three-breed cross pigs. On average, 93.7 % of the alleles of three-breed cross pigs were assigned a breed of origin without using pedigree information and 94.6 % with using pedigree information. The assignment percentage could be improved by allowing a percentage (f_r) of the copies of a haplotype to be observed in a purebred population different from the assigned breed of origin. Changing f_r from 0 to 20 %, increased assignment of breed of origin by 0.6 and 0.7 % when pedigree information was and was not used, respectively, which indicates the benefit of setting f_r to 20 %. Breed of origin of alleles of three-breed cross pigs can be derived empirically without the need for pedigree information, with around 93.0 % of alleles assigned a breed of origin. Pedigree information is useful to reduce computation time and can slightly increase the percentage of assignments. Knowledge on the breed of origin of alleles allows the use of models that implement breed-specific effects of SNP alleles in genomic prediction, with the aim of improving selection of purebred animals for crossbred offspring performance.

Key words: SNP, crossbred, imputation, phasing, origin of alleles, pig

2.1 Introduction

The genetic correlation between purebred and crossbred performance (r_{pc}) is a crucial parameter that determines the effect of selection at the nucleus level, where purebred animals are used, on the rate of genetic change at the production level, where crossbred animals are used (Wei and van der Steen, 1991; Brandt and Täubert, 1998). In many cases, r_{pc} deviates from 1 because of (1) different genetic backgrounds, and (2) different management procedures for purebreds and crossbreds (Wei and van der Steen, 1991; Lutaaya et al., 2001; Bastiaansen et al., 2014). As r_{pc} decreases, the benefit of using crossbred information increases (Wei and van der Steen, 1991; Bijma and van Arendonk, 1998), e.g. Dekkers (2007) reported that even with a r_{pc} as low as 0.7 using crossbred information was advantageous. When information on crossbred animals is used, effects of single nucleotide polymorphisms (SNPs) may be breed-specific because linkage disequilibrium (LD) patterns between a SNP and a quantitative trait locus (QTL) (Bastiaansen et al., 2014) and allele frequencies and allele substitution effects of a QTL may differ between breeds (Wientjes et al., 2015). With genomic prediction, it is possible to determine the effect of alleles from different breeds and, thus, it can be used to select purebred animals for crossbred performance. An additive model that accounts for breed-specific SNP effects for genomic prediction using crossbred information was proposed by Ibáñẽz-Escriche et al. (2009) and Christensen et al. (2014; 2015). Ibáñẽz-Escriche et al. (2009) and Esfandyari et al. (2015) showed with simulated data that, under some conditions (i.e., low SNP density, large training data size, and low breed relatedness), the model that accounts for breed-specific SNP effects outperformed models in which SNP effects are assumed to be the same across breeds. If the above-mentioned conditions that favor the model that accounts for breed-specific effects with simulated data are met in real then it is important to determine whether such models are also superior in real data.

To estimate the effect of a SNP allele that is present in a crossbred animal and originates from a purebred animal, the breed of origin of alleles in crossbreds must be known. While breed of origin of alleles was assumed to be known without error by Ibáñẽz-Escriche et al. (2009), errors in breed of origin of alleles and the total percentage of alleles assigned to a breed of origin likely impact the accuracy of subsequent analyses such as genomic prediction.

For a two-way cross, determining the breed of origin of alleles is relatively easy, especially when both parents are genotyped (Lopes et al. 2016). However, in pig

and chicken production, three-way crosses are commonly used. Bastiaansen et al. (2014) developed an approach to assign breed of origin to alleles in three-breed cross animals. They used a long-range phasing method (Hickey et al., 2011) to relax the dependency on genotyped parents and available pedigree information. Haplotypes that were derived from the long-range phasing method were assigned to a breed if they were present in only one of the purebred populations, which subsequently allowed assigning the breed of origin of alleles when that haplotype was observed in crossbred animals. Vandenplas et al. (2016) improved and tested the approach to assign breed of origin of alleles on simulated data and obtained highly accurate allele assignments in three-breed cross animals without using pedigree information. Our aim was to investigate the performance of assignment of breed of origin of alleles on real data of three-breed cross pigs. The impact of using pedigree information of the crossbred animals on the assignment of breed of origin of alleles was also tested because in this dataset the pedigree was completely known and this approach is able to use such information when available.

2.2 Methods

2.2.1 Ethics statement

The data used for this study was collected as part of routine data recording in a commercial breeding program. Samples collected for DNA extraction were only used for routine diagnostic purposes of the breeding program. Data recording and sample collection were conducted strictly in line with the Dutch law on the protection of animals (Gezondheids- en welzijnswet voor dieren).

2.2.2 Genotypic data

We used pigs that originated from a three-way crossbreeding design, in which Landrace (LR) pigs were crossed with Large White (LW) pigs to produce F1 (LR x LW or LW x LR) crossbred pigs, which in turn were crossed with synthetic boar (S) pigs to produce three-breed cross pigs (S (LR x LW) or S (LW x LR)). Genotyping data was available for 14,187 purebred, 1354 F1, and 1723 three-breed cross pigs (Table 2.1). All pigs were genotyped using one of the three following SNP panels: Illumina PorcineSNP60.v2 BeadChip (60K.v2), Illumina PorcineSNP60 BeadChip (60K), or Illumina PorcineSNP10 BeadChip (10K) (see Table 2.1 for details). LR, LW and S pigs were primarily genotyped with the 60K (N = 2352), 10K (N = 3618), and 10K (N = 1233) chips, respectively. F1 pigs were primarily genotyped with the 60K.v2 (N = 786) chip and three-breed cross pigs with the 10K (N = 1432) chip. SNPs were removed from the data if they had the same position as another SNP (only one

removed), if they had no position assigned, or if they were present on *Sus scrofa* chromosome (SSC) X or SSCY. The SNP set for subsequent analyses consisted of SNPs from the 60K.v2 that had a call rate higher or equal to 90 % across all purebred lines. Pigs genotyped with the 60K or 10K chips were imputed to the 60K.v2 panel. SNPs with low imputation accuracy across all purebred lines and F1 crossbreds (concordance < 0.80) were removed from the final set of SNPs. Finally, 52,164 SNPs remained for the analyses (Fig. 2.1).

Table 2.1 Number of genotyped pigs available per SNP panel, and per crossbred line or cross

SNP panel	S	LR	LW	F1	3-breed cross	Total
60K.v2	810	914	878	786	0	3388
60K	782	2352	2687	543	291	6655
10K	1233	913	3618	25	1432	7221
Total	2825	4179	7183	1354	1723	17 264

S = Synthetic boar, LR = Landrace, LW = Large White, F1 = LR x LW or LW x LR, 3-breed cross = S (LR x LW) or S (LW x LR).

2.2.3 Imputation

FImpute Version 2.2 software (Sargolzaei et al., 2014) was chosen for imputation with default parameter settings and using pedigree information because it is one of the most efficient available software programs for imputation (Sargolzaei et al., 2014; Ventura et al., 2014). Within each of the three purebred lines, LR, LW, and S, imputation was performed in two steps: (1) pigs genotyped with the 10K chip were imputed to 60K, and (2) all pigs with 60K data were imputed to 60K.v2. For F1 and three-breed cross pigs, imputation was done in a single step, i.e. pigs genotyped with the 10K and 60K chips were directly imputed to 60K.v2, because all ancestors were genotyped or already imputed to 60K.v2. The numbers of SNPs from each panel that were used in each imputation step are in Fig. 2.1.

2.2.3.1 Validation of imputation

Imputation accuracy was assessed in 80 pigs from each of the purebred lines, LR, LW, and S, and in 80 F1 crossbred pigs, which were all genotyped with the 60K.v2 panel. Accuracy of imputation in three-breed cross pigs was not assessed because none of them were genotyped with the 60K.v2. All pigs that were chosen to assess imputation accuracy had no offspring and both their parents were genotyped with the 60K.v2, 60K, or 10K chips. In these pigs, the genotypes of all SNPs on the 60K.v2 panel were set to missing, except for the SNPs that were also on the 10K panel.

Imputation accuracy was calculated for each SNP in two ways, based on concordance and Pearson correlation, using the real and imputed genotypes. Pearson correlations per SNP between the real and imputed genotypes were corrected for minor allele frequency (MAF), i.e., real genotype – 2*MAF and imputed genotype – 2*MAF. The MAF for each SNP was calculated using the data for the 80 pigs tested from each population. SNPs with low imputation accuracy across all purebred lines and F1 crossbreds (concordance < 0.80) were removed from the final set of SNPs.

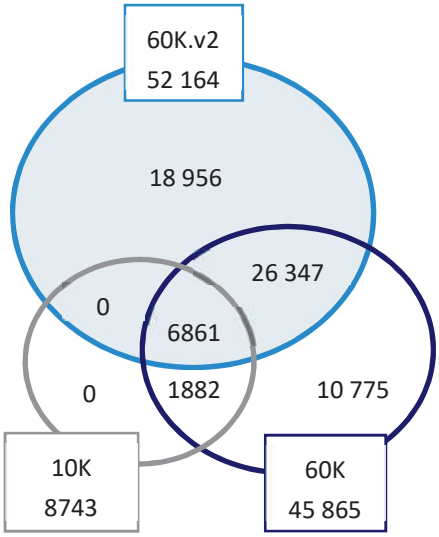

Figure 2.1 Distribution of SNPs across the three different SNP panels after pruning. SNPs within the shadowed blue circle are included in the final set of SNPs. SNPs outside the shadowed blue circle were used during the imputation procedure.

2.2.4 Assignment of breed of origin of alleles

To assign breed of origin of alleles to three-breed cross pigs, we used an approach that consisted of three steps : (1) phasing the haplotypes of both purebred and crossbred pigs, (2) determining the unique haplotypes among the pure breeds, and (3) assigning the breed of origin for each allele carried on the haplotypes of crossbred pigs, i.e. F1 and three-breed cross pigs. For these steps, we used all the 52 164 SNPs in the final set.

2.2.4.1 Phasing

AlphaPhase1.1 software (Hickey et al., 2011) that implements a long-range phasing and haplotype library imputation algorithm was used to phase the genotypes. Although FImpute (Sargolzaei et al., 2014) also searches for long shared haplotypes and builds a haplotype library, the breed of origin approach cannot use this program because it also searches for short shared haplotypes. Short shared haplotypes can be difficult to assign to a breed because they may be shared across breeds. Long-range phasing is of particular interest because it does not rely on pedigree information. However, we tested both scenarios, phasing with and without pedigree information, to assess if allele assignment was improved when using pedigree information. Due to computational limitations, assigning breed of origin without using pedigree information was performed only for chromosomes 3, 4, 9, 12, and 16. For both scenarios, haplotypes were built using nine combinations of core and tail lengths: 350:50, 250:100, 300:100, 350:100, 150:200, 200:200, 250:200, 300:200, 350:200. The concepts of core and tails are outlined in detail in Hickey et al. (2011). Briefly, a core is a consecutive string of SNPs that are phased simultaneously, while tails are consecutive strings of SNPs that are immediately adjacent to either end of a core and that are used together with the core SNPs to identify which pigs in the data carry the same haplotype. Each combination of core and tails was run both considering “Offset” and “NotOffset” modes. The “Offset” mode shifts the start of the cores to halfway along the first core, creating 50 % overlaps between cores. These settings were chosen based on results of Vandenplas et al. (2016) and allowed each allele to be considered 18 times through different haplotypes of variable length. Varying the haplotype lengths may improve the overall phasing when some animals do and others do not have close relatives present in the genotype data. For all phasing analyses, 1 % of genotype errors and 1 % disagreement between genotypes and haplotypes were allowed.

2.1.4.2 Assignment of breed for haplotypes and alleles

Assignment of breed of origin to haplotypes was performed as in Vandenplas et al. (2016). To assign a breed of origin to a haplotype, it was necessary that most of its copies were present in a specific breed. We tested two relaxation factors (f_r), i.e. 0 and 20 %, which is the maximum percentage of the copies of a haplotype that may be observed in a different purebred population. When the percentage of the copies of a haplotype that was observed in a single breed was less than (100- f_r) %, the breed of origin for that haplotype was set to unknown.

Assignment of breed of origin to each allele that is carried on the haplotypes of crossbred animals is based on the knowledge available for the breed of origin of the haplotypes, the zygosity (i.e., homozygosity or heterozygosity) of the locus, and the breed composition of the crossbred animals (see Vandenplas et al. (2016) for the algorithm). Each allele at each locus can receive 18 breed-of-origin assignments, but, in some analyses, this number can be smaller when no breed is assigned to the haplotype.

2.2.5 Principal components analysis

A principal components analysis (PCA) was performed to check if three-breed cross pigs with a low assignment of breed of origin to their alleles were genetically distinct to the three-breed cross population. The PCA was performed by eigen decomposition of the genomic relationship matrix (G-matrix). The G-matrix was computed as in Yang et al. (2010), using our in-house software calc_grm (Calus and Vandenplas, 2015).

2.3 Results

2.3.1 Imputation and accuracies of imputation

Accuracies of imputation were very close to 1, both when based on concordance and Pearson correlation (Table 2.2). The Pearson correlation between imputed and real genotypes per SNP was greater than 0.96 across all pure lines and F1 pigs (Table 2.2). The Pearson correlation per SNP was very similar across different MAF (Fig. 2.2). Some individual SNPs (N = 406) showed poor imputation accuracy (concordance < 0.80) and were removed from the set of SNPs. The final set of SNPs considered for imputation and assignment of breed of origin for alleles of three-breed cross pigs included 52,164 SNPs from the 60K.v2 panel.

Table 2.2 Average imputation accuracies computed across pigs or SNPs

	Pig	SNP	
	Concordance	**Correlation**	**Concordance**
Landrace	0.99	0.97	0.98
Large White	0.99	0.97	0.98
Synthetic boar	0.98	0.96	0.98
F1 crossbred	0.98	0.97	0.98

Accuracy was computed for the masked loci as the proportion of pigs or loci that had the same observed and imputed genotype (concordance), or the same Pearson correlation between the observed and imputed genotypes.

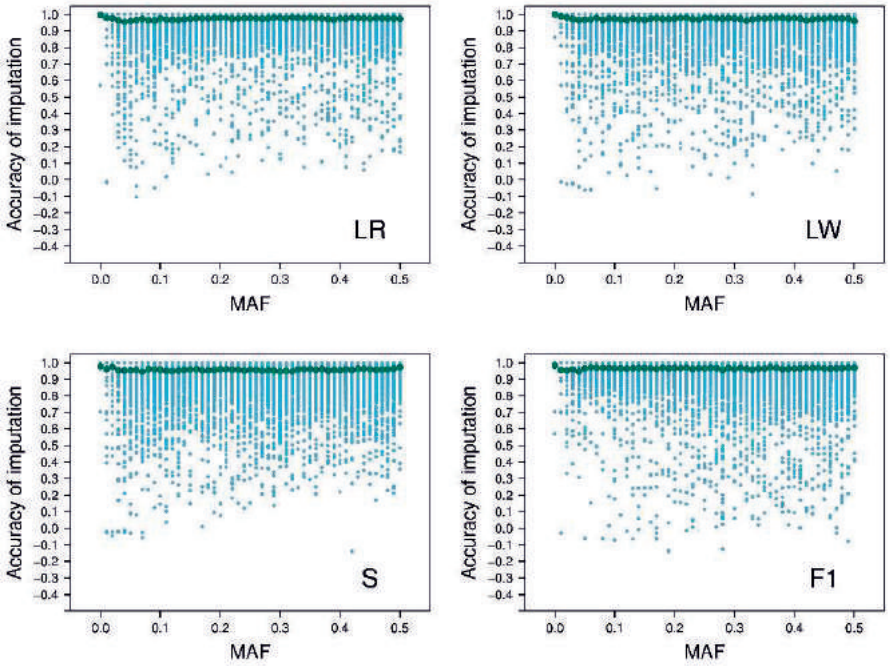

Figure 2.2 Accuracy of imputation according to minor allele frequencies. Minor allele frequencies (MAF) of genotyped SNPs versus the accuracy (Pearson correlation) of imputation from the PorcineSNP10 BeadChip panel to Illumina PorcineSNP60.v2 BeadChip for 80 pigs of each purebred line, i.e. synthetic boar (S), Landrace (LR), Large White (LW), and crossbred F1 pigs. The dark green dots are the average accuracy for different MAF.

2.3.2 Assignment of breed of origin for alleles

2.3.2.1 Comparison of different settings used for assignment of breed of origin

All pigs were used to assign the breed of origin of alleles but the results are presented only for three-breed cross pigs. Breed-of-origin assignments were obtained from analyses without pedigree information for chromosomes 3, 4, 9, 12, and 16, and from analyses with pedigree information for all autosomes. For chromosomes 3, 4, 9, 12, and 16, on average 93.0 % (±1.0 %) of the alleles of a three-breed cross pig were assigned to a breed of origin without using pedigree

information and 94.6 % (±1.0 %) with using pedigree information, both with a f_r equal to 0 % (Table 3). For all autosomes, on average 93.9 % (±1.4 %) of the alleles of a three-breed cross pig were assigned to a breed of origin when using pedigree and f_r set at 0 %. Relaxing f_r from 0 to 20 % increased the assignment by 0.6 and 0.7 % with and without using pedigree information, respectively, for chromosomes 3, 4, 9, 12, and 16, and increased the assignment by 1.3 % with using pedigree information for all autosomes (Table 2.3). In general, increases in assignment percentage were small regardless of whether pedigree information was used or not or whether f_r was set to 0 or 20 %.

Table 2.3 Allele assignment (%) to purebred lines as breed of origin when pedigree information is used or not, and with a relaxation factor (f_r) of 0 or 20 %. SD are in parenthesis.

Ped[+]	f_r	Paternal		Maternal		Total
	(%)	Line S	Line LR	Line LW	Total	
No[1]	0	49.5 (0.25)	22.4 (0.59)	21.1 (0.38)	43.5 (0.80)	93.0 (1.04)
	20	49.6 (0.23)	22.5 (0.64)	21.6 (0.42)	44.1 (0.82)	93.7 (1.03)
Yes[1]	0	49.7 (0.26)	23.2 (0.48)	21.8 (0.33)	45.0 (0.71)	94.6 (0.97)
	20	49.7 (0.25)	23.0 (0.61)	22.6 (0.83)	45.5 (0.67)	95.2 (0.91)
Yes[2]	0	49.5 (0.46)	22.5 (0.90)	21.8 (0.53)	44.4 (1.13)	93.9 (1.44)
	20	49.6 (0.42)	23.0 (0.65)	22.7 (0.59)	45.7 (0.73)	95.2 (0.95)

Synthetic boar (S), Landrace (LR), Large White (LW)

[+]Ped = Pedigree information (yes or no).

[1]Averages estimated using chromosomes 3, 4, 9, 12, and 16.

[2]Averages estimated using all 18 autosomes.

The assigned breed of origin of alleles for heterozygous genotypes may differ depending on the approach used. To assess the effect of using pedigree information, breed-of-origin assignments with or without the use of pedigree information were compared. Both scenarios included only chromosomes 3, 4, 9, 12, and 16 (Table 2.4, comparison A). Only 0.3 % of the assignments displayed a change in their breed of origin depending on the use of pedigree information or not. Assignments were concordant for 94.2 % of the genotypes and 5.5 % of the genotypes were assigned a breed of origin by only one of the two approaches.

To assess the impact of increasing the relaxation factor, assignments of breed of origin obtained with f_r set at 0 and 20 % were compared. In this case, both scenarios included pedigree information and only chromosomes 3, 4, 9, 12, and 16

were used (Table 2.4, comparison B). Only 0.1 % of the assignments displayed a change in their breed of origin between setting f_r at 0 or 20 %. The assignments were concordant for 99.2 % of the genotypes and 0.7 % of the genotypes were assigned a breed of origin by only one of the approaches. Because differences in breed-of-origin assignments between options were small, only results obtained with pedigree information and an f_r set at 20 % will be presented in the following.

Table 2.4 Comparison between different scenarios for the assignment of breed of origin of alleles

Comparison A			**Comparison B**		
Pedigree	**No pedigree**	**%**	**f_r 20 %**	**f_r 0 %**	**%**
Concordance		94.16	Concordance		99.24
Assigned	Not assigned	3.57	Assigned	Not assigned	0.63
Not assigned	Assigned	1.97	Not assigned	Assigned	0.07
Disagreement		0.30	Disagreement		0.06

(A) Breed-of-origin approach with vs. without pedigree (relaxation factor (f_r) of 0 %)

(B) Breed-of-origin approach with f_r set to 20 % vs. f_r se to 0 % (with pedigree)

Concordance, same allele assigned to the same breed of origin by both scenarios or same allele not assigned to a breed of origin by both scenarios.

Disagreement, same allele assigned to different breed of origins by both scenarios

Allele assigned to a breed of origin by only one scenario (assigned – not assigned or not assigned - assigned).

2.3.2.2 Performance of assignment of breed of origin

Average assignment percentages were similar across three-breed cross pigs. On average, for each chromosome, at least 80 % of alleles were assigned a breed of origin to 98.7 % of the three-breed cross pigs. Of the three-breed cross pigs, 8 % (N = 141) had a chromosome for which less than 80 % of the alleles were assigned and 4 % (N = 66) had multiple such chromosomes. The assignment percentage of these 207 three-breed cross pigs is illustrated in Fig. 2.3. The chromosome that has the lowest percentage of assignment varied across these 207 pigs. The lowest assignment for a chromosome was observed in a three-breed cross pig for which only 19.0 % of the alleles on chromosome 9 were assigned to a breed. For this pig, chromosome 6 had the highest assignment, for which 67 % of the alleles were assigned to a breed. Two three-breed cross pigs, including the one mentioned above, had a low percentage of assignment for all 18 chromosomes (Fig. 2.3).

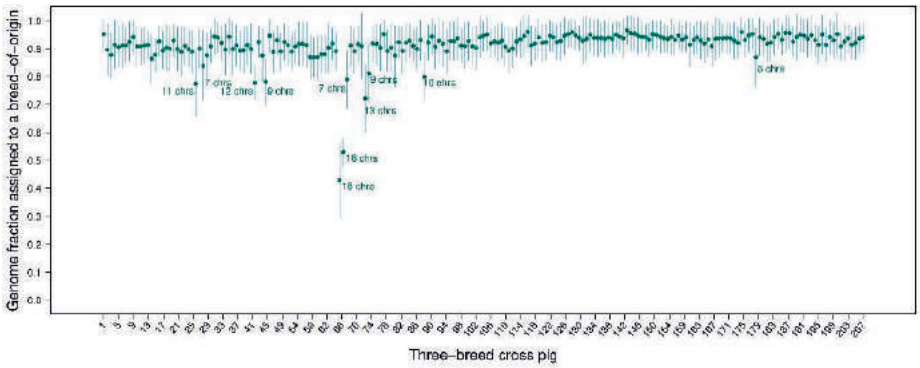

Figure 2.3 Average (±SD) assignment of breed of origin of alleles for 207 three-breed cross pigs. All three-breed cross pigs had at least one of their chromosomes with less than 80 % breed-of-origin assignment of alleles. Numbers of chromosomes per pig with poor assignment are written next to the averages (number is omitted if number of chromosomes is lower than five).

The average assignment of breed of origin of alleles was similar across chromosomes, with a standard deviation (SD) of 0.95 % among the 18 chromosomes. Within chromosome, the SD ranged from 3.36 % for chromosome 1 and 13, to 5.97 % for chromosome 2. The highest assignment was obtained for chromosome 17 (96.5 %) and the lowest for chromosome 12 (93.6 %) [See Supplementary material, Additional file S2.1]. For chromosome 17, 49.8 % of the alleles were assigned to the paternal S purebred line, 23.1 % to the maternal LR purebred line, and 23.6 % to the maternal LW purebred line. For chromosome 12, 49.3 % of the alleles were assigned to the paternal S purebred line, 21.7 % to the maternal purebred LR line, and 22.6 % to the maternal LW purebred line. The main differences between chromosomes were due to differences in the percentage assigned to the maternal purebred lines.

For most three-breed cross pigs, one chromosome of each pair was almost completely assigned to the paternal S purebred line, as shown for 25 random pigs in Fig. 2.4, while the other chromosome showed large blocks that were assigned to the maternal LR or LW purebred line. While it is expected that 50 % of the maternal chromosome originates from one of the two maternal purebred lines, these percentages can deviate strongly from this value for individual animals. The pattern in Fig. 2.4 is as expected based on the 1.2 recombination rate of *Sus scrofa* chromosome (SSC) 12 (Tortereau et al., 2012), and we observed on average one recombination per chromosome. However, near the ends of the maternal

chromosomes, the number of alternate assignments of breed of origin of alleles between the maternal LR or LW purebred lines increased, which is consistent with the higher levels of recombination that are observed in these chromosome regions (Tortereau et al., 2012). Assignment of breed of origin to each allele is also based on the breed composition of the crossbred animals. For one three-breed cross pig, if the origin of the maternal allele is assigned, the algorithm always assigns the paternal origin to the other allele at the same locus, i.e. in Fig. 2.4 no dark grey region is observed opposite to an assigned maternal allele. The other way around, if the origin of the paternal allele is assigned, the algorithm does not necessarily always assign the maternal origin to the other allele at the same locus, because it cannot choose between the 2 maternal purebred lines, as can be observed from dark grey regions opposite to an assigned paternal allele.

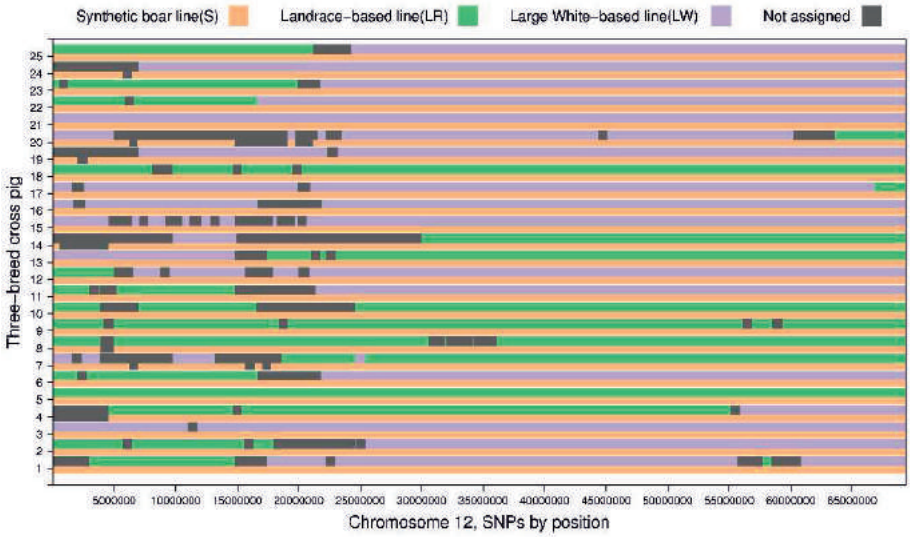

Figure 2.4 Breed of origin of alleles in 25 three-breed cross pigs. Each three-breed cross pig is represented in two rows, one row representing the paternal and one row the maternal chromosome. Dark grey regions indicate unassigned allelic origin. White regions indicate regions that not covered with SNPs.

2.3.3 Principal component analysis

The principal component analysis of the genomic relationship matrix provided a clear separation between the purebred lines and between the three-breed cross pigs (Fig. 2.5). The first and second principal components together explained 16.9 % of variation, while the third principal component only explained 1.9 % of the

variation, which is mainly associated with variation within the LR purebred line population. Previously, we detected two three-breed cross pigs with a low percentage of assignment for all 18 chromosomes. In Fig. 2.5, we plotted the first three principal components of the genomic relationship matrix and we observed that one of these pigs was placed within the paternal S purebred line population, while the other pig was placed outside the three-breed cross population, but also outside all purebred line populations. This indicates that these two pigs were genetically distinct from the three-breed cross population.

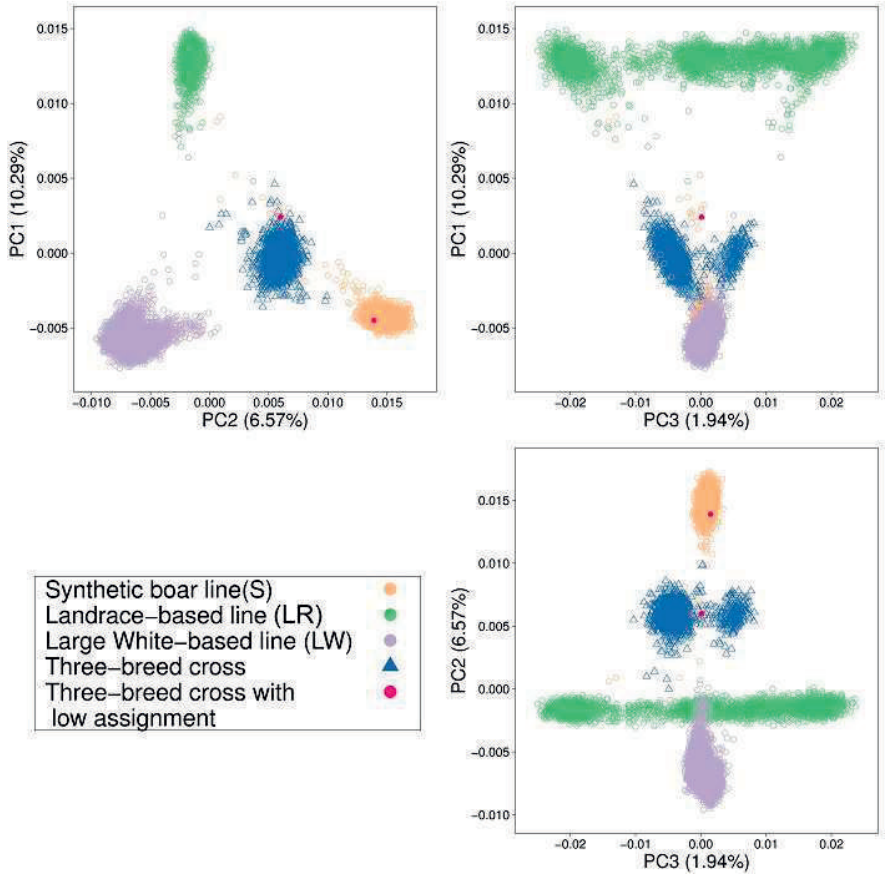

Figure 2.5 Three first principal components (PC) for the three purebred lines and three-breed cross pigs. Each circle (o) or triangle (Δ) represents a pig. The two pink dots represent the two three-breed cross pigs with a low percentage of assignment for all 18 chromosomes.

2.4 Discussion

2.4.1 Imputation

For the three purebred lines, LR, LW, and S, imputation was performed in two steps, 10K genotypes were imputed to 60K, and the output of the first step was imputed to 60K.v2. This strategy was chosen because the 10K panel shared more SNPs (8743) with the 60K panel than with the 60K.v2 panel (6861). Pedigree information was used for the imputation because it was available. However, in the absence of pedigree information and with high-density panels, family information can be captured by searching for long haplotypes and used for imputation (Sargolzaei et al., 2014). The accuracies of imputation that were obtained in our study, using related pigs that were genotyped with high-density panels (60K or 60K.v2) and using pedigree information, were close to accuracies reported in the literature with similar datasets (Gualdron Duarte et al., 2013; Ventura et al., 2014). Gualdron Duarte et al. (2013), imputed 9K genotypes of F2 individuals from a Duroc x Pietrain population to 60K, and obtained an accuracy of imputation higher than 0.94. With our data, the accuracy of imputation per SNP was very similar across different values of MAF, which indicates that rare variants were also accurately imputed. Similarly, Gualdron Duarte et al. (2013), observed that SNPs with a MAF lower than 0.10 were imputed with reasonably good accuracy in the F2 population. Ventura et al. (2014), imputed crossbred beef cattle from 6K to 50K, and concluded that the accuracy of imputation of crossbred animals can be high if the number of reference animals genotyped with high-density panels is sufficiently large and if all breeds that have led to the crossbred animals are included. They also used the FImpute software (Sargolzaei et al., 2014) and obtained imputation accuracies higher than 0.94. However, accuracy of imputation was based only on concordance. Concordance estimates for imputation accuracy are generally higher than Pearson correlations. Imputation errors are generally due to the assignment of the major instead of the minor allele, and the probability of such errors decreases as MAF decreases. Therefore, SNPs with a low MAF generally show high concordance (Lin et al., 2010). Moreover, the slightly lower accuracies reported by Ventura et al. (Ventura et al., 2014) compared to those found in our study, can be explained by the fact that they lacked pedigree information. Another reason may be the higher levels of genomic divergence between the reference population and the group of animals to be imputed. In addition, the structure of the populations may have also contributed to this difference since pig breeding populations have a small effective population size, few boars with large family sizes, and generally complete pedigree information, while beef cattle populations have a larger number of sires with smaller family sizes and incomplete pedigree information (Ventura et

al., 2014). Accuracy of imputation in our three-breed cross pigs data was not assessed because none of these animals were genotyped with the 60K.v2 chip. However, we would expect high imputation accuracies, i.e. similar to the results obtained for the purebreds and F1 pigs. Shared haplotypes should have been found easily and accurately because the reference group was large and related to the target group (Sargolzaei et al., 2014). Moreover, high imputation accuracy of rare variants was also expected in the three-breed cross pigs, because alleles present in the crossbreds must be present in the purebred parental lines (Gualdron Duarte et al., 2013).

2.4.2 Assignment of breed of origin

Percentage of assignment of breed of origin to alleles increased only slightly when pedigree information was used (1.6 % with f_r set at 0 %, and 1.5 % with f_r set at 20 %). Using pedigree information is recommended, first to increase allelic assignments, and second to reduce computation time during the phasing analyses. Only a small difference in assignment of breed of origin between using pedigree information or not was expected, because this information was only used for the phasing step, and it has been shown that long-range phasing, as implemented in AlphaPhase1.1 software, performs well in the absence of pedigree information (Hickey et al., 2011). Percentages of assignments were in line with the results based on simulated data that were reported by Vandenplas et al. (2016). In their simulation study (Vandenplas et al., 2016), the distantly-related breeds scenario is comparable to our real data analysis. We obtained the highest percentage of assignment when using pedigree information and f_r equal to 20 %. Based on the simulation study of Vandenplas et al. (2016), relaxing the maximum percentage of copies of the haplotype observed in another purebred population from 0 to 10 %, and then to 20 %, slightly increased the percentage of correct assignments but did not influence the percentage of incorrect assignments, and consequently slightly decreased the percentage of unknown assignments for crossbred animals that originated from distantly-related breeds. Across our results in Tables 2.3 and 2.4, 91 % of the alleles were always assigned and 2.8 % were never assigned, regardless of whether pedigree information was used or not. Therefore, 6.2 % of the alleles might switch from not being assigned to being assigned or vice versa, depending on whether pedigree information is used or not and the value set for f_r. Furthermore, we observed that assignments of breed of origin obtained with f_r set at 0 or 20 % were consistent. Therefore, relaxing the maximum percentage of copies of the haplotype to be observed in another purebred population from 0 to 20 % did

appear to have resulted in extra assignments rather than rearrangement of assignments.

2.4.3 Animals with a low percentage of assignment of breed of origin

The percentage of assignment of breed of origin to alleles was high and constant across chromosomes. Two hundred and seven three-breed cross pigs had at least one chromosome for which less than 80 % of the alleles were assigned. It is difficult to characterize these animals, since 115 of these have only one or none of their parents genotyped. However across the whole data, 221 three-way crossbred animals also had only one or none of their parents genotyped, which means that 106 of them still achieved more than 80 % assignment for all chromosomes. Relatedness within these three-way crossbred animals does not seem to be the issue either. We found a maximum of 16 half- or full-sibs (in the scenario with common sire A) and 11 half- or full-sibs (in the scenario with common sire B), however, sires A and B also produced 13 and 15 other half- or full-sibs, respectively, with more than 80 % assignment for all chromosomes. A low assignment percentage was found for the whole genome for two three-breed cross pigs. A principle component analysis of the genotype data showed that these two pigs do not overlap with the three-breed cross population. Thus, the approach used was not able to assign an origin to most of their haplotypes. We suspect that this absence of overlap of these two three-breed cross pigs with the three-breed cross population may be due to erroneous pig identification, i.e. the first pig might have originated from the paternal S purebred line and the second pig from a cross with another line that was not included in this study. This absence of overlap with the correspondent population was also observed for some purebred line pigs, likely for the same reasons. Pigs with low assignment of breed of origin to alleles along the whole genome should be removed from the dataset because they do not add information about breed of origin of alleles when it is used in further analyses, and because the low assignment may indicate an error in the data. The assignment of breed of origin to alleles of other three-breed cross pigs in the dataset should not be affected by these apparently incorrectly labelled pigs, even if the incorrect assignment occurs for the purebred line pigs, i.e. using breed-of-origin assignment with a f_r of 20 %, we still expected that at least 80 % of the alleles from the other purebred line pigs would be assign the correct breed.

With the third principal component, we observed that pigs from the LR line were more variable compared to those from the other purebred lines (Fig. 2.5). This is

probably because the recent history of the LR pigs used in our study involves animals that originated from two populations. As a result, the three-breed cross pigs were also sub-divided into two sub-groups, which probably depended on which of these two sub-populations the LR grand-dam came from. This variation within the LR population was mainly captured by the third principal component but it explained only 1.94 % of the extra variation.

2.4.4 Phasing and haplotypes library

The first step to assign the breed of origin of alleles, was to phase the genotypes using the long-range phasing and haplotype library algorithm AlphaPhase1.1 (Hickey et al., 2011). Phasing using pedigree information was on average three times faster than phasing without pedigree information. For the starting analysis, which includes phasing of the purebred animals, and the first batch of the crossbred animals, the assignment of breed of origin can still be accurately obtained without pedigree information, but one has to account for the increased computational demand. AlphaPhase1.1 builds a library of all unique haplotypes that long-range phasing has found in the dataset. This library can then be used in subsequent analyses for phasing new crossbred animals that are added to the dataset and that may or may not have pedigree information, without the need to phase the reference population again. Hickey et al. (2010) tested this phasing strategy with simulations and obtained 81 to 94 % of correctly phased SNPs with low error rate (< 0.08 %). This phasing strategy can be applied for breed-of-origin assignment to speed up the assignment of alleles of new crossbred animals that are added to the dataset.

2.4.5 Application

Using crossbred performance for genetic predictions could be beneficial in breeding systems where production animals are crossbred, especially for traits with a low genetic correlation between purebred and crossbred performance. Genomic selection outperforms selection based on pedigree relationships and allows the use of crossbred performance information, even when pedigree information is not available. However, when using crossbred information for genomic prediction, we must take into account that effects of SNPs may be breed-specific because LD patterns between a SNP and a QTL may differ between breeds (Bastiaansen et al., 2014), and allele frequencies and allele substitution effects of QTL may also differ between breeds (Wientjes et al., 2015). To include these differences between breeds in a prediction model, we first need to determine the breed of origin of alleles in three-breed cross animals with high accuracy, as in this study, and then

use prediction models that estimate breed-specific SNP effects, as proposed by Ibáñẽz-Escriche et al. (2009) and Christensen et al. (2014). The benefit of this approach, training with crossbred data and using breed-specific SNP effects models, is that allele substitution effects of purebred alleles will be estimated against the genetic background that they will be expressed in. Thus, this approach can potentially incorporate the additive components of dominance and epistasis (Ibáñẽz-Escriche et al., 2009; Kinghorn et al., 2010). This could be used in combination with reciprocal recurrent selection (Wei and van der Steen, 1991) using phenotypes and genotypes of crossbred animals instead of only phenotypes (Kinghorn et al., 2010). Under some conditions (i.e., low SNP density, large crossbred training data size, and low breed relatedness), Ibáñẽz-Escriche et al. (2009) and Esfandyari et al. (2015), reported improved predictions using a model that accounts for the breed of origin of alleles compared to an additive or dominance model where SNP effects are assumed the same across breeds. In Ibáñẽz-Escriche et al. (2009) and Esfandyari et al. (2015), simulated data were used and breed of origin of alleles was assumed to be known a priori. With the results obtained in our study, the genomic model that accounts for breed of origin of alleles can be tested with real data. Since applications of genomic prediction require frequent re-estimation of SNP effects to maintain prediction accuracy, genomic prediction based on crossbred performance and breed-of-origin knowledge would also require repeated derivation of breed-specific SNP effects.

In addition to genomic prediction analyses, knowledge of breed of origin of alleles can also be used in genome-wide association studies (GWAS), accounting for the fact that the effect of causative mutations on phenotypes may depend on breed of origin. The approach can be similar to that using parental origin of sequence variants (Kong et al., 2009), in which genomic imprinting restricts the effect to the allele inherited from a parent of a specific sex; however, to be able to distinguish between parental origin and breed of origin, reciprocal crosses will be needed.

The genomic prediction model and GWAS that account for breed of origin can be also tested using haplotypes instead of single SNPs, which can increase prediction accuracies in genomic prediction (Calus et al., 2008), and increase power and precision in GWAS (Pryce et al., 2010). However, although the output of the breed-of-origin approach provided 18 haplotypes libraries, it will be still necessary to combine them and redefine the start and endpoints of the haplotypes so that they are suitable for these types of analyses.

2.5 Conclusion

Breed of origin of alleles of crossbred animals can be empirically derived without pedigree information. Pedigree information is, however, useful to reduce computation time and slightly improves assignment percentage. Around 94 % of the alleles of three-breed cross pigs were assigned a breed of origin. The results of this approach for assigning breed of origin to alleles allows the use of models that implement breed-specific effects of SNP alleles in genomic prediction, with the aim to improve selection of purebred animals for crossbred offspring performance. Breed-of-origin information also opens new possibilities to study associations between SNPs and production traits.

Chapter 3

Genomic evaluation for a three-way crossbreeding system considering breed-of-origin of alleles

Claudia A. Sevillano[1,2], Jeremie Vandenplas[1], John W.M. Bastiaansen[1], Rob Bergsma[2], Mario P.L. Calus[1]

[1] Wageningen University & Research Animal Breeding and Genomics, 6700AH, Wageningen, the Netherlands; [2] Topigs Norsvin Research Center, 6640AA, Beuningen, the Netherlands

Genetics Selection Evolution (2017) 49:75

Abstract

Genomic prediction of purebred animals for crossbred performance can be based on a model that estimates effects of single nucleotide polymorphisms (SNPs) in purebreds on crossbred performance. For crossbred performance, SNP effects might be breed-specific due to differences between breeds in allele frequencies and linkage disequilibrium patterns between SNPs and quantitative trait loci. Accurately tracing the breed of origin of alleles (BOA) in three-way crosses is possible with a recently developed procedure called BOA. A model that accounts for breed-specific SNP effects (BOA model), has never been tested empirically on a three-way crossbreeding scheme. Therefore, the objectives of this study were to evaluate the estimates of variance components and the predictive accuracy of the BOA model compared to models in which SNP effects for crossbred performance were assumed to be the same across breeds, using either breed-specific allele frequencies (G_A model) or allele frequencies averaged across breeds (G_B model). In this study, we used data from purebred and three-way crossbred pigs on average daily gain (ADG), back fat thickness (BF), and loin depth (LD). Estimates of variance components for crossbred performance from the BOA model were mostly similar to estimates from models G_A and G_B. Heritabilities for crossbred performance ranged from 0.24 to 0.46 between traits. Genetic correlations between purebred and crossbred performance (r_{pc}) across breeds ranged from 0.30 to 0.62 for ADG and from 0.53 to 0.74 for BF and LD. For ADG, prediction accuracies of the BOA model were higher than those of the G_A and G_B models, with significantly higher accuracies only for one maternal breed. For BF and LD, prediction accuracies of models G_A and G_B were higher than those of the BOA model, with no significant differences. Across all traits, models G_A and G_B yielded similar predictions. The BOA model yielded a higher prediction accuracy for ADG in one maternal breed, which had the lowest r_{pc} (0.30). Using the BOA model was especially relevant for traits with a low r_{pc}. In all other cases, the use of crossbred information in models G_A and G_B, does not jeopardize predictions and these models are more easily implemented than the BOA model.

Key words: origin of alleles, crossbred, genomic prediction, finisher, pig

3.1 Introduction

Genomic selection (GS) is more accurate than pedigree-based selection, and thus was developed for purebred (PB) populations of many farm species (Hayes et al., 2009; Forni et al., 2010; Jannink et al., 2010; Wolc et al., 2011). However, many production systems use crossbreeding schemes to produce crossbred (CB) individuals for commercial production. Crossbreeding in plants is common practice in many crops, such as maize. Crossbreeding in animals is common practice for pigs and poultry, and, in cattle, the use of crosses or composite breeds contributes largely to the beef and dairy industry. If selection is based on the performance measured on PB individuals, the rate of genetic change observed in CB individuals may be reduced because of differences in additive variance between PB and CB individuals, and because the genetic correlation between performance in PB and CB individuals (r_{pc}) is lower than 1 (Wei and Van der Steen, 1991; Brandt and Täubert, 1998). With r_{pc} values of 0.7 or lower, using only PB performance was predicted to yield considerably less genetic progress in CB performance compared to using performance of both PB and CB (Dekkers, 2007; Van Grevenhof and Van der Werf, 2015). In pigs, r_{pc} lower than 0.7 were reported for daily gain, daily feed intake, feed conversion ratio and residual feed intake (Lutaaya et al., 2001; Nakavisut et al., 2005; Knap and Wang, 2012), and also in poultry for egg number (Wei and Van der Werf, 1995), and in cattle for weight-related traits (Newman et al., 2002). In maize, the correlation between PB and CB performance for grain yield (GY) is lower than that for grain dry matter content (GDMC), and it was observed that models that do not include CB information failed to predict the performance of CB for GY but for GDMC yielded a high prediction accuracy (Schrag et al., 2009).

With GS, training with CB information is facilitated because GS eliminates the disadvantages of having to record pedigree data on CB individuals (Dekkers, 2007). Moreover, GS using CB information could benefit from models that estimate the effects on CB performance of markers that segregate within the parental breeds, as suggested by Dekkers (2007), Ibáñez-Escriche et al. (2009), Kinghorn et al. (2010) and Christensen et al. (2014; 2015) in the context of animal breeding, and by Schrag et al. (2009) in the context of hybrid performance in maize.

A commonly used GS model, known as genomic best linear unbiased prediction (GBLUP) (Meuwissen et al., 2001), replaces the pedigree-based relationship matrix by a genomic relationship matrix. The values in the genomic relationship matrix are a function of allele content and allele frequencies (VanRaden, 2008). Consequently,

the genomic relationship matrix is built under the assumption that all individuals belong to the same population, with the same average allele contents. Moreover, GBLUP implicitly assumes a single value for the linkage disequilibrium between a single nucleotide polymorphism (SNP) and a quantitative trait locus (QTL). When individuals originate from different populations, as in the crossbreeding context, these assumptions are violated because allele frequencies and the linkage disequilibrium patterns across the genome differ between breeds (De Roos et al., 2008; Makgahlela et al., 2013; Veroneze et al., 2014). Models that account for breed-specific allele frequencies were tested with simulated and real data and showed no improvement in prediction accuracies (Makgahlela et al., 2014; Moghaddar et al., 2014; Lourenco et al., 2016). Models that, in addition to including breed-specific allele frequencies, also account for breed-specific SNP effects did outperform models in which SNP effects were assumed to be the same across breeds. However, these results were only observed in simulation studies under some conditions (i.e., low SNP density, large training data size, and low breed relatedness) and where breed of origin of alleles was assumed to be known without error (Ibánẽz-Escriche et al., 2009; Esfandyari et al., 2015). With real data from a two-way crossbreeding scheme, Xiang et al. (2016) and Lopes et al. (2017) reached different conclusions. When using a model that accounted for breed-specific SNP effects compared to a model in which SNP effects were assumed to be the same across breeds, Xiang et al. (2016) found improved prediction accuracies and reduced bias of prediction, whereas Lopes et al. (2017) found similar prediction accuracies between the two models. The benefit of a two-way CB is that tracing the breed of origin of alleles is relatively straightforward. However, many crossbreeding schemes are based on a three-way cross, for which tracing the breed of origin of alleles is considerably more complicated (Vandenplas et al., 2016). Recently we have developed a procedure that enables breed-of-origin assignment (BOA) of alleles in three-way CB animals (Sevillano et al., 2016). BOA allows empirical testing of the model that accounts for breed-specific SNP effects in real data. Therefore, the objectives of this study were to evaluate the estimates of variance components and the accuracy of a model that accounts for breed-specific SNP effects using information from both PB and three-way CB pigs for average daily gain (ADG), back fat thickness (BF), and loin depth (LD).

3.2 Methods

3.2.1 Data

The pig data consisted of three PB populations: Synthetic boar (S), Landrace (LR), and Large White (LW), and a three-way CB population: (S (LR x LW) or S (LW x LR)),

produced by crossing the above-mentioned PB populations. The numbers of available genotypes and phenotypes per trait and per population are in Table 3.1. All pigs were genotyped using one of the three following SNP panels: Illumina PorcineSNP60.v2 BeadChip (60K.v2), Illumina PorcineSNP60 BeadChip (60K), or Illumina PorcineSNP10 BeadChip (10K). Pigs genotyped with the 60K or 10K chips were imputed to the 60K.v2 panel using FImpute Version 2.2 software (Sargolzaei et al., 2014). SNP quality control and imputation were applied on the same dataset in a previous study (Sevillano et al., 2016), in which more details are provided. The final SNP set for subsequent analyses consisted of 52,164 SNPs. Phenotypes for ADG (g/d), BF (mm), and LD (mm), were measured for most of the PB and CB pigs. ADG for PB was calculated as the difference of on-test body weight measured on average at 60 days of age and off-test body weight measured on average at 173 days of age. ADG for CB was calculated as the difference of on-test body weight measured on average at 70 days of age and body weight at the end of the finishing period, which was on average 120 kg. BF and LD for PB were measured on average at 173 days of age using an ultrasound instrument, while BF and LD for CB were measured on the carcass after slaughter using a probe, named "capteur gras maigre" (CGM; Sydel, France). For all phenotyped pigs, four generations of pedigree information were included.

Table 3.1 Number of genotypes and phenotypes available for each trait and population

Population	Genotypes	ADG	BF	LD
S	2733	2575	2616	2595
LR	4148	2333	3605	2386
LW	7103	5294	6769	5469
CB	1706	1675	1676	1681
Total	15,690	11,877	14,666	12,131

S = Synthetic boar, LR = Landrace, LW = Large White, and CB = three-way crossbred pigs. ADG = average daily gain, BF = back fat thickness, and LD = loin depth.

3.2.2 Analyses

3.2.2.1 GBLUP model with breed-specific partial relationship matrices (BOA model)

To account for the breed-specific effect of SNPs, the following 4-trait animal model with three breed-specific partial relationship matrices ($\mathbf{G}^{(\mathbf{S})}$, $\mathbf{G}^{(\mathbf{LR})}$ and $\mathbf{G}^{(\mathbf{LW})}$)

$$\mathbf{y_S} = \mathbf{X_S b_S} + \mathbf{W_S u_S} + \mathbf{Z_S a_S} + \mathbf{e_S},$$

$$\mathbf{y_{LR}} = \mathbf{X_{LR} b_{LR}} + \mathbf{W_{LR} u_{LR}} + \mathbf{Z_{LR} a_{LR}} + \mathbf{e_{LR}},$$

$$\mathbf{y_{LW}} = \mathbf{X_{LW} b_{LW}} + \mathbf{W_{LW} u_{LW}} + \mathbf{Z_{LW} a_{LW}} + \mathbf{e_{LW}},$$

$$\mathbf{y_{CB}} = \mathbf{X_{CB} b_{CB}} + \mathbf{W_{CB} u_{CB}} + \mathbf{Z_{CB} g_{CB}^{(S)}} + \mathbf{Z_{CB} g_{CB}^{(LR)}} + \mathbf{Z_{CB} g_{CB}^{(LW)}} + \mathbf{e_{CB}},$$

where $\mathbf{y_S}$, $\mathbf{y_{LR}}$, $\mathbf{y_{LW}}$, and $\mathbf{y_{CB}}$ are the vectors of the phenotypes for S, LR, LW, and CB pigs, respectively; $\mathbf{b_S}$, $\mathbf{b_{LR}}$, $\mathbf{b_{LW}}$, $\mathbf{b_{CB}}$ represent the vectors of fixed effects (listed in Table 3.2) and $\mathbf{X_S}$, $\mathbf{X_{LR}}$, $\mathbf{X_{LW}}$, $\mathbf{X_{CB}}$ are the respective incidence matrices relating pig records to fixed effects; $\mathbf{u_S}$, $\mathbf{u_{LR}}$, $\mathbf{u_{LW}}$, $\mathbf{u_{CB}}$ represent the vectors of random common litter effects, and $\mathbf{W_S}$, $\mathbf{W_{LR}}$, $\mathbf{W_{LW}}$, $\mathbf{W_{CB}}$ are the respective incidence matrices relating pig records to litter effects; $\mathbf{a_S}$, $\mathbf{a_{LR}}$, $\mathbf{a_{LW}}$, are the vectors of additive genetic effects in PB, $\mathbf{g_{CB}^{(S)}}$, $\mathbf{g_{CB}^{(LR)}}$, $\mathbf{g_{CB}^{(LW)}}$ are the vectors of the additive genetic effect of PB gametes in CB, and $\mathbf{Z_S}$, $\mathbf{Z_{LR}}$, $\mathbf{Z_{LW}}$, $\mathbf{Z_{CB}}$ are the respective incidence matrices. Because each model was run for each trait and only pigs with phenotypes were included, $\mathbf{Z}$ incidence matrices relating pig records to additive genetic effects were identity matrices when variance components were estimated. Finally, $\mathbf{e_S}$, $\mathbf{e_{LR}}$, $\mathbf{e_{LW}}$, $\mathbf{e_{CB}}$ represent the vectors of random residual effects. The variance-covariance of the common litter effect and residual effect were:

$$\text{Var}\begin{bmatrix}\mathbf{u_S}\\ \mathbf{u_{LR}}\\ \mathbf{u_{LW}}\\ \mathbf{u_{CB}}\end{bmatrix} = \begin{bmatrix} \mathrm{I}_S\sigma^2_{u_S} & 0 & 0 & 0\\ 0 & \mathrm{I}_{LR}\sigma^2_{u_{LR}} & 0 & 0\\ 0 & 0 & \mathrm{I}_{LW}\sigma^2_{u_{LW}} & 0\\ 0 & 0 & 0 & \mathrm{I}_{CB}\sigma^2_{u_{CB}}\end{bmatrix},$$

$$\text{and Var}\begin{bmatrix}\mathbf{e_S}\\ \mathbf{e_{LR}}\\ \mathbf{e_{LW}}\\ \mathbf{e_{CB}}\end{bmatrix} = \begin{bmatrix} \mathrm{I}_S\sigma^2_{e_S} & 0 & 0 & 0\\ 0 & \mathrm{I}_{LR}\sigma^2_{e_{LR}} & 0 & 0\\ 0 & 0 & \mathrm{I}_{LW}\sigma^2_{e_{LW}} & 0\\ 0 & 0 & 0 & \mathrm{I}_{CB}\sigma^2_{e_{CB}}\end{bmatrix}.$$

The variance-covariance of additive genetic effect for breed S origin was:

$$\text{Var}\begin{bmatrix}\mathbf{a_S}\\ \mathbf{a_{CB}^{(S)}}\\ \mathbf{g_S}\\ \mathbf{g_{CB}^{(S)}}\end{bmatrix} = \begin{bmatrix}\sigma^2_{a_S} & \sigma_{a_S,g_S}\\ \sigma_{g_S,a_S} & \sigma^2_{g_S}\end{bmatrix} \otimes \mathbf{G^{(S)}} = \begin{bmatrix}\sigma^2_{a_S} & \sigma_{a_S,g_S}\\ \sigma_{g_S,a_S} & \sigma^2_{g_S}\end{bmatrix} \otimes \begin{bmatrix}\mathbf{G_{S,S}} & \mathbf{G_{S,CB}^{(S)}}\\ \mathbf{G_{CB,S}^{(S)}} & \mathbf{G_{CB,CB}^{(S)}}\end{bmatrix},$$

where S pigs have additive effects (i.e. breeding values), $\mathbf{a_S}$ for PB performance and $\mathbf{a_{CB}^{(S)}}$ for CB performance. The CB pigs have additive effects from the breed S gametes, $\mathbf{g_{CB}^{(S)}}$ for CB performance and $\mathbf{g_S}$ for PB performance. This last effect, $\mathbf{g_S}$, is an artificial random vector that is added to be able to define the variance-covariance of additive genetic effects with the above Kronecker product, but does not have practical relevance. The matrix $\mathbf{G^{(S)}}$ is a breed-specific partial relationships matrix for breed S which contains four blocks, one for within S pigs ($\mathbf{G_{S,S}}$), two for S with CB pigs ($\mathbf{G_{S,CB}^{(S)}}$ and $\mathbf{G_{CB,S}^{(S)}}$), and one for within CB pigs ($\mathbf{G_{CB,CB}^{(S)}}$).

The variance-covariance structures for the origin of breeds LR and LW are defined similarly, and the three variance-covariance structures are assumed independent, i.e. no covariances are considered between S, LR, and LW effects (Christensen et al., 2015). There are six genetic variance components, two for each breed of origin, and three covariance components, one for each breed of origin. To construct the three breed-specific partial relationship matrices, $\mathbf{G^{(S)}}$, $\mathbf{G^{(LR)}}$, and $\mathbf{G^{(LW)}}$, we used the breed of origin of phased alleles in CB pigs. Then, the breed-specific partial relationship submatrices are defined as, e.g. breed S origin:

$$\mathbf{G_{S,S}} = (\mathbf{M^S} - 2\mathbf{1p^{S\prime}})\mathbf{D^S}(\mathbf{M^S} - 2\mathbf{1p^{S\prime}})'/N,$$

$$\mathbf{G_{S,CB}} = (\mathbf{M^S} - 2\mathbf{1p^{S\prime}})\mathbf{D^S}\left(\mathbf{M^{CB(S)}} - \mathbf{1p^{S\prime}}\right)'/N,$$

$$\mathbf{G_{CB,CB}} = (\mathbf{M^{CB}} - \mathbf{1p^{S\prime}})\mathbf{D^S}\left(\mathbf{M^{CB(S)}} - \mathbf{1p^{S\prime}}\right)'/N,$$

where $\mathbf{M^S}$ is a matrix containing breed-specific allele content for breed S pigs (coded as 0, 1, or 2), $\mathbf{M^{CB(S)}}$ is a matrix containing breed S allele content for CB pigs (coded as 0, or 1), alleles that were not assigned a breed of origin were set to missing, $\mathbf{p^S}$ is the vector of breed S specific frequencies of the counted allele (p_j^S). p_j^S was calculated across S and CB pigs by counting the occurrences alleles originating from the S breed and coded as 1, across the S breed and in CB, divided by the total number of S alleles in the S breed and CB on locus *j*. $\mathbf{D^S}$ is diagonal with $D_{jj}^S = \frac{1}{2p_j^S(1-p_j^S)}$. N is the number of SNPs.

The breed-specific partial relationship submatrices $\mathbf{G^{(LR)}}$ and $\mathbf{G^{(LW)}}$ are defined similarly to $\mathbf{G^{(S)}}$. However, the entries of the $\mathbf{M^{CB}}$ matrix containing the breed-specific allele content for CB pigs are set to a missing value if the origin of the allele corresponds to the other maternal line, and effectively does not contribute to the breed-specific partial relationship matrix.

Table 3.2 Fixed effects used in the GBLUP models for average daily gain (ADG), back fat thickness (BF), and loin depth (LD), for purebred (PB) (i.e. S, LR, LW) and three-way crossbred (CB) pigs

Trait	Population	Fixed effects
ADG	PB	$\text{farm} * \text{breed} * \text{sex} + b_a \times \text{birth weight}$
	CB	$\text{trial} + \text{farm} * \text{sex} + b_a \times \text{birth weight}$
BF, LD	PB	$\text{farm} * \text{breed} * \text{sex} + b_b \times \text{off\_test BW}$
	CB	$\text{trial} + \text{farm} * \text{sex} + b_c \times \text{hot carcass weight}$

b_a, b_b, b_c, are regression coefficients for birth weight, off-test BW, and hot carcass weight, respectively.

3.2.2.2 Assigning breed of origin to alleles in crossbreds

To infer the breed of origin of the alleles in CB pigs, we used the BOA approach that was developed by Vandenplas et al. (2016). It consists of three steps: (1) phasing the haplotypes of both PB and CB pigs with AlphaPhase1.1 software (Hickey et al., 2011), (2) determining the unique haplotypes among the PB, and (3) assigning the breed of origin for each allele carried on the haplotypes of CB. This approach was applied to the same dataset in a previous study (Sevillano et al., 2016). On average, 95.2% of the alleles of the three-way CB pigs were assigned a breed of origin. These alleles with their assigned breed of origin were used to build the breed-specific partial relationship matrices. Alleles that were not assigned a breed of origin were set to missing, and effectively did not contribute to any of the breed-specific partial relationship matrices.

3.2.2.3 GBLUP model with the genomic relationship matrix

For comparison to the BOA model, the following 4-trait animal model was fitted (G model):

$$\mathbf{y_S} = \mathbf{X_S b_S} + \mathbf{W_S u_S} + \mathbf{Z_S a_S} + \mathbf{e_S},$$

$$\mathbf{y_{LR}} = \mathbf{X_{LR} b_{LR}} + \mathbf{W_{LR} u_{LR}} + \mathbf{Z_{LR} a_{LR}} + \mathbf{e_{LR}},$$

$$\mathbf{y_{LW}} = \mathbf{X_{LW} b_{LW}} + \mathbf{W_{LW} u_{LW}} + \mathbf{Z_{LW} a_{LW}} + \mathbf{e_{LW}},$$

$$\mathbf{y_{CB}} = \mathbf{X_{CB} b_{CB}} + \mathbf{W_{CB} u_{CB}} + \mathbf{Z_{CB} a_{CB}} + \mathbf{e_{CB}},$$

where vectors and matrices are defined as in the BOA model, with the only difference being that the additive genetic effect in CB pigs was defined only by one vector, $\mathbf{a_{CB}}$. Therefore, the variance-covariance matrix of genetic effects was:

$$\text{Var}\begin{bmatrix}\mathbf{a_S}\\ \mathbf{a_{LR}}\\ \mathbf{a_{LW}}\\ \mathbf{a_{CB}}\end{bmatrix} = \begin{bmatrix}\sigma^2_{a_S} & \sigma_{a_S,a_{LR}} & \sigma_{a_S,a_{LW}} & \sigma_{a_S,a_{CB}}\\ \sigma_{a_S,a_{LR}} & \sigma^2_{a_{LR}} & \sigma_{a_{LR},a_{LW}} & \sigma_{a_{LR},a_{CB}}\\ \sigma_{a_S,a_{LW}} & \sigma_{a_{LR},a_{LW}} & \sigma^2_{a_{LW}} & \sigma_{a_{LW},a_{CB}}\\ \sigma_{a_S,a_{CB}} & \sigma_{a_{LR},a_{CB}} & \sigma_{a_{LW},a_{CB}} & \sigma^2_{a_{CB}}\end{bmatrix} \otimes \mathbf{G}.$$

This model was implemented using two different genomic relationship matrices ($\mathbf{G}$) as explained in the next sections.

3.2.2.4 Genomic relationship matrix using allele frequencies across all genotyped pigs (G_A matrix)

The $\mathbf{G_A}$ matrix was constructed using the second method in VanRaden (2008):

$$\mathbf{G_A} = (\mathbf{M} - 2\mathbf{1p}')\mathbf{D}(\mathbf{M} - 2\mathbf{1p}')'/N,$$

where $\mathbf{M}$ is a matrix containing SNP genotypes for each pig (coded as 0, 1, or 2), $\boldsymbol{p}$ is the vector of the frequencies of the counted allele (p_j), calculated across the genotyped population, $\mathbf{D}$ is diagonal with $D_{jj} = \frac{1}{2p_j(1-p_j)}$, and N is the number of SNPs.

3.2.2.5 Genomic relationship matrix using breed-specific allele frequencies (G_B matrix)

To account for population structure, we also used a genomic relationship matrix based on genotypes centered and scaled by breed-specific allele frequencies ($\mathbf{G_B}$):

$$\mathbf{G_B} = (\mathbf{M} - 2\mathbf{1p_B}')\mathbf{D^B}(\mathbf{M} - 2\mathbf{1p_B}')'/N,$$

where each $\mathbf{p_B}$ is the vector of the frequencies of the counted allele at locus j (p_{Bj}). p_{Bj} was specific for each breed B (i.e. S, LR, and LW), and the weighted contribution of each breed for the CB. The weights considered for the CB were 0.5 for S, 0.25 for LR and 0.25 for LW. $\mathbf{D^B}$ is diagonal with $D^B_{jj} = \frac{1}{2p_{Bj}(1-p_{Bj})}$.

3.2.3 Estimation of variance components and BLUP

Implementation of the aforementioned GBLUP models required estimates for all variance components involved. Variance components were estimated for each of the three models using the ASReml software (Gilmour et al., 2009). Instead of one 4-trait multivariate model, three bivariate models were fitted to overcome workspace memory limitation of the software. Each analysis included PB of one of the three breeds and all CB. As a consequence, genetic co-variances between

breeds were not estimated. For the BOA model, these genetic co-variances are not considered and thus are effectively equal to 0. For the other two models, we also assumed that these co-variances were not significant, and therefore, we set them to 0 in the subsequent BLUP analyses. Variance components of the bivariate models were combined to obtain the full variance-covariance matrices for the 4-trait model. The variance-covariance matrices were combined by averaging the three CB variance components estimated in each of the bivariate models. If necessary, the combined variance-covariance matrices were bended to make them positive definite (Jorjani et al., 2003). Bending changed the variance-covariance components on average by 7.5% (0.3 to 18.5%). BLUP for the three models were obtained using the MiXBLUP software (Ten Napel et al., 2016).

3.2.4 Cross-validation

The accuracy of EBV of PB pigs for CB performance from the three models was evaluated as the average accuracy obtained from 4-fold cross-validation. Because of different degrees of relationship between PB and CB, genotyped S, LR, or LW pigs were first divided into four mutually exclusive clusters, using the K-means clustering method applied to a dissimilarity matrix computed from elements of the $\mathbf{G_A}$ matrix (Saatchi et al., 2011). Then, each CB pig was assigned to the PB cluster with the closest relationship based on the $\mathbf{G_A}$ matrix. For the maternal breed LW, the CB pigs were not very evenly distributed across the clusters, with one cluster including most of the CB. Therefore, for this breed, the cluster with the largest number of CB pigs was randomly split into four groups and each of those groups was joined with one of the other clusters.

In each training analysis, the data excluded PB and CB pigs from one fold to train on the remaining three folds to predict EBV for CB performance of the excluded PB pigs (validation set). This resulted in every PB pig having EBV for CB performance that were obtained without using performance of the most closely-related CB pigs for training. Thus, the information coming from the most closely-related CB pigs could be used for validation. The number of pigs in the validation and training sets for each of the folds of the cross-validation and for each trait are in Tables 3.3, 3.4 and 3.5 for S, LR, and LW, respectively.

3.2.4.1 Validation set

The PB pigs cannot have an own performance for CB performance, and also in our data, they do not have large offspring groups, which would allow to compute a phenotype as average offspring performance. Therefore, we calculated

deregressed proofs (DRP) for PB pigs within the validation sets to validate the predictions of our models. For this, first we obtained EBV from the G model with a pedigree-based relationship matrix. This resulted in an EBV for CB performance for each PB pig. The EBV were estimated based on performance of the CB pigs assigned to each of the validation folds (Tables 3.3, 3.4, and 3.5 for S, LR, and LW, respectively). Phenotype information was also available for an additional 501 CB pigs (CB-extra) that were not genotyped. These records were used in each of the four validation folds (Tables 3.3, 3.4, and 3.5 for S, LR, and LW, respectively). Within each validation fold, the EBV of PB pigs for CB performance were then deregressed according to Calus et al. (2016). The deregression involved removal of all effects of relatives in the same validation set, and correction for regression to the mean, to obtain a more accurate estimate of the expected phenotype. In addition, a weighting factor (*w*) was estimated for each DRP value based on the reliability of the calculated DRP. These w are the effective record contributions (Přibyl et al., 2013), and reflect the amount of information in the DRP contributed by the animal itself, correcting for any information of the relatives that contributed to its EBV before deregression.

Table 3.3 Cross-validation strategy for crossbred performance of Synthetic boar (S)

Folds	**Training**		**Validation**		
	S	**CB**	**S**	**CB**	**CB-extra***
Average daily gain					
1	2115	1535	460	140	199
2	2119	1341	456	334	268
3	1895	605	680	1070	297
4	1596	1544	979	131	145
Back fat thickness					
1	2132	1536	484	140	188
2	2144	1344	472	332	246
3	1932	604	684	1072	289
4	1640	1544	976	132	145
Loin Depth					
1	2128	1541	467	140	200
2	2132	1348	463	333	272
3	1921	605	674	1076	299
4	1604	1549	991	132	145

Numbers of individuals for Synthetic boar (S), three-way crossbred (CB) and extra three-way crossbred pigs (CB-extra) in the training and validation sets per trait.

*Three-way crossbred pigs with only phenotypic information, no genotype.

Table 3.4 Cross-validation strategy for crossbred performance of Landrace (LR)

Folds	Training		Validation		
	LR	CB	LR	CB	CB-extra*
Average daily gain					
1	1584	1564	748	111	456
2	1825	1523	507	152	465
3	1762	1531	570	144	456
4	1825	407	507	1268	471
Back fat thickness					
1	2829	1565	775	111	463
2	2492	1523	1112	153	472
3	3002	1532	602	144	463
4	2489	408	1115	1268	478
Loin Depth					
1	1631	1570	754	111	463
2	1891	1528	494	153	472
3	1823	1537	562	144	463
4	1810	408	575	1273	478

Numbers of individuals for Landrace (LR), three-way crossbred (CB), and extra three-way crossbred pigs (CB-extra) in the training and validation sets per trait.

*Three-way crossbred pigs with only phenotypic information, no genotype.

Table 3.5 Cross-validation strategy for crossbred performance of Large White (LW)

Folds	Training		Validation		
	LR	CB	LR	CB	CB-extra*
Average daily gain					
1	3628	1193	1666	482	468
2	3612	1269	1682	406	468
3	4008	1111	1286	564	468
4	4634	1452	660	223	468
Back fat thickness					
1	4870	1191	1899	485	475
2	4954	1271	1815	405	475
3	4381	1113	2388	563	475
4	6102	1453	667	223	475
Loin Depth					
1	3759	1196	1710	485	475
2	3678	1275	1791	406	475
3	4162	1114	1307	567	475
4	4808	1458	661	223	475

Numbers of individuals for Large White (LW), three-way crossbred (CB), and extra three-way crossbred pigs (CB-extra) in the training and validation sets per trait.
*Three-way crossbred pigs with only phenotypic information, no genotype.

3.2.4.2 Predictive ability

Accuracies of the BOA and G models were calculated as the weighted correlation between the DRP and the EBV of PB pigs for CB performance, where the weighting factor w was used to account for differences in the amount of available information on relatives to estimate DRP. The standard error (SE) of the correlations were approximated as $(1-\mathrm{r}^2)/\sqrt{\mathrm{N}}$, were r is the accuracy of the model, and N is the number of validation animals (Stuart and Ord, 1994).

3.3 Results

3.3.1 Genotyped population and relationship matrices

The three breeds, S, LR, and LW, were clearly different populations as shown in Fig. 3.1 based on the first two principal components of the $\mathbf{G_A}$ matrix. The CB population appeared intermediate among the PB populations. The divergence among the three populations estimated with Weir and Cockerham's F_{ST} (Weir and Cockerham, 1984), were equal to 0.17 between S and LR, 0.12 between S and LW, and 0.14 between LW and LR, which indicated that they are distantly-related breeds.

The relationships between breeds, calculated with the $\mathbf{G_A}$ matrix were mainly negative (Table 3.6), with average relationships between breeds ranging from -0.13 to -0.07. When using the $\mathbf{G_B}$ matrix, the average relationships between all breeds are zero by definition. When using breed-specific partial relationship matrices ($\mathbf{G^{(S)}}$, $\mathbf{G^{(LR)}}$ and $\mathbf{G^{(LW)}}$), only the relationships based on common alleles originating from the same breed were considered and, consequently no relationships were estimated between breeds. For CB pigs, the diagonal elements of the $\mathbf{G_A}$ and $\mathbf{G_B}$ matrices had an average of 0.96 and 0.94, respectively. For the $\mathbf{G^{(S)}}$, $\mathbf{G^{(LR)}}$ and $\mathbf{G^{(LW)}}$ matrices, as they are partial relationship matrices, the diagonal elements for CB pigs had averages of 0.49, 0.23, and 0.23 for $\mathbf{G^{(S)}}$, $\mathbf{G^{(LR)}}$ and $\mathbf{G^{(LW)}}$, respectively. These averages are close to the expected values, i.e. 0.50 for the S breed and 0.25 for the LR and LW breeds.

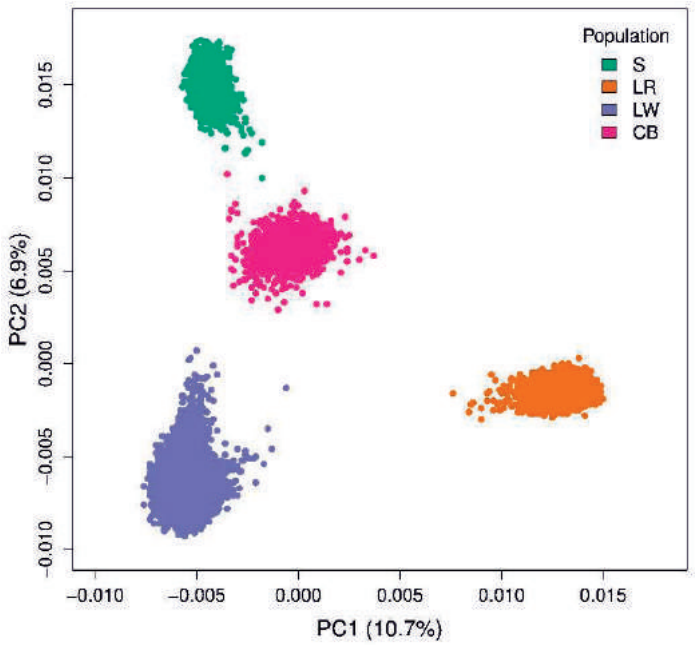

Figure 3.1 The two first principal components (PC) from the genomic relationship matrix between the different populations. Synthetic boar (S), Landrace (LR), Large White (LW), and three-way crossbred (CB) pigs. Each circle (o) represents a pig.

Table 3.6 Descriptive statistics for relationship between populations based on different genomic relationship matrices

Relationship between	Matrix[a]	Mean	Median	Min	Max	SD
S-LR	$\mathbf{G_A}$	-0.13	-0.13	-0.22	0.00	0.02
	$\mathbf{G_B}$	0.00	0.00	-0.09	0.09	0.02
S-LW	$\mathbf{G_A}$	-0.07	-0.07	-0.18	0.12	0.02
	$\mathbf{G_B}$	0.00	0.00	-0.11	0.11	0.02
LR-LW	$\mathbf{G_A}$	-0.13	-0.13	-0.23	0.16	0.02
	$\mathbf{G_B}$	0.00	0.00	-0.13	0.23	0.02
CB (diagonal)	$\mathbf{G}^{(S)}$	0.49	0.49	0.40	0.80	0.04
	$\mathbf{G}^{(LR)}$	0.23	0.23	0.02	0.40	0.04
	$\mathbf{G}^{(LW)}$	0.23	0.23	0.07	0.39	0.04
	$\mathbf{G_A}$	0.96	0.95	0.88	1.07	0.03
	$\mathbf{G_B}$	0.94	0.93	0.86	1.08	0.03

[a] $\mathbf{G}^{(S)}$ = partial relationship matrix for breed Synthetic boar (S); $\mathbf{G}^{(LR)}$ = partial relationship matrix for breed Landrace (LR); $\mathbf{G}^{(LW)}$ = partial relationship matrix for breed Large White (LW); $\mathbf{G_A}$ = genomic relationship matrix by allele frequencies obtained across the genotyped population; $\mathbf{G_B}$ = genomic relationship matrix by breed-specific allele frequencies.

3.3.2 Variance components, heritabilities, and genetic correlations

Estimated variance components for ADG, BF, and LD using the BOA model with the $\mathbf{G}^{(S)}$, $\mathbf{G}^{(LR)}$ and $\mathbf{G}^{(LW)}$ matrices, the G model with the $\mathbf{G_A}$ matrix ($\mathrm{G_A}$ model), and the G model with $\mathbf{G_B}$ matrix ($\mathrm{G_B}$ model) are in Table 3.7. The standard errors of the estimated variance components in Table 3.7 are provided in Table S3.1 [see Supplementary material, Additional file S3.1.]. Regardless of the model and trait, the PB additive genetic variance estimated for the maternal breeds, i.e. LR and LW, were very similar. For the maternal breeds, CB additive genetic variance was larger than PB additive genetic variance for all traits. For the paternal breed, the opposite was observed, i.e. CB additive genetic variance was smaller than PB additive genetic variance, for all traits except BF. Estimates of CB heritability tended to be higher than estimates of PB heritability for all traits except LD.

A comparison between models showed that PB and CB additive genetic variances for the maternal breeds were similar between the $\mathrm{G_A}$ and $\mathrm{G_B}$ models. For the paternal breed S, compared to the $\mathrm{G_B}$ model, the $\mathrm{G_A}$ model estimated a larger PB additive genetic variance, and smaller CB additive genetic variance. Estimated PB additive genetic variances with the BOA model were similar to those obtained with

the G_A or G_B models and the estimated CB additive genetic variances with the BOA model, on average across the three breeds, were larger than those obtained with the G_A or G_B models. The estimates of PB and CB heritability were similar across models, while estimates obtained with the BOA model tended to be slightly lower and those with the G_B model tended to be slightly higher than with the G_A model. The genetic correlations for traits between PB and CB pigs estimated with the BOA model were generally similar to those of the G_A and G_B models, except for the genetic correlation between LR and CB pigs for ADG that was much higher than that estimated with the G_A and G_B models. The genetic correlations between PB and CB pigs estimated with the G_A or G_B models were similar. In general, the SE of PB additive genetic variances and heritabilities were similar across models, although the SE of the three CB additive genetic variances estimated with the BOA model were much larger than the SE of the single CB additive genetic variance estimated with the G_A or G_B models. The SE of the estimated genetic correlations were relatively large, ranging from 0.10 to 0.29, across all models and traits.

For the BOA model, the CB variance for litter effect was about three times larger than that obtained with the G_A or G_B models. Estimates of the CB residual variance were also slightly larger when using the BOA model compared to the G_A and G_B models. Estimates of PB variance for litter and residual effects by the G_A and G_B models were similar among breeds. Estimates of CB variance for litter and residual effects by the G_A and G_B models were similar among the maternal breeds, while for breed S, the CB variance for litter and residual effects was lower with the G_A model than with the G_B model. In summary, estimated variance components were mostly similar across models, apart from the CB litter variance that was considerably larger with the BOA model compared to the other two models.

3.3.3 Predicting breeding values of PB pigs for CB performance with different models

For each breed S, LR, and LW, four validation groups were formed to perform the 4-fold cross-validation. Figure 3.2 represents the first two principal components from the $\mathbf{G_A}$ matrix and shows that the grouping for the cross-validation was done correctly. The first two principal components explained 6.3% of the variability among S pigs, 8.8% among LR pigs and 4.65% among LW pigs.

Table 3.7 Additive genetic variance (σ_a^2), litter variance (σ_u^2), residual variance (σ_e^2), and heritabilities for each breed for PB and CB performance, and genetic correlation between purebred and CB pigs (r_{PC}), estimated for each trait using the BOA[a], G_A[b], anG_B[c] models

Model	Breed	$\sigma^2_{a_{PB}}$	$\sigma^2_{u_{PB}}$	$\sigma^2_{e_{PB}}$	h^2_{PB}	$\sigma^2_{a_{CB}}$	$\sigma^{2\ *}_{u_{CB}}$	$\sigma^{2\ *}_{e_{CB}}$	h^2_{CB}	r_{pc}
Average daily gain										
BOA	S	2699	2925	6124	0.23	2316	853	4192	0.34**	0.50
	LR	2165	2291	3778	0.26	3566				0.62
	LW	2123	1595	4602	0.26	2258				0.57
G_A	S	3386	2850	6068	0.28	2053*	258	3576	0.35	0.52
	LR	2461	2282	3718	0.29					0.31
	LW	2336	1563	4595	0.28					0.61
G_B	S	2775	2846	6082	0.24	2261*	262	3592	0.37	0.52
	LR	2248	2287	3703	0.27					0.30
	LW	2154	1640	4568	0.26					0.59
Back fat thickness										
BOA	S	0.82	0.55	1.27	0.31	1.90	0.88	3.96	0.38**	0.74
	LR	1.09	0.60	1.73	0.32	3.74				0.67
	LW	1.33	0.86	1.67	0.34	4.16				0.58
G_A	S	1.18	0.55	1.26	0.40	2.18*	0.33	3.32	0.37	0.73
	LR	1.38	0.59	1.71	0.38					0.72
	LW	1.57	0.85	1.64	0.39					0.65
G_B	S	0.98	0.54	1.26	0.35	2.40*	0.34	3.34	0.39	0.69
	LR	1.26	0.59	1.70	0.35					0.70
	LW	1.44	0.85	1.64	0.37					0.62
Loin Depth										
BOA	S	10.59	6.00	8.43	0.42	11.59	3.20	31.45	0.24**	0.53
	LR	5.72	3.00	6.65	0.37	7.23				0.58
	LW	6.04	3.55	6.93	0.37	12.86				0.53
G_A	S	12.78	5.93	8.41	0.47	9.05*	0.11	28.89	0.24	0.57
	LR	6.58	2.98	6.60	0.41					0.57
	LW	6.82	3.56	6.89	0.40					0.68
G_B	S	10.58	5.87	8.33	0.43	10.00*	0.05	28.89	0.26	0.55
	LR	5.82	2.97	6.57	0.38					0.56
	LW	6.09	3.55	6.86	0.37					0.62

S = Synthetic boar, LR = Landrace, LW = Large White, and CB = three-way crossbred pigs

[a]BOA model, model with breed-specific relationship matrices

[b]G_A model, model with genomic relationship matrix by allele frequencies obtained across the genotyped population

[c]G_B model, model with genomic relationship matrix by breed-specific allele frequencies

*Average from the three bivariate models

**$(0.5\sigma^2_{a_S} + 0.25\sigma^2_{a_{LR}} + 0.25\sigma^2_{a_{LW}})/(0.5\sigma^2_{a_S} + 0.25\sigma^2_{a_{LR}} + 0.25\sigma^2_{a_{LW}} + \sigma^2_{u_{CB}}{}^* + \sigma^2_{e_{CB}}{}^*)$

Accuracies of the three models for the estimated breeding values of S pigs for CB performance are in Table 3.8. For ADG, the BOA model yielded slightly better accuracies than the G_A and G_B models. The opposite was observed for BF and LD, where the G_A and G_B models yielded slightly better accuracies than the BOA model. Accuracies of the three models for the estimated breeding values of LR pigs for CB performance are in Table 3.9. For ADG, the BOA model yielded higher accuracies than the G_A and G_B models. For BF and LD, there was no difference in accuracies between the three models. Accuracies of the three models for the estimated breeding values of LW pigs for CB performance are in Table 3.10. The trait ADG is not included, because the reliabilities of the EBV of LW pigs within the validation groups for CB performance for this trait were too low to be used for proper validation. Similar to the results for the LR breed, there was no difference in accuracies between the three models for the traits BF and LD. In general, accuracies from models G_A and G_B were similar.

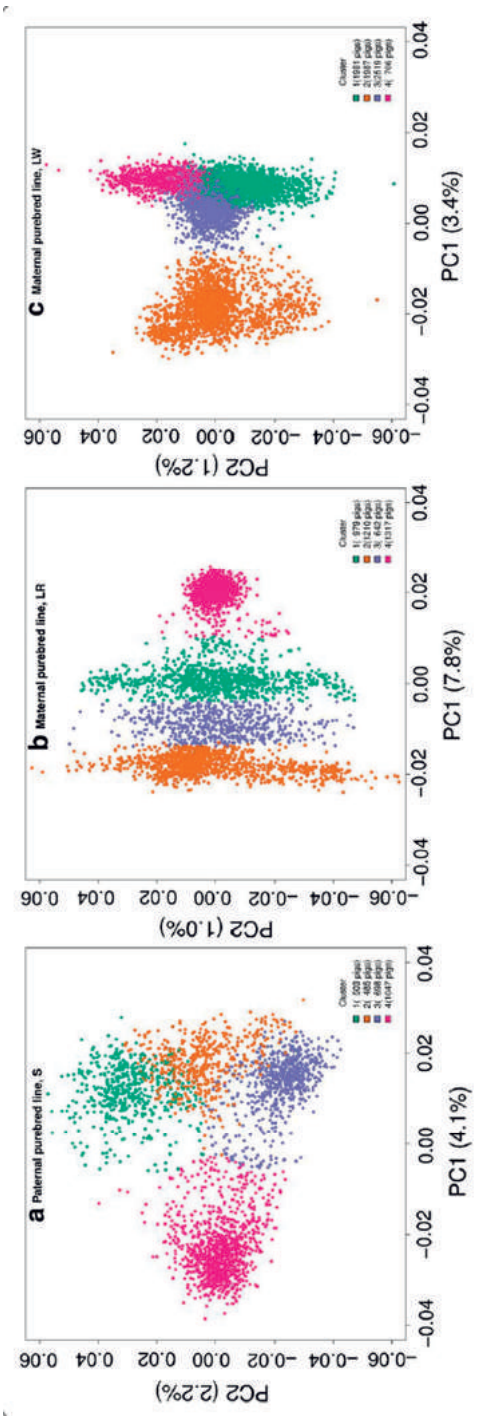

Figure 3.2 The two first principal components (PC) from the genomic relationship matrix between the four validation groups of Synthetic boar (S) pigs (a), Landrace (LR) pigs (b) and Large White (LW) pigs (c). Each circle (o) represents a pig.

Table 3.8 Accuracies* of BOA[a], G_A[b], and G_B[c] models calculated for each of the four folds of cross-validation for estimating breeding values of the paternal breed Synthetic boar pigs for crossbred performance for each trait, and average weighting factor (*w*) of the calculated DRP per validation fold

Folds	*w*	BOA	$\mathbf{G_A}$	$\mathbf{G_B}$
Average daily gain				
1	0.49	0.055	0.055	0.057
2	0.12	0.128	0.111	0.094
3	0.21	0.170	0.156	0.152
4	0.07	0.063	0.084	0.082
Mean		*0.104*	*0.102*	*0.096*
Back fat thickness				
1	0.31	0.168	0.168	0.162
2	0.39	0.201	0.157	0.159
3	0.52	0.191	0.294	0.280
4	0.25	0.150	0.179	0.177
Mean		*0.178*	*0.199*	*0.195*
Loin Depth				
1	0.55	0.204	0.234	0.236
2	0.67	0.212	0.209	0.207
3	0.88	0.127	0.140	0.134
4	0.45	0.088	0.135	0.142
Mean		*0.158*	*0.179*	*0.180*

*Accuracies measured as weighted correlation between DRP and EBVs of S pigs for crossbred performance. Approximate standard errors SE, computed as $(1 - \mathrm{r}^2)/\sqrt{\mathrm{N}}$, were equal to 0.023 to 0.024 for the mean accuracies across the folds, for all combinations of traits and methods

[a]BOA model, model with breed-specific relationship matrices

[b]$\mathrm{G_A}$ model, model with genomic relationship matrix by allele frequencies obtained across the genotyped population

[c]$\mathrm{G_B}$ model, model with genomic relationship matrix by breed-specific allele frequencies

Table 3.9 Accuracies* of BOA[a], G_A[b], and G_B[c] models calculated for each of the four folds of cross-validation for estimating breeding values of the maternal breed Landrace pigs for crossbred performance for each trait, and weighting factor (w) of the calculated DRP

Folds	w	BOA	$\mathbf{G_A}$	$\mathbf{G_B}$
Average daily gain				
1	0.20	0.133	0.106	0.099
2	0.23	0.190	0.095	0.111
3	0.21	0.159	0.106	0.106
4	0.22	0.094	0.007	0.014
Mean		*0.144*	*0.079*	*0.083*
Back fat thickness				
1	0.09	0.185	0.169	0.171
2	0.07	0.186	0.210	0.199
3	0.10	0.223	0.216	0.215
4	0.09	0.144	0.149	0.141
Mean		*0.184*	*0.186*	*0.181*
Loin Depth				
1	0.43	0.224	0.206	0.203
2	0.47	0.085	0.107	0.107
3	0.45	0.239	0.232	0.228
4	0.47	0.170	0.208	0.207
Mean		*0.179*	*0.188*	*0.186*

*Accuracies measured as weighted correlation between DRP and EBVs of LR pigs for crossbred performance. Approximate SE's, computed as $(1 - r^2)/\sqrt{N}$, were equal to 0.024 for the mean accuracies across the folds, for all combinations of traits and methods

[a]BOA model, model with breed-specific relationship matrices

[b]G_A model, model with genomic relationship matrix by allele frequencies obtained across the genotyped population

[c]G_B model, model with genomic relationship matrix by breed-specific allele frequencies

Table 3.10 Accuracies* of BOA[a], G_A[b], and G_B[c] models calculated for each of the four folds of cross-validation for estimating breeding values of the maternal breed Large White pigs for crossbred performance for each trait, and weighting factor (*w*) of the calculated DRP

Folds	*w*	BOA	$\mathbf{G_A}$	$\mathbf{G_B}$
Back fat thickness				
1	0.21	0.217	0.221	0.216
2	0.13	0.095	0.094	0.089
3	0.28	0.190	0.175	0.170
4	0.23	0.219	0.242	0.243
Mean		*0.180*	*0.183*	*0.180*
Loin Depth				
1	0.62	0.235	0.234	0.232
2	0.38	0.103	0.126	0.126
3	0.74	0.226	0.229	0.228
4	0.64	0.297	0.318	0.318
Mean		*0.215*	*0.227*	*0.226*

*Accuracies measured as weighted correlation between DRP and EBVs of LR pigs for crossbred performance. Approximate SE's, computed as $(1-\mathrm{r}^2)/\sqrt{\mathrm{N}}$, were 0.023-0.024 for the mean accuracies across the folds, for all combinations of traits and methods.

[a]BOA model, model with breed-specific relationship matrices.

[b]$\mathrm{G_A}$ model, model with genomic relationship matrix by allele frequencies obtained across the genotyped population.

[c]$\mathrm{G_B}$ model, model with genomic relationship matrix by breed-specific allele frequencies

3.4 Discussion

3.4.1 Properties of the relationship matrices

Genomic relationships within and across populations are defined differently depending on how the genetic covariance between individuals is calculated. Using across-breed allele frequencies when the correlations of allele frequencies between breeds differ from 1, could lead to genomic relationships between animals of different breeds that are on average negative (Lourenco et al., 2016), as observed for the $\mathbf{G_A}$ matrix. This was not the case for the $\mathbf{G_B}$ matrix, in which the genomic relationships between animals of different breeds was on average 0, as expected for distantly-related breeds.

Diagonal elements (D) from a pedigree-based relationship matrix have a value of 1 when there is no inbreeding. Because a genomic relationship matrix is built to resemble a pedigree-based relationship matrix and the current genotyped population is considered the base population (VanRaden, 2008), the average D from a genomic relationship matrix is expected to be 1, as we observed for the $\mathbf{G_A}$

and $\mathbf{G_B}$ matrices. To calculate the partial relationship matrices, $\mathbf{G}^{(S)}$, $\mathbf{G}^{(LR)}$ and $\mathbf{G}^{(LW)}$, the D for CB pigs were expected to be 0.5 for $\mathbf{G}^{(S)}$, and 0.25 for $\mathbf{G}^{(LR)}$ and $\mathbf{G}^{(LW)}$, expressing the proportion of the genome in CB pigs contributed by each breed S, LR, and LW, respectively. Using all the 52,164 SNPs, Fig. 3.3 shows how the diagonal elements among CB pigs from the $\mathbf{G}^{(LR)}$, and $\mathbf{G}^{(LW)}$ matrices increased as the percentage of alleles of CB pigs assigned to the respective maternal breed as breed of origin increased.

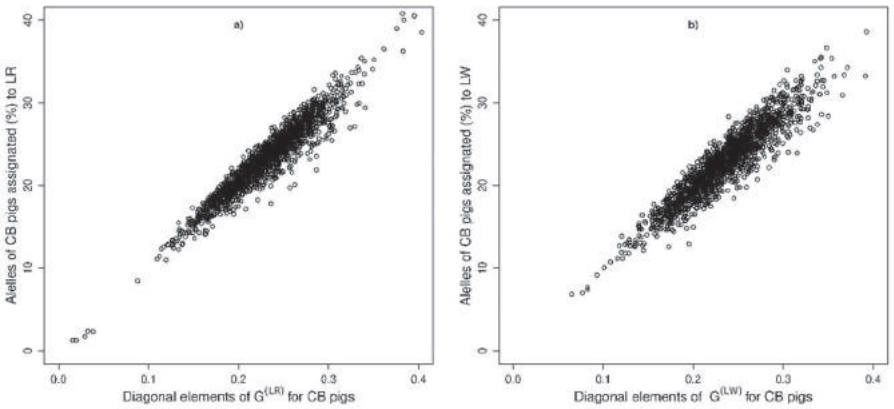

Figure 3.3 Relation between percentage of assigned alleles to a breed of origin and diagonal elements of partial relationship matrices. (a) Observed percentage of assigned alleles of crossbred pigs to Landrace (LR) as breed of origin on the y-axis compared to the diagonal elements of the $\mathbf{G}^{(LR)}$ partial relationship matrix for the same crossbred pigs on the x-axis. (b) Observed percentage of assigned alleles of crossbred pigs to Large White (LW) as breed of origin on the y-axis compared to the diagonal elements of the $\mathbf{G}^{(LW)}$ partial relationship matrix for the same crossbred pigs on the x-axis.

3.4.2 Variance components across models

Estimating variance components for the 4-trait multivariate models was not possible due to workspace memory limitation when trying to run the full BOA model with the three partial relationship matrices or the G models with the relationship matrices containing the four populations. Therefore, for the G models, the construction of a full variance-covariance matrix based on sub-models was required, in this case three bivariate models. This procedure of constructing a full variance-covariance matrix is often used in genetic evaluation (Jorjani et al., 2003). The combined variance-covariance matrices in the G_A and G_B model for BF were

considerably bended (variance components changed up to 10.9%) and this may have affected the results. The combined variance-covariance matrix in the G_A model for LD was also bended, but, in this case, the components changed only up to 2.5%. For ADG, no bending of the variance-covariance matrix was required for any of the models. An advantage of the BOA model, since variance-covariance matrices are by breed, is that it allows the estimates of the CB additive genetic variance contributed by the different parental breeds to differ. With the G_A and G_B models, these differences cannot be observed because there is only one estimate for CB additive genetic variance across the three breeds. A disadvantage of the BOA model is that estimates must be based on half the information (for the paternal breed) or on a quarter of the information (for the maternal breeds) compared to estimates from the G_A or G_B models. Therefore, the SE of CB additive genetic variances estimated with the BOA model were much larger than the SE of CB additive genetic variances estimated with the G_A and G_B models. With the BOA model, we could observe that estimates of CB additive genetic variance differed between the three breeds for all traits. This means that r_{pc} should also be interpreted separately by breed. The estimates of r_{pc} differed slightly across models. In theory, the CB additive variance components estimated with the BOA model comprises the variation observed in CB pigs due only to the alleles coming from the analyzed breed. Therefore, differences in r_{pc} estimated with the G_A or G_B model rather than the BOA model were expected. For instance, for ADG, the r_{pc} estimated with the BOA model for S and LW were slightly smaller than those estimated with the G_A and G_B models. However, the r_{pc} estimated with the BOA model for LR was twice as high compared to that of the other two models. One explanation is that a large part of the CB additive variance can come mainly from variation observed among the alleles originating from a specific breed and this is not captured when all alleles are assumed to have the same origin.

In the literature, r_{pc} for production traits have been calculated from pedigree information only (Brandt and Täubert, 1998; Zumbach et al., 2007) and vary greatly, but on average they are higher than our estimates, probably because the breeds were different or the estimates were an average across different breeds. In general, the investigated traits showed a moderate r_{pc} indicating that using CB information together with PB information in the reference population might be beneficial for selection of PB pigs for CB performance. Using CB information is expected to be most important for combinations of trait and breed for which r_{pc} is low, for instance for ADG in breed LR.

From the estimates of the BOA model, we observed that CB litter effect and residual variance were much larger than those obtained with the G_A or G_B models. Because the genotypes of only one breed at a time were used in the bivariate BOA model, the litter and residual effect variance in the BOA model likely absorbed the variance coming from the genetic relationships from the breeds that were absent in the model. To investigate the impact of these possibly inflated litter and residual variances, we tried to correct this by setting the CB litter effect and residual variance of the BOA model equal to the average estimates from the G_A and G_B models. Using these new variance estimates did not affect the accuracies of the BOA model compared to the G_A and G_B models (results not shown).

3.4.3 Predictive ability across models

The three breeds used in this study are distantly related and correlations between breed-specific allele frequencies were low: 0.31 for breeds S and LR, 0.54 for breeds S and LW, and 0.39 for breeds LR and LW. However, taking population structure into account by accounting for different allele frequencies in the three different breeds (G_B model) did not improve the accuracy for predicting EBV compared with using allele frequencies obtained across genotyped populations (G_A model). In a study with CB sheep, Moghaddar et al. (2014) reported limited impact on prediction accuracy when adjusting for breed-specific allele frequency, also when differences in allele frequencies between breeds were large. Makgahlela et al. (2014) and Lourenco et al. (2016) also observed no advantage of using breed-specific allele frequencies for constructing the relationship matrix, even when this led to observable changes in the coefficients of the relationship matrix. Although correlations between breed-specific allele frequencies were low, correlations between these breed-specific allele frequencies and the across-breed frequency were relatively high, simply because the breed-specific allele frequencies are included in the across-breed allele frequency. In our study, the correlations between the breed-specific allele frequencies and the across-breed frequency were equal to 0.74, 0.68, and 0.89, for breeds S, LR, and LW, respectively. The correlation between breed LW allele frequency and the across-breed frequency was higher than the others, because the LW breed has the largest number of pigs (Table 3.1), therefore, it has a larger contribution to the across-breed allele frequencies across breeds. The correlation between crossbred allele frequency and the across-breed frequency was equal to 0.93. Therefore, using breed-specific or across-breed frequencies in the calculation of the relationship coefficient between a PB and CB pig will have little effect on predicted EBV of PB for CB performance.

In the G_A and G_B models, genetic co-variances between breeds were assumed to be zero. To test if this was a correct assumption, covariances between PB lines were also estimated by fitting three additional bivariate models (one for each pair of PB) for the trait ADG using the G_A model. Variance components of the six bivariate models were combined to obtain the full variance-covariance matrices for the 4-trait model. This combination was performed by averaging the three variance components estimated for each population, i.e. S, LR, LW and CB. In this case, it was not necessary to bend the combined variance-covariance matrix to make it positive definite. The genetic correlations between PB performance for ADG were 0.13 (±0.24) between S and LR, 0.39 (±0.14) between S and LW, and 0.36 (±0.16) between LR and LW. These estimates were in line with estimated values of 0.23 and 0.30 between a Danish Landrace and Danish Yorkshire population (Xiang et al., 2017). Moreover, for breeds S and LR, the value of zero was within one SD. Accuracies of the G_A model taking into account the covariance between PB for estimating breeding values of S pigs for CB performance, were similar to prediction accuracy of the G_A model assuming the covariances between PB to be zero (Table 3.11). This was expected because relationships between pigs from different breeds were low and showed very little variation (Table 3.6). Therefore, the G_A model assuming the covariances between PB to be zero are not expected to affect accuracies, even when genetic correlations between PB are moderate.

The BOA model assumes that relationships between PB are zero, and thus, also effectively assumes that the covariances between PB are zero. A study from Xiang et al. (2017) compares the BOA approach in a single-step model against a single-step model with metafounders, where the last model defines relationships between the pedigree base populations across breeds but also takes genomic relationships across breeds into account. Taken together their conclusions that both models perform similarly and our findings, these results suggest that considering genomic relationships and covariances between PB lines has limited relevance in models for predicting crossbred performance for pig crossbreeding programs.

Compared to the G_A and G_B models, taking population structure into account by using breed-specific partial relationships as in the BOA model, including breed-specific allele frequencies, had some impact on the accuracy of EBV. The BOA model had a positive impact for traits with a low r_{pc} as for ADG in breed LR (0.30). BF and LD showed higher r_{pc} (0.55 to 0.73), and accuracies of the BOA model for

these traits was similar to those of the G_A or G_B models. Comparing PB lines, somewhat higher accuracies could have been expected for the S line, because the sire line contributes 50% of the genome of the CB, while the dam lines contribute only 25%. Thus, the sire line will have a larger variance in genomic relationships with the CB pigs used for training, which is expected to yield higher accuracies (Ibánẽz-Escriche et al., 2009). Nevertheless, in our study, accuracies were very comparable across the sire and dam lines. The BOA model was previously tested on simulated data (Ibánẽz-Escriche et al., 2009; Esfandyari et al., 2015), and on real data but for a two-breed cross scheme (Xiang et al., 2016; Lopes et al., 2017). These studies also compared the BOA model to models similar to G_A and G_B. Ibánez-Escriche et al. (2009) used a simulated population of two-way and three-way CB, for a trait with a heritability of 0.3. They observed that the prediction accuracy of EBV of PB pigs for CB performance with the G_A model was often equal or higher compared to that with the BOA model. The superiority of the BOA model was only observed when PB populations were distant or unrelated, and SNP density was low. Similarly, Esfandyari et al. (2015) tested the BOA model with a simulated two-way CB population for a trait with a heritability of 0.3 and a r_{pc} of 0.78. They observed a higher response to selection in CB animals when the BOA model was used compared to the G_A model, but, again, only when PB populations were distantly related. Vandenplas et al. (2017) predicted the average reliability of EBV for CB performance obtained from the G_B and BOA models using simulated PB and two-way CB data and different heritabilities (0.20, 0.40, and 0.95), r_{pc} (0.30 and 0.70), and population relatedness. In their study, average reliabilities of the BOA model were always lower than those of the G_B model. The difference in reliabilities between the BOA and G_B models also increased with increasing heritability, r_{pc} and with the population relatedness. Using real data of two-way CB, Xiang et al. (2016) and Lopes et al. (2017) tested the BOA approach. Xiang et al. (2016) used a single-step model with a trait that had a CB heritability of 0.10 and r_{pc} of 0.59 and 0.73 between each breed. They obtained up to 13% higher accuracy for EBV of PB pigs for CB performance considering breed-specific SNP effects. Lopes et al. (2017) tested the BOA approach with two traits that had a CB heritability of 0.14 and 0.37, respectively and r_{pc} higher than 0.88. They obtained similar prediction accuracies with the BOA approach than with a model that did not account for breed-specific SNP effects in CB animals. The results from these studies indicate that breeding values are better estimated with the BOA model for traits with a low heritability and low r_{pc}. In our study, CB and PB heritabilities were higher than 0.22, which may have limited the positive impact of the BOA model. Therefore, already considering

distantly-related breeds, the BOA model seems to outperform the G_A and G_B models for predicting breeding values of PB animals for CB performance, only when the r_{pc} and heritabilities of the analysed trait are low.

3.5 Conclusions

A positive impact of the BOA model was observed for ADG in breed LR, which showed a low r_{pc} (0.30). Results from the literature and from our study suggest that, in cases where traits have a combination of low r_{pc} and low heritabilities, and breeds are distantly related, the use of the BOA model is justified. In other cases, using CB information in a model that does not account for breed-specific SNP effects in CB animals, such as the G_A and G_B models, does not seem to jeopardize predictions and may be preferred because it can be more easily implemented than the BOA model.

Chapter 4

Effects of alleles in crossbred pigs estimated for genomic prediction depend on their breed of origin

Claudia A. Sevillano[1,2], Jan ten Napel[1], Simone E.F. Guimarães[3], Fabyano F. Silva[3], Mario P.L. Calus[1]

[1] Wageningen University & Research Animal Breeding and Genomics, 6700AH, Wageningen, the Netherlands; [2] Topigs Norsvin Research Center, 6640AA, Beuningen, the Netherlands; [3]Department of Animal Science, Universidade Federal de Viçosa, 36570-000 Viçosa, Minas Gerais, Brazil

BMC Genomics (2018) 19:740

Abstract

This study investigated if the allele effect of a given single nucleotide polymorphism (SNP) for crossbred performance in pigs estimated in a genomic prediction model differs depending on its breed of origin, and how these are related to estimated effects for purebred performance. SNP-allele substitution effects were estimated for a commonly used SNP panel using a genomic best linear unbiased prediction model with breed-specific partial relationship matrices. Estimated breeding values for purebred and crossbred performance were converted to SNP-allele effects by breed of origin. Differences between purebred and crossbred, and between breeds-of-origin were evaluated by comparing percentage of variance explained by genomic regions for back fat thickness (BF), average daily gain (ADG), and residual feed intake (RFI). From ten regions explaining most additive genetic variance for crossbred performance, 1 to 5 regions also appeared in the top ten for purebred performance. The proportion of genetic variance explained by a genomic region and the estimated effect of a haplotype in such a region were different depending upon the breed of origin. To illustrate underlying mechanisms, we evaluated the estimated effects across breeds-of-origin for haplotypes associated to the melanocortin 4 receptor (MC4R) gene, and for the MC4Rsnp itself which is a missense mutation with a known effect on BF and ADG. Although estimated allele substitution effects of the MC4Rsnp mutation were very similar across breeds, explained genetic variance of haplotypes associated to the MC4R gene using a SNP panel that does not include the mutation, was considerably lower in one of the breeds where the allele frequency of the mutation was the lowest. To conclude, similar regions explaining similar additive genetic variance were observed across purebred and crossbred performance. Moreover, there was some overlap across breeds of origin between regions that explained relatively large proportions of genetic variance for crossbred performance; albeit that the actual proportion of variance deviated across breeds of origin. Results based on a missense mutation in MC4R confirmed that even if a causal locus has similar effects across breeds of origin, estimated effects and explained variance in its region using a commonly used SNP panel can strongly depend on the allele frequency of the underlying causal mutation.

Key words: crossbred, pig, breed of origin, genomic prediction, association study.

4.1 Introduction

In pig production, as selection is performed in purebred lines, while the final product is a crossbred animal, there is an anticipated benefit of using crossbred information for estimating breeding values of purebred for crossbred performance (Hidalgo et al., 2015a; Lopes et al., 2017). The genetic correlation between purebred and crossbred performance (r_{pc}) determines the effect of selection in the purebred animals on the rate of genetic change in the crossbred animals (Wei and Van der Steen, 1991; Brandt and Täubert, 1998). As r_{pc} decreases, the benefit of using crossbred information increases (Bijma and Van Arendonk, 1998; Dekkers, 2007).

Moreover, crossbred genomic information is composed of a mosaic of genomic regions inherited from the different parental breeds (i.e. breed of origin). As a result, depending from which breed of origin an allele was inherited from, it might have different effects. These different allele effects can be due to: (1) quantitative trait loci (QTL) may be in linkage disequilibrium with different single nucleotide polymorphisms (SNPs) depending from which parental breed the QTL was inherited (Lopes, 2016), (2) the functional variation that underlies the inherited QTL may have different minor allele frequencies (MAF) in different parental breeds, with the extreme case where it is not segregating in one or more breeds (Wientjes et al., 2015), (3) epistatic interactions in one parental breed may be different due to other genes that modify the effect of the inherited QTL in that breed (Mackay, 2014), and above all these reasons (4) multiple and different quantitative trait nucleotides (QTN) could be underlying a QTL in different parental breeds. Therefore, the allele effect of a given SNP for purebred performance might differ from its effect for crossbred performance, and an allele of that given SNP could have a different effect on crossbred performance depending on the breed it was inherited from. Thus, SNP by genetic background interactions may be relevant when training with crossbred information to estimate breeding values of purebred animals for crossbred performance.

Several studies support that effects of SNPs may be breed-specific. Firstly, in many cases, estimated effects of SNPs in an association study for a given breed are not replicated by studies in other breeds (Diniz et al., 2014; Sevillano et al., 2015; Hidalgo et al., 2016). Secondly, associations found in a breed often are not replicated in crossbred populations derived from this breed (Kumar et al., 2005; Bolormaa et al., 2013). Finally, the proportion of genetic variance in crossbred

performance that is explained by each parental purebred population appears to deviate from the breed proportions (Hidalgo, 2015b).

With crossbreeding, SNPs can be observed in the different genetic backgrounds. Estimation of background specific effects, however, requires that the SNP alleles present in the crossbred animal are assigned to one of the parental breeds. Recently, we have developed a procedure that enables breed-of-origin assignment of alleles in three-way crossbred animals (Sevillano et al., 2016; Vandenplas et al., 2016). Knowing the breed of origin enables to estimate SNP effects for crossbred performance depending on the breed of origin, and to compare those within breed to estimated effects for purebred performance.

For traits with low r_{pc} (<0.70), tracing the breed of origin of alleles and using this information in a genomic best linear unbiased prediction model that accounts for breed-specific SNP effects for crossbred performance (BOA model) tended to show better predictive abilities compared to models in which SNP effects are assumed to be the same across breeds (Sevillano et al., 2017). This is another indication that the effect of alleles estimated for crossbred performance might be different depending upon the breed of origin. The objective of this study, therefore, was to investigate if the allele effect of a given SNP for crossbred performance in pigs estimated in a genomic prediction model using a commonly used SNP panel differs depending on its breed of origin, and how these related to estimated effects for purebred performance. For this, we estimated breed-specific SNP effects from the solutions of a BOA model. Based on previous results (Sevillano et al., 2017; Godinho et al., 2018) we chose three traits, back fat thickness (BF) with an r_{pc} of 0.82, a heritability for crossbred performance of 0.43 and no better predictions observed when using the BOA model; average daily gain (ADG) with an r_{pc} of 0.61, a heritability for crossbred performance of 0.26 and better predictions observed when using the BOA model; and residual feed intake (RFI) with an r_{pc} of 0.62, a heritability for crossbred performance of 0.19, but not tested previously with the BOA model. To illustrate how the effect of SNP-alleles in crossbred pigs depend on their breed of origin, we evaluated the estimated effects across breeds of origin for the melanocortin 4 receptor (MC4Rsnp) which has a missense mutation that is known to affect BF and ADG.

4.2 Methods

4.2.1 Data

The data consisted of three purebred-based pig populations; S, LR, and LW, and one crossbred population (S (LR x LW) or S (LW x LR)). S is a synthetic sire line created as a combination of Large White and Pietrain. LR is a Landrace based dam line and LW is a Large White based dam line. All pigs were genotyped using one of the three following SNP panels: Illumina PorcineSNP60.v2 BeadChip (60K.v2), Illumina PorcineSNP60 BeadChip (60K), or Illumina PorcineSNP10 BeadChip (10K). Pigs genotyped with the 60K or 10K chips were imputed to the 60K.v2 panel using FImpute Version 2.2 software (Sargolzaei et al., 2014) with default parameter settings and using pedigree information. The imputation strategy was similar to Sevillano et al. (2016), where each of the three purebred populations; LR, LW, and S, were imputed in two steps: (1) pigs genotyped with the 10K chip were imputed to 60K, and (2) all pigs with 60K data (imputed or genotyped) were imputed to 60K.v2. This strategy was chosen because the 10K panel shared more SNPs (8743) with the 60K panel than with the 60K.v2 panel (6861). For the crossbred population, imputation was done in a single step, crossbred pigs genotyped with the 10K chip were directly imputed to 60K.v2, because all ancestors were genotyped or already imputed to 60K.v2.

Performance from purebred pigs were available from 52 nucleus and combined crossbred purebred system (CCPS) farms recorded from August 2005 until August 2016. Performance from crossbred pigs were available from 7 CCPS and experimental farms recorded from January 2009 until March 2016. Phenotypes for BF and ADG were measured in most of the purebred and crossbred pigs. BF for purebred pigs was measured on average at 173 days of age using an ultrasound instrument, while BF for crossbred pigs was measured on the carcass using a probe, named “capteur gras maigre” (CGM; Sydel, France), crossbred pigs were slaughtered when they achieved 120 kg (at an average age of 169 days). BF was measured approximately at the third to fourth rib from the last rib position. ADG for purebred pigs was calculated as the difference of on-test body weight at an average age of 60 days and off-test body weight at an average age of 173 days divided by the phase length. ADG for crossbred pigs was calculated as the difference of on-test body weight at an average age of 70 days of age and body weight at end of the finishing period, which was on average 120 kg, divided by the phase length. RFI was obtained as the estimated residual term from the following regression model (Cai et al., 2008):

$$ADFI = \mu + b_1 BW_{on} + b_2 BW_{off} + b_3 BF + b_4 ADG + e,$$

in which ADFI is the average daily feed intake, μ is the mean, BW_{on} is the on-test body weight, BW_{off} is the off-test body weight, BF and ADG are the previously described traits, b_1, b_2, b_3, and b_4 are the linear coefficients of the regression on covariates, and *e* is the RFI. The numbers of available genotypes and phenotypes per trait and per population are summarized in Table 4.1. For all phenotyped pigs, four generations of pedigree information were included for analysis.

Table 4.1 Number of genotypes and phenotypes available per trait and per population.

Population	Genotypes	BF/ADG	RFI
S	8079	7547	2102
LR	5233	3288	56
LW	15 727	12 794	1133
Crossbred	3352	2816	2695
Total	32 391	26 445	5986

BF = back fat thickness, ADG = average daily feed intake, and RFI = residual feed intake

4.2.2 Estimation of SNP-allele effects

SNP-allele substitution effects were estimated using best linear unbiased predictions (BLUP) similar to Wang et al. (Wang et al., 2014). However, instead of using a single-step BLUP, we used a genomic BLUP (GBLUP) with breed-specific partial relationship matrices (BOA model) (Sevillano et al., 2017). With this approach, genomic estimated breeding values (GEBV) for purebred and crossbred performance could be calculated, and posteriorly converted to SNP-allele effects by breed of origin. SNP-allele effects were derived using the following steps:

1. Determine breed of origin of alleles to calculate breed-specific partial relationship matrices, $\mathbf{G}^{(\mathbf{S})}$, $\mathbf{G}^{(\mathbf{LR})}$, and $\mathbf{G}^{(\mathbf{LW})}$.
2. Calculate GEBVs for purebred and crossbred performance using a GBLUP model with breed-specific partial relationship matrices (BOA model).
3. Back-solve SNP-allele effects for purebred and crossbred performance from GEBVs.
4. Calculate proportion of variance explained by non-overlapping blocks of SNPs.

4.2.2.1 Inference of the breed of origin of alleles

To infer breed of origin of alleles in crossbred pigs we used the BOA approach developed by Vandenplas et al. (2016) and assuming the parameter settings recommended by Sevillano et al. (2016). The BOA approach consisted of three steps: (1) Phasing the haplotypes of both purebred and crossbred pigs with

AlphaPhase1.1 software (Hickey et al., 2011). Phasing was performed using pedigree, and using nine combinations of haplotypes length and each combination was run both considering "Offset" and "NotOffset" modes, the "Offset" mode shifts the start of the cores to halfway along the first core, creating 50% overlaps between cores. These settings allowed each allele to be considered 18 times through different haplotypes of variable length. (2) Determining the unique haplotypes among the purebred pigs. For assigning a breed of origin to a haplotype, at least 80% of its copies were required to be observed in a specific breed. (3) Assigning the breed of origin for each allele carried on the haplotypes of crossbred pigs based on the knowledge of the breed of origin of the haplotypes, on the zygosity (i.e., homozygosity or heterozygosity) of the locus, and on the breed composition of the crossbred. Alleles that were not assigned a breed of origin were set to missing. SNPs for which the paternal or maternal allele were assigned a breed of origin in less than 90% of the cases were removed. Crossbred pigs with assigned breed of origin for less than 90% of their genome were removed. If an allele was observed less than 5 times in any of the breeds-of-origin, the corresponding SNP was also removed from the final set of SNPs. The final SNP set for subsequent analyses consisted of 41,557 SNPs. All populations were analysed with the same set of SNPs.

4.2.2.2 Model with three breed-specific partial relationship matrices

To account for breed-specific effect of alleles, a 4-trait animal model (i.e., S trait, LR trait, LW trait and crossbred trait) with three breed-specific partial relationship matrices ($\mathbf{G}^{(\mathbf{S})}, \mathbf{G}^{(\mathbf{LR})}, \mathbf{G}^{(\mathbf{LW})}$) was fitted (BOA model) (Sevillano et al., 2017). The three breed-specific partial relationship matrices, $\mathbf{G}^{(\mathbf{S})}$, $\mathbf{G}^{(\mathbf{LR})}$, and $\mathbf{G}^{(\mathbf{LW})}$, were built using the breed of origin of phased alleles in crossbred pigs and the first method from VanRaden (2008). The breed-specific partial relationship submatrices are defined, considering e.g. the breed S origin, as:

$$\mathbf{G}_{\mathbf{S},\mathbf{S}} = (\mathbf{M}^{\mathbf{S}} - 2\mathbf{1}\mathbf{p}^{\mathbf{S}\prime})(\mathbf{M}^{\mathbf{S}} - 2\mathbf{1}\mathbf{p}^{\mathbf{S}\prime})'\mathrm{F}^{-1},$$

$$\mathbf{G}_{\mathbf{S},\mathbf{CB}} = (\mathbf{M}^{\mathbf{S}} - 2\mathbf{1}\mathbf{p}^{\mathbf{S}\prime})\left(\mathbf{M}^{\mathbf{CB}(\mathbf{S})} - \mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right)'\mathrm{F}^{-1}, \text{ and}$$

$$\mathbf{G}_{\mathbf{CB},\mathbf{CB}} = (\mathbf{M}^{\mathbf{CB}} - \mathbf{1}\mathbf{p}^{\mathbf{S}\prime})\left(\mathbf{M}^{\mathbf{CB}(\mathbf{S})} - \mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right)'\mathrm{F}^{-1},$$

where $\mathbf{M}^{\mathbf{S}}$ is a matrix containing breed-specific allele content for purebred S pigs (coded as 0, 1, or 2). $\mathbf{M}^{\mathbf{CB}(\mathbf{S})}$ is a matrix containing breed S allele content for crossbred pigs (coded as 0, or 1), so that alleles not assigned a breed of origin were

set to missing, meaning that they had an entry of zero in the centered matrix represented by $(\mathbf{M^{CB}} - \mathbf{1p^{S\prime}})$; $\mathbf{p^S}$ is the vector of breed S specific frequencies of the counted allele (p_j^S), where p_j^S was calculated across S and crossbred pigs by counting the occurrences of alleles originating from the S breed and coded as 1, divided by the total number of S alleles in the S breed and crossbred on locus *j*. Finally, the scaling factor was defined as $\mathrm{F} = \sum_j 2p_j^S(1 - p_j^S)$. The breed-specific partial relationship submatrices $\mathbf{G^{(LR)}}$ and $\mathbf{G^{(LW)}}$ are defined similarly to $\mathbf{G^{(S)}}$. Other effects in the model included fixed effects partially depending on the trait (Table 4.2), and random common litter effects. The BOA model was implemented in the MiXBLUP software (Ten Napel et al., 2016). To estimate variance components we used the same BOA model in the ASReml software (Gilmour et al., 2015), after reducing each of the purebred populations to 3500 pigs most closely related to the crossbred population.

Table 4.2 Fixed effects used in the models for each trait for purebred and crossbred pigs.

Trait	Population	Fixed effects
BF	Purebred	$\mathrm{farm} * \mathrm{breed} * \mathrm{sex} + \mathrm{b_a} \times \mathrm{off-test\ BW}$
	Crossbred	$\mathrm{trial} + \mathrm{farm} * \mathrm{sex} + \mathrm{b_b} \times \mathrm{hot\ carcass\ weight}$
ADG	Purebred	$\mathrm{farm} * \mathrm{breed} * \mathrm{sex} + \mathrm{b_c} \times \mathrm{birth\ weight}$
	Crossbred	$\mathrm{trial} + \mathrm{farm} * \mathrm{sex} + \mathrm{b_c} \times \mathrm{birth\ weight}$
RFI	Purebred	$\mathrm{farm} * \mathrm{breed} * \mathrm{sex} + \mathrm{b_d} \times \mathrm{on-test\ BW}$
	Crossbred	$\mathrm{trial} + \mathrm{farm} * \mathrm{breed} * \mathrm{sex} + \mathrm{b_d} \times \mathrm{on-test\ BW}$

BF = back fat thickness, ADG = average daily gain, and RFI = residual feed intake.
b_a, b_b, b_c, b_d, are regression coefficients for off-test BW, hot carcass weight, birth weight, and on-test BW, respectively.

4.2.2.3 Back-solve SNP-allele effects from GEBV

GEBV of purebred S pigs for purebred performance ($\hat{\mathbf{a}}_\mathbf{S}$) were converted to SNP-allele effects ($\hat{\alpha}_s$), considering that:

$$\hat{\mathbf{a}}_\mathbf{S} = \mathbf{W^S}\hat{\boldsymbol{\alpha}}_\mathbf{s},$$

where $\mathbf{W^S}$ contains centered genotypes, which can be obtained respectively by:

$$\mathbf{W^S} = (\mathbf{M^S} - 2\mathbf{1p^{S\prime}}) \text{ and}$$

$$\hat{\boldsymbol{\alpha}}_\mathbf{s} = \mathbf{W^{S\prime}}(\mathbf{W^S W^{S\prime}})^{-1}\hat{\mathbf{a}}_\mathbf{S} = \mathrm{F}^{-1}\mathbf{W^{S\prime}G^{(S)-1}}\hat{\mathbf{a}}_\mathbf{S}.$$

The SNP-allele effects for crossbred performance and for the other purebred populations were calculated similarly.

4.2.2.4 Variance proportion explained by SNP regions

Under a back-solving approach, all SNPs are considered simultaneously in the model, therefore, the effect of a QTL is likely distributed across all SNPs that have a nonrandom association with the QTL. For this reason, it is recommended to calculate the proportion of variance explained by a group of SNPs in nonrandom association instead of reporting effects of single SNPs (Lopes, 2016). Groups of SNPs in nonrandom association will hereafter be called LD blocks. LD blocks were built per breed of origin, therefore, nonrandom association between alleles at two loci was tested in the crossbred population between all pair of loci coming from the same breed of origin. Significant nonrandom association between alleles at two loci was tested with Fisher's exact test on a contingency table made for counts of the four gametic types at the two loci (Slatkin, 1994). If statistical significant nonrandom association is detected (P-value<0.01), then it can be concluded that the coefficient of linkage disequilibrium, D, is significantly different from zero and that pair of loci are in linkage disequilibrium (Slatkin and Excoffier, 1996). Breakpoints between LD blocks were defined when D between two consecutive SNPs was not significantly different from zero. Estimation of D and Fisher's exact test was performed using the Arlequin software (Excoffier et al., 2005).

Percentage of genetic variance explained by the *i*-th LD block was calculated as in Wang et al. (2014):

$$\frac{\mathrm{Var}(\mathrm{a}_i)}{\sigma_{\mathrm{a}}^2} \times \frac{\mathrm{x_n}}{\mathrm{n}_i} \times 100\% = \frac{\mathrm{Var}(\sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{z_j}\hat{\alpha}_{\mathrm{j}})}{\sigma_{\mathrm{a}}^2} \times \frac{\mathrm{x_n}}{\mathrm{n}_i} \times 100\%,$$

where $\mathrm{a_i}$ is genetic value of the *i*-th LD block, σ_{a}^2 is the total genetic variance, $\mathrm{z_j}$ is a vector of genotypes of the *j*-th SNP for all purebred individuals of the same breed, $\hat{\alpha}_{\mathrm{j}}$ is the estimated effect of the *j*-th SNP within the *i*-th LD block that contains *n* SNPs, $\mathrm{x_n}$is the mean number of SNPs across LD blocks and $\mathrm{n}_i$is the number of SNPs of the *i*-th LD block. With the back-solving approach we can identify peaks that explain the most variance, in our case, we took the top 10 LD blocks for comparison across scenarios.

4.2.3 Candidate genes

Putative candidate genes within the top 10 LD blocks and in the neighbouring upstream and downstream 1-Mb regions were identified based on the Sscrofa11.1

genome assembly, using the NCBI Map Viewer (https://www.ncbi.nlm.nih.gov/genome/gdv/?org=sus-scrofa) and based on literature.

4.2.4 MC4R

To illustrate the mechanisms underlying breed of origin specific estimated SNP effects, we investigated the estimated effects across breeds of origin for haplotypes associated to the MC4R gene, and the allele substitution effects for the MC4Rsnp itself. The MC4R gene has a missense mutation that is known to have a strong effect on BF and ADG (Kim et al., 2000). The genotypes at the MC4Rsnp were available for 4996 S, 1363 LR, 7663 LW, and 1478 crossbred pigs. The MC4Rsnp is biallelic (A|G) and located in the MC4R gene at 160,772,887 bp of the SSC1; allele A is the mutant allele (hereafter denoted as allele m) and allele G is the wild type allele (hereafter denoted as allele w). The MC4Rsnp was imputed in pigs that were not genotyped for it and the breed of origin of both alleles were inferred with the BOA approach. After quality control we had information for 7469 S, 3257 LR, 12 707 LW, and 2763 crossbred pigs. Allele frequencies of the MC4Rsnp were computed in each of the purebred populations and in the crossbred population considering breed of origin. In order to build LD blocks that co-segregate with the MC4R gene, linkage disequilibrium was tested between the alleles of MC4Rsnp and the alleles from all the other loci in the SSC1 of the crossbred population (Slatkin, 1994). Unlike the LD blocks previously built, breakpoints to define the MC4R LD blocks were not defined when D between two consecutive SNPs was not significantly different from zero, but when D between the MC4Rsnp and any of the other SNPS in the SSC1 was not significantly different from zero. The effect of each haplotype present in the LD block that co-segregate with the MC4R gene was calculated per breed of origin for crossbred performance for ADG.

To enable comparison to the estimated haplotype effects we also estimated the effect of the MC4Rsnp itself. The effect was estimated with the software ASReml (Gilmour et al., 2015) by applying the following model:

$$ADG_{ij} = \mu + b_S MC4Rsnp_S + b_{LR} MC4Rsnp_{LR} + b_{LW} MC4Rsnp_{LW} + c_i^2 + u_j + e_{ij},$$

where ADG_{ij} was the pre-corrected ADG phenotype of crossbred pig *j*, ADG phenotypes were corrected for fixed effects listed in Table 4.2; $MC4Rsnp_S$, $MC4Rsnp_{LR}$, and $MC4Rsnp_{LW}$ were the centered allele content of MC4Rsnp (0 or 1) of crossbred *j* for breed of origin S, LR, and LW, respectively; b_S, b_{LR}, and b_{LW}

were the unknown allele substitution effect of MC4Rsnp for breed of origin S, LR, and LW, respectively; c_i^2 was the random effect of common litter *i*, assumed to be normally distributed ~ $N(0, \mathbf{I}\sigma_c^2)$, where **I** was an identity matrix and σ_c^2 was the unknown variance between litters; a_j was the random additive genetic effect of crossbred *j* assumed to be normally distributed ~ $N(0, \mathbf{A}\sigma_u^2)$, where **A** was a known matrix of additive genetic relationship among pigs (pedigree-based) and σ_u^2 was the genetic variance between pigs that was estimated in the BOA model; and e_{ij} was the random residual effect assumed to be normally distributed ~ $N(0, \mathbf{I}\sigma_e^2)$, where σ_e^2 was the unknown residual variance.

4.3 Results

4.3.1 Heritabilities and genetic correlations

Estimated variance components and standard errors for BF, ADG, and RFI using the BOA model are shown in Table 4.3. Estimates of crossbred heritability tended to be larger than estimates of purebred heritability. The lowest heritability for crossbred performance was observed for ADG (0.29), while BF and RFI showed similar heritabilities of 0.41 and 0.40, respectively. The lowest r_{pc} was observed for RFI (0.37-0.60), followed by ADG (0.60-0.69), while BF showed the highest r_{pc} (0.71-0.89). Because of the limited number of RFI records from LR pigs, genetic parameters estimated in this population had very high standard errors, therefore, estimates are not shown and were not further used in this study.

Table 4.3 Heritability estimates for purebred (h_{PB}^2) and crossbred (h_{CB}^2) performance, and genetic correlation between performance in purebred and crossbred (r_{PC}).

Trait	Breed	$\boldsymbol{h_{PB}^2}$	$\boldsymbol{h_{CB}^2}$	$\boldsymbol{r_{PC}}$
BF	S	0.31 (0.02)	0.41 (0.04)	0.80 (0.07)
	LR	0.33 (0.03)		0.71 (0.10)
	LW	0.34 (0.03)		0.89 (0.09)
ADG	S	0.09 (0.02)	0.29 (0.04)	0.69 (0.12)
	LR	0.22 (0.02)		0.60 (0.16)
	LW	0.20 (0.02)		0.68 (0.13)
RFI	S	0.15 (0.03)	0.40 (0.05)	0.37 (0.14)
	LR	-		-
	LW	0.61 (0.05)		0.60 (0.18)

BF = back fat thickness, ADG = average daily gain, and RFI = residual feed intake.

4.3.2 Proportion of genetic variance explained by a region

The number and size of the LD blocks are shown per breed of origin in Table 4.4. The LD blocks coming from the S breed of origin were on average the longest (7.1 SNPs), followed by the LD blocks coming from the LW breed of origin (6.4 SNPs), while the LD blocks coming from the LR breed of origin were on average the shortest (5.3 SNPs).

Table 4.4 Description of LD blocks built per breed of origin.

Breed of origin	Number of blocks	Length (number of SNPs)		
		Mean	Min	Max
S	5516	7.1	1	115
LR	7495	5.3	1	56
LW	6296	6.4	1	86

Figures 4.1–4.3 show for each breed genetic variances for all LD blocks for purebred and crossbred performance for BF, ADG, and RFI, respectively. Depending on the breed and across traits, the proportion of genetic variance jointly explained by the top 10 LD blocks with most explained genetic variance ranged, across breeds and traits, from 1.73% to 4.51% for purebred performance and from 1.71% to 4.51% for crossbred performance (Table 4.5). Depending on the trait, and considering that the haploid genome of the domesticated pig is estimated to be 2800 Mb, the top 10 LD blocks covered at least 0.19% and at the most 0.47% of the genome. Proportion of genetic variance and position of each of the top 10 LD blocks for purebred and crossbred performance by breed are detailed in Supplementary material, additional files S4.1-4.3 for BF, ADG and RFI, respectively.

Table 4.5 Percentage of genetic variance explained by top ten LD blocks for purebred and crossbred (CB) performance.

%*	BF	ADG	RFI
Purebred			
gVar S	4.51	3.57	2.80
gVar LR	1.73	1.81	-
gVar LW	2.61	3.00	2.33
CB, breed of origin			
gVar CB,S	4.51	3.80	2.50
gVar CB,LR	2.35	1.71	2.42
gVar CB,LW	2.85	2.71	2.28

*Percentage of genetic variance for purebred performance by breed and for crossbred (CB) performance by breed of origin.
BF = back fat thickness, ADG = average daily gain, and RFI = residual feed intake.

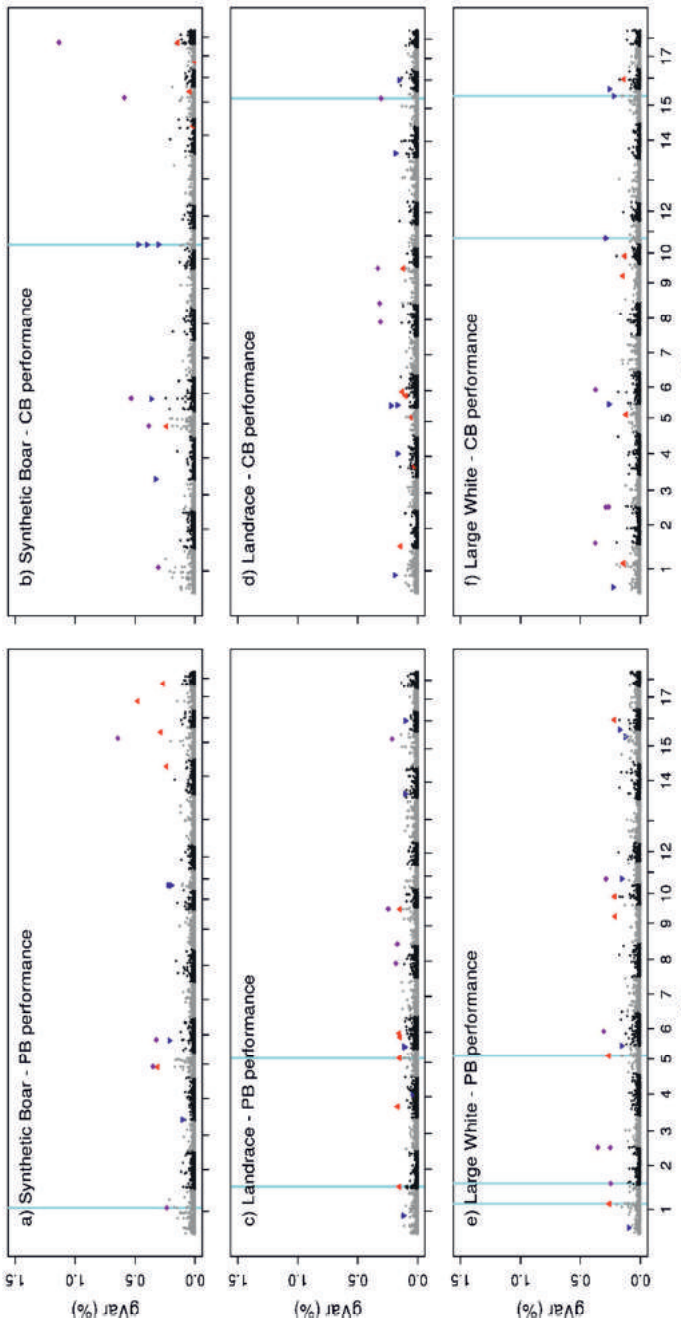

Figure 4.1. Proportion of genetic variance for back fat thickness explained by each LD block. Observed in S (a), LR (c), and LW (e) for purebred performance (PB) and when alleles originate from S (b), LR (d), or LW (f) for crossbred performance (CB). Top 10 LD blocks explaining most variance for PB (red ▲), and top 10 LD blocks explaining most variance for CB performance (blue ▼). LD blocks belonging to the top 10 in both, PB and CB performance (purple ◆). Regions explaining the variance for PB in more than one breed or explaining the variance for CB in more than one breed-of-origin (light blue strip).

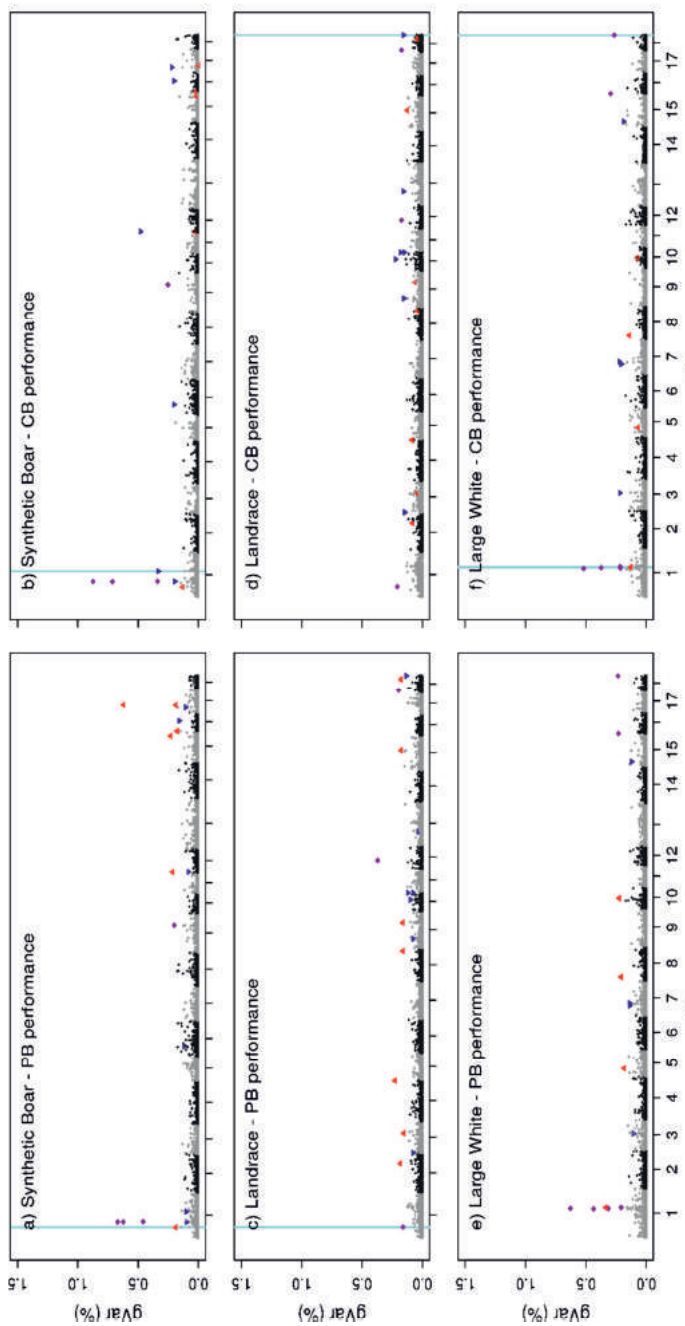

Figure 4.2. Proportion of genetic variance for average daily gain explained by each LD block. Observed in S (a), LR (c), and LW (e) for purebred performance (PB) and when alleles originate from S (b), LR (d), or LW (f) for crossbred performance (CB). Top 10 LD blocks explaining most variance for PB (red ▲), and top 10 LD blocks explaining most variance for CB performance (blue ▼). LD blocks belonging to the top 10 in both, PB and CB performance (purple ♦). Regions explaining the variance for PB in more than one breed or explaining the variance for CB in more than one breed-of-origin (light blue strip).

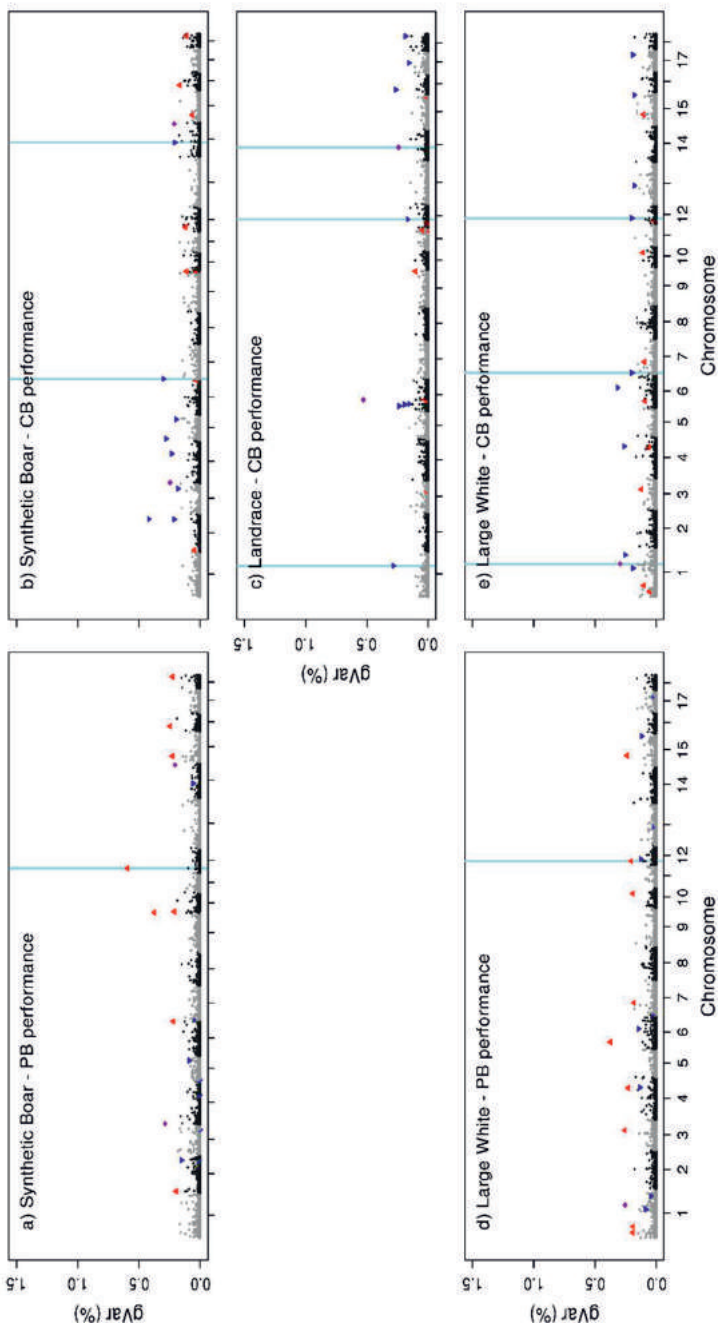

Figure 4.3. Proportion of genetic variance for residual feed intake explained by each LD block. Observed in Synthetic boar (a), and Large White (d) for purebred performance (PB) and when alleles originate from S (b), LR (c), or LW(e) for crossbred performance (CB). Top 10 LD blocks explaining most variance for PB (red ▲), and top 10 LD blocks explaining most variance for CB performance (blue ▼). LD blocks belonging to the top 10 in both, PB and CB performance (purple ◆). Regions explaining the variance for PB in more than one breed or explaining the variance for CB in more than one breed-of-origin (light blue strip).

LD blocks that appeared for both purebred and crossbred performance in the top 10 with most explained genetic variance, are shown per breed of origin in Table 4.6. Depending on the breed, the number of LD blocks from the top 10 that appeared for both purebred and crossbred performance, was 4 to 5 for BF, 3 to 6 for ADG, and at most one for RFI. For the LD blocks that appeared for both purebred and crossbred performance in the top 10, the percentage of genetic variance they explained for both purebred and crossbred performance was quite similar.

As LD blocks across breed of origin were not the same, because of different patterns of linkage disequilibrium, comparisons across breeds for crossbred performance were done regarding whether the top 10 LD blocks across breeds overlapped or were less than 1-Mb distance apart (Table 4.7). These regions can be observed in Figures 4.1-4.3 in light blue. From the top 10 LD blocks, at most, one region in common was observed between breeds for crossbred performance per trait. For both BF and ADG performance in crossbred, there were no common regions between the top 10 LD blocks from S breed of origin and the top 10 LD blocks from LR breed of origin.

A similar comparison was made across breeds for purebred performance. For BF, there was one common region between the top 10 LD blocks from S and LW and there were two common regions between the top 10 LD blocks from LR and LW. For ADG, there was one common region between the top 10 LD blocks from S and LR. For RFI, comparisons could be only made between S and LW because the SNP-allele effects for the LR population were not estimated, there was one common region between the top 10 LD blocks from S and LW. These regions can be observed in Figures 4.1-4.3 in light blue.

Table 4.6 LD blocks in common* between crossbred and purebred performance per breed-of-origin.

Trait	S_{PB}-S_{CB}			LR_{PB}-LR_{CB}			LW_{PB}-LW_{CB}		
	Position Chr:Mb	gVar PB	gVar CB	Position Chr:Mb	gVar PB	gVar CB	Position Chr:Mb	gVar PB	gVar CB
BF	SSC1:158,9 – 160,2	0.24	0.31	SSC8: 60,6 – 64,0	0.18	0.31	SSC2: 9,8 – 10,5	0.25	0.38
	SSC5: 30,1 – 30,8	0.35	0.39	SSC9: 0 – 0,6	0.17	0.32	SSC2: 144,6 – 144,7	0.25	0.29
	SSC6: 47,8 – 49,9	0.32	0.53	SSC9: 128,3 – 128,9	0.25	0.33	SSC2: 144,8 – 145,1	0.35	0.26
	SSC15: 102,3 – 104,5	0.64	0.59	SSC15: 119,3 – 119,8	0.21	0.30	SSC6: 77,5 – 78,4	0.31	0.37
	SSC18: 10,1 – 10,6	1.62	1.14				SSC11: 7,6 – 7,9	0.29	0.29
ADG	SSC1: 51,5 – 52,7	0.62	0.87	SSC1: 24,2 – 25,6	0.16	0.21	SSC1: 148,2 – 150,4	0.44	0.52
	SSC1: 50,8 – 51,4	0.67	0.71	SSC12: 16,5 – 17,3	0.37	0.17	SSC1:152,4 – 153,1	0.32	0.21
	SSC1: 53,8 – 54,1	0.46	0.34	SSC18: 9,8 – 9,9	0.20	0.17	SSC1: 154,1 – 155,5	0.63	0.37
	SSC9: 76,1 – 78,8	0.20	0.25				SSC1: 160,9 – 162,4	0.21	0.22
							SSC15: 132,5 – 132,9	0.23	0.30
							SSC18: 53,2 – 53,9	0.23	0.27
RFI	SSC14: 125,3 – 125,9	0.21	0.21				SSC1: 183,7 – 184,8	0.25	0.29

*Considering only the top 10 LD blocks explaining most of the genetic variance for purebred or crossbred performance.
gVar PB = percentage of genetic variance explained by a LD block for purebred performance.
gVar CB= percentage of genetic variance explained by a LD block for crossbred performance.
BF = back fat thickness, ADG = average daily gain, and RFI = residual feed intake.
S_{PB} = PB performance for S breed, LR_{PB} = PB performance for LR breed, and LW_{PB} = PB performance for LW breed.
S_{CB} = CB performance for S breed-of-origin, LR_{CB} = CB performance for LR breed-of-origin, and LW_{CB} = CB performance for LW breed-of-origin.

Table 4.7 LD blocks in common* across breed-of-origin for crossbred performance

Trait	S_{CB}-LR_{CB}		S_{CB}-LW_{CB}		LR_{CB}-LW_{CB}	
	Chr:Mb_S Chr:Mb_{LR}	gVar S gVar LR	Chr:Mb_S Chr:Mb_{LW}	gVar S gVar LW	Chr:Mb_{LR} Chr:Mb_{LW}	gVar LR gVar LW
BF			SSC11: 7,6 – 9.9	1.18	SSC15: 132,0 – 132,9	0.32
			SSC11: 7,6 – 9,4	0.58	SSC15: 130,9 – 131,4	0.20
ADG			SSC1: 158,9 – 160,2	0.33	SSC18: 54,3 – 54,7	0.15
			SSC1: 160,9 – 162,4	0.22	SSC18: 53,2 – 53,9	0.27
RFI	SSC14: 46,2 – 48,0	0.21	SSC7: 2,2 – 2,8	0.30	SSC1: 183,9 – 185,0	0.29
	SSC14: 45,8 – 47,8	0.24	SSC7: 2,6 – 3,0	0.20	SSC1: 183,7 – 184,8	0.29

*Because LD blocks are different between breed-of-origin, comparisons were done regarding whether the top 10 LD blocks explaining most of the genetic variance across breed-of-origin overlapped or were less than 1-Mb distance apart.
BF = back fat thickness, ADG = average daily gain, and RFI = residual feed intake.
S_{CB} = CB performance for S breed-of-origin, LR_{CB} = CB performance for LR breed-of-origin, and LW_{CB} = CB performance for LW breed-of-origin
gVar = percentage of genetic variance explained by a LD block according to the breed-of-origin.

4.3.3 Candidate genes

Putative candidate genes within the top 10 LD blocks either for purebred or crossbred performance and in the neighbouring upstream and downstream 1-Mb regions were identified based on the Sscrofa11.1 genome assembly and based on literature. The melanocortin 4 receptor (MC4R) was identified as a candidate gene for ADG and BF. The MC4R gene was previously associated with feed intake and growth rate in pigs, as well as with BF (Kim et al., 2000; Meidtner et al., 2006; Fan et al., 2010; Onteru et al., 2013). The MC4R gene controls energy balance (Seeley et al., 2004). MC4R are broadly distributed in the central neuronal system and an agonist stimulation at MC4R leads to a decrease in feed intake and loss of body weight (Seeley et al., 2004). The MC4R gene is located on SSC1 at 160,771,802 – 160,774,335 bp. For S, the gene was contained in an LD block located at 160.2 – 161.4 Mb. However, the LD block in the previous position (158.9 – 160.2 Mb) was in the top 10 with most explained genetic variance for crossbred performance for ADG and BF. This LD block explained a large variance for purebred performance for BF although it did not make it into the top 10 LD blocks. For LR, this region seems not to contain any QTL. For LW, the MC4R gene was located in an LD block located at 160.2 – 160.7 Mb. However, a second LD block, located immediately before (159.2 – 160.2 Mb) was in the top 10 with most explained genetic variance for purebred performance for ADG and BF. Additionally, a third LD block, located immediately after (160.9– 162.4 Mb) was in the top 10 with most explained genetic variance for both in purebred and crossbred performance for ADG.

The StAR-related lipid transfer domain containing 13 (STARD13) was identified as a candidate gene for BF. The STAR gene family is involved with lipids and lipid hormones binding to be exchanged between biological membranes (Thorsell et al., 2011). STARD13 seems to regulate FOS gene expression, which is a gene functionally related with intramuscular fatty acid composition (Puig-Oliveras et al., 2016). The STARD13 gene is located on SSC11 at 9,496,111 – 9,760,394 bp. For S, the gene was located in a LD block located at 8.9 – 9.9 Mb. This LD block was in the top 10 with most explained variance for crossbred performance for BF. Two contiguous LD blocks (7.6 -7.9 Mb and 8.0 – 8.8 Mb) were also in the top 10 with most explained variance for crossbred performance for BF. These three LD blocks explained a relatively large part of the variance for purebred performance for BF although they did not make it to the top 10 LD blocks. For LR, this region does not seem to contain any QTL. For LW, the STARD13 gene overlapped one LD block (9.5 – 9.7 Mb). However, the LD blocks in the previous positions (7.6 – 7.9 Mb and 8.0 –

9.4 MB) were in the top 10 with most explained variance for BF performance in purebred and crossbred, and crossbred, respectively.

The porcine insulin-like growth factor binding protein (IGFBP-5) was also identified as a candidate gene for BF. IGFBP-5 is a focal regulatory factor during the development of several key cell types, e.g., myoblasts and neural cells (Salih et al., 2004). The IGFBP-5 gene might be involved in intramuscular fat development in cattle (Wang et al., 2009), and was also associated with fat deposition in pigs (Fan et al., 2009). The IGFBP-5 gene is located on SSC15 at 118,860,219 – 118,879,384 bp. For S, this region does not seem to contain any QTL. For LR, the gene was contained in a LD block located at 118.6 – 118.9 Mb. However, the LD block in a following position (119.3 – 119.8 Mb) was in the top 10 with most explained variance for purebred and crossbred performance for BF. For LW, the gene was contained in a LD block located at 118.8 – 119.0 Mb. However, the LD block in the previous position (118.2 – 118.8 Mb) was in the top 10 with most explained variance for crossbred performance for BF.

We did not identify any candidate gene for RFI. For RFI, there are few GWAS studies in pigs and they all revealed different regions associated with this trait (Fan et al., 2010; Gilbert et al., 2017; Onteru et al., 2013; Do et al., 2014). RFI is a complex trait and the biology behind it seems difficult to unravel, as we were unable to find LD blocks explaining a large percentage of genetic variance or patterns across purebred and crossbred performance within the same breed.

4.3.4 MC4R

From all evaluated candidate genes, only for the MC4R gene the underlying causal mutation is known. Allele frequencies of this MC4Rsnp were quite similar between observed frequencies in purebred compared to crossbred pigs, but considerable differences were observed between breeds within the purebred or between breeds-of-origin within the crossbred (Table 4.8). In the S population and among alleles originating from S in the crossbred population, the m allele is highly prevalent (0.81-0.85), whereas in the LR population or among alleles originating from LR in the crossbred population, the m allele is almost absent (0.06-0.11).

Table 4.8 Frequency of MC4Rsnp alleles* in purebreds and in crossbreds (CB) within breed of origin.

	m	w
Purebred		
S	0.85	0.15
LR	0.06	0.94
LW	0.39	0.61
CB, breed of origin†		
CB,S	0.81	0.19
CB,LR	0.12	0.88
CB,LW	0.44	0.56

*m is associated with the mutant allele and allele w is associated with the wild allele of MC4R.

†Expressed as frequency within each breed of origin.

For S breed of origin, the MC4Rsnp was in LD with 31 flanking loci, which resulted in a LD block from 158.9 to 161.5 Mb. For LR breed of origin, the MC4Rsnp was in LD with 49 flanking loci, which resulted in a LD block from 158.8 to 163.3 Mb. For LW breed of origin, the MC4Rsnp was in LD with 42 flanking loci, which resulted in a LD block from 158.9 to 162.6 Mb. For comparison across breed of origin, we only considered the overlapping SNPs across the three LD blocks which resulted in a block of 31 SNPs (158.9 - 161.5 Mb). It is worthwhile noting that this MC4R based block contains the LD block spanning 158.9-160.2 Mb that was identified to be associated with ADG and BF in S and the LD block spanning 159.2-160.2 Mb associated with ADG and BF in LW. The block contained 74 different haplotypes, each unique haplotype was always exclusively co-segregating either the m or w allele of MC4Rsnp. The only exception was a haplotype that was observed in 83 crossbred pigs originating from S, in 260 crossbred pigs originating from LR, and in 1993 crossbred pigs originating from LW. This haplotype carried the m allele for all these crossbred pigs, except for two who received the haplotype from S and carried the w allele. These two cases, however, may simply be genotyping errors and were not used further for the MC4R analysis. Therefore, after including the MC4Rsnp in the LD block we still observed 74 different haplotypes. From the 74 haplotypes, 44 were observed in the S breed of origin, 19 in the LR breed of origin and 31 in the LW breed of origin.

In Figure 4.4, the effect of each haplotype that co-segregates with the MC4R gene is shown per breed of origin for crossbred performance for ADG. Within breed of origin, haplotypes co-segregating with the m allele had different effects compared

to haplotypes co-segregating with the w allele (T-test, P-value <0.05). Haplotypes co-segregating with the m allele, in general, had a positive effect, while haplotypes co-segregating with the w allele had a negative effect. Effects of specific haplotypes were similar if they originated from the S or the LW population, however, their effects were smaller if they originated from the LR population (paired T-test, P < 0.05). For each breed the average effects of the m and w allele, weighted according to the haplotype frequencies, are shown as red numbers in Figure 4.4. The difference of the averages is an approximation of the allele substitution effect, substituting an m allele for a w allele has an effect on ADG of -2.5 g/d, -0.5 g/d and -1.6 g/d, when the allele originates from S, LR, or LW, respectively. Using the MC4Rsnp itself, the effect of substituting an m allele for a w allele at MC4Rsnp was -22.60 g/d, -14.21 g/d, or -21.67 g/d, when the allele originated from S, LR or LW, respectively. Figure 4.5 shows the number of times each haplotype was observed per breed of origin versus its effect on crossbred performance for ADG. For S breed of origin, there is one very common haplotype accounting for 73% of the observations and this haplotype had the largest effect (+1.52 g/d) among all the haplotypes in this LD block. For LR breed of origin, the 19 haplotypes observed had small effects, from -0.40 to +0.54 g/d, and the most common haplotype accounted for 37% of the observations and had an effect of -0.11 g/d. For LW breed of origin, the haplotypes had more variable estimated effects, and the most common haplotype accounted for 28% of the observations and had an effect of -1.16 g/d.

4.4 Discussion

The objective of this study was to show how the effect of SNP-alleles, estimated in a genomic prediction model using commonly used SNP panels, varies when observed in different genetic backgrounds. With crossbreeding, the effects of SNP-alleles can be observed both against purebred and crossbred background. Moreover, the degree of allelic differentiation among the three populations estimated with Weir and Cockerham's F_{ST} was previously estimated by Sevillano et al. (2017) and were equal to 0.17 between S and LR, 0.12 between S and LW, and 0.14 between LW and LR, which indicates that they are distantly-related breeds. Since the three breeds are distantly-related, the effects of the SNP-alleles is expected to vary in the three distinctive backgrounds provided by each of the breeds-of-origin.

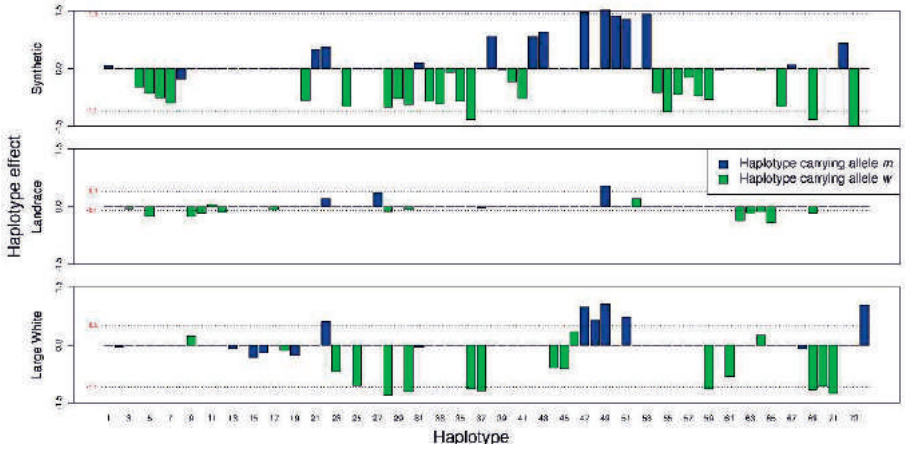

Figure 4.4. Haplotypes effect on average daily gain (g/d) per breed of origin. From the 74 haplotypes observed in the LD block associated with the MC4R gen. Average effects of the m and w allele, weighted according to the haplotype frequencies, are shown as red numbers.

To observe the estimated effects of SNP-alleles for crossbred performance in different genetic backgrounds, we traced the breed of origin of alleles in crossbred animals and estimated breed-specific SNP effects from the solutions of a BOA model for three traits. For traits with low heritability (<0.20) and low r_{pc} (<0.70) the BOA model tended to show better predictive abilities (Sevillano et al., 2017). Therefore, based on the heritability and r_{pc} estimates with pedigree information from Godinho et al. (2018), BF, ADG and RFI, were chosen to be studied. Only for RFI, the estimated heritability for crossbred performance differed from the expected value of ~0.2 (Godinho et al., 2018) as it was considerably higher (0.40) in our data. Genetic parameters estimated for LR pigs had high standard errors because of the limited number of RFI records, therefore, GEBVs of LR pigs for purebred performance were not further used in this study. For all the other traits, estimates of r_{pc} and heritability for crossbred performance were as expected.

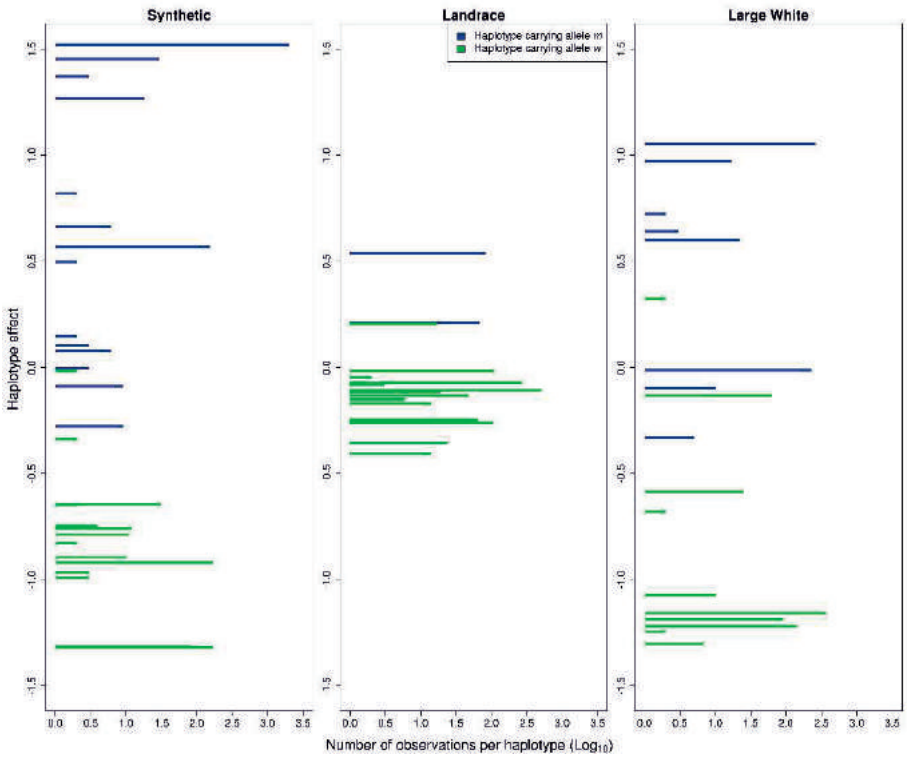

Figure 4.5. Number of observations (Log10) of each of the 74 haplotypes. Number of observations are presented according to the effect of the haplotype on average daily gain (g/d). From the 74 haplotypes observed in the LD block associated with the MC4R gen.

4.4.1 Proportion of genetic variance explained by a region

The proportion of total genetic variance explained was calculated for each LD block instead of reporting effects of single SNPs. The LD blocks built based on alleles originating from the S paternal population were, on average, longer than the LD blocks built based on alleles originating from the maternal LR and LW populations. This is in line with linkage disequilibrium estimations made by Veroneze et al. (2014) using the same populations as in our study, where they showed that the S population showed the highest level of linkage disequilibrium, followed by LW, and LR having the lowest level of linkage disequilibrium. Their populations named SL2, DL1 and DL2 correspond to S, LR and LW populations in the present study.

4.4.1.1 Regions associated with purebred and crossbred performance

Within the same breed of origin, we observed some LD blocks that appeared for both purebred and crossbred performance in the top 10 with most explained genetic variance. Across traits, this number of common LD blocks in the top 10 is expected to be related to the r_{pc} for that trait, as the correlation between allele substitution effects of the causal variants of two traits is expected to be the same as the genetic correlation between two traits (Wientjes and Calus, 2017). Our results are in line with this, as RFI showed the lowest r_{pc} (0.37-0.60) and had at most only one LD block that appeared for both purebred and crossbred performance, while BF showed the highest r_{pc} (0.71-0.89) and had 4 to 5 LD blocks that appeared for both purebred and crossbred performance.

For LD blocks that appeared for both purebred and crossbred performance in the top 10 with most explained variance, we observed that they explained a similar percentage of additive genetic variance. Despite the fact that percentages of additive genetic variance were quite similar, differences in allele frequencies between purebred and crossbred can explain r_{pc} values below unity. However, as shown in Table 4.8, allele frequency of the MC4Rsnp between purebred and crossbred were quite similar. One of the possible reasons for r_{pc} values below unity is the presence of genotype by environment interactions (GxE) (Wientjes and Calus, 2017; Godinho et al., 2018). GxE might have been present because some purebred pigs were housed in high-health status farms (nucleus farms, free of a number of specific diseases), while some crossbred pigs were housed in experimental farms with environmental conditions similar to commercial farms with these specific diseases prevalent. Another environmental difference between purebred and crossbred is that trait measurement methods were different (Wientjes and Calus, 2017; Godinho et al., 2018), as explained earlier in the methods section. ADG and BF were measured in a different way for purebred and crossbred pigs, and as these are the components traits for RFI, RFI was also derived differently for purebred and crossbred. It is unclear to which extent the genetic ranking is affected by these differences in measurements. Nevertheless, using crossbred information in the training set avoids that the difference in measurements affects the breeding decisions.

4.4.1.2 Regions associated with crossbred performance by breed

Next to the comparison between purebred and crossbred background, comparison across breed of origin backgrounds was also performed. For all traits, there was at most one region in common between breeds-of-origin for crossbred performance. This indicates that the proportion of genetic variance for crossbred performance explained by a genomic region depends upon the breed of origin. Differences in genetic variance across breeds of origin can be due to differences in allele frequencies that affect the contribution of dominance effects to the additive genetic variance. The allele frequency of the MC4Rsnp (see Table 4.8) is quite different in crossbred pigs depending upon the breed of origin. In addition, for the block co-segregating with the MC4R gene originated from the LR population, we observed a relatively small variance among the effect sizes of the different haplotypes, caused by the low frequency of the m allele of MC4Rsnp (Figure 4.5). For S and LW, we observed that the haplotypes in this region had a larger variance of effect size for ADG performance in crossbred, because the MAF of the MC4Rsnp was considerably higher. We hypothesize that for other genomic regions, similar differences in MAF may be one important source of differences in how the genetic variance is distributed across the genome for different breeds, and therefore having different contribution to genomic prediction.

We also observed that the effect of a haplotype associated with crossbred performance is different depending upon from which population it originates. In the case of MC4R, identical haplotypes co-segregating uniquely either with the mutant or the wild type allele, yielded different effects for LR compared to S and LW (Figure 4.4). Similarly, the effect difference between haplotypes co-segregating with the m allele and the w allele was five and three times larger for haplotypes originating from S and LW compared to haplotypes originating LR, respectively (Figure 4.4). Differences in haplotype effects across breed of origin can be due to differences in linkage between the haplotype and any QTL in the vicinity, however, this was not the case for MC4R. Another reason for these differences in haplotypes effects across breed of origin might be that the haplotypes are not identical between the breeds, they only appear to be so due to the genotype resolution used. If that is the case, the difference can be due to distinct interactions of the MC4R allele with different local genetic background, i.e. epistasis (Mackay, 2014); or because the unobserved differences between the haplotypes directly give rise to additional additive effects. So, what is observed as a breed of origin effect may actually be different haplotypes which can be only differentiated with a higher

density genotype. However, when we estimated the allele substitution effect of the MC4Rsnp itself, we still observed that the largest effect was when the allele originated from S, followed by LW origin, while alleles from LR origin had the smallest effect. But the magnitude was much larger than when we approximate the allele substitution effect from the haplotypes estimates. These differences might arise from the methodology as SNP effects in the haplotypes were estimated jointly as random effects via BLUP, being subjected to considerable shrinkage, whereas MC4Rsnp effects were estimated using fixed regression. For the MC4Rsnp we can conclude that the main difference across breeds are the allele frequencies which can reflect selection pressure for other performance traits, as also observed by Kim et al. (Kim et al., 2000).

In general we observed few regions strongly associated with ADG, RFI, or BF for crossbred performance, and these are mainly breed-specific. Conversely, we observed many regions that did not have a large effect on ADG, RFI, or BF for crossbred performance. Hypothesizing that only for regions with large effect breed-specific modelling is beneficial, using SNP effects averaged across breeds may be more realistic than considering breed-specific SNP effects. We previously compared the BOA model, which considers breed-specific SNP effects in crossbred animals, to a model that does not account for breed-specific SNP effects in crossbred animals (Sevillano et al., 2017), and found similar or slightly higher accuracies of estimated breeding values with the BOA model. This suggests that few regions, such as the region containing the MC4R, may benefit from accounting for breed-specific SNP effects.

4.5 Conclusions

Some similar regions explaining similar additive genetic variance were observed across purebred and crossbred performance. The number of similar regions was related to the trait r_{pc}. Observed r_{pc} values below one can be due to differences in housing and trait measurements between purebred and crossbred as they can affect the genetic ranking. Therefore crossbred information is valuable in the training set to account for the environmental background differences between crossbred and purebred performance.

Moreover, there was some overlap across breeds of origin between regions that explained relatively large proportions of genetic variance for crossbred performance of ADG, RFI, and BF; albeit that the actual proportion of variance

deviated across breeds-of-origin. This variation is due to differences in allele frequencies across population and epistasis can be also playing a role. Results based on a missense mutation in MC4R confirmed that even if a causal locus has similar effects across breeds-of-origin, estimated effects and explained variance in its region estimated using a genomic prediction model relying on a SNP panel can strongly depend on the allele frequency of the underlying causal mutation.

These results are valuable to understand the limited benefit obtained when predicting breeding values of purebred animals for crossbred performance with models that account for breed-specific effect of alleles, as the BOA model, compared to a model using crossbred information but without accounting for breed-specific effect of alleles. However, selecting important regions associated with crossbred performance and differentiating their SNP-allele effects according to their breed of origin, might improve prediction models for crossbred performance.

Chapter 5

Genomic evaluation for a crossbreeding system implementing breed of origin for targeted markers

Claudia A. Sevillano[1,2], Henk Bovenhuis[1], Mario P.L. Calus[1]

[1] Wageningen University & Research Animal Breeding and Genomics, 6700AH, Wageningen, the Netherlands; [2] Topigs Norsvin Research Center, 6640AA, Beuningen, the Netherlands

Submitted

Abstract

When using crossbred genomic information for estimating breeding values of purebred individuals for crossbred performance, it is important to realize that the genome in crossbred animals is a mosaic of genomic regions inherited from the different parental breeds. We previously showed that effects of haplotypes strongly associated with crossbred performance are different depending upon from which parental breed they are inherited, however, the majority of the genomic regions are not or only weakly associated with crossbred performance. Therefore, our objective was to develop a model that distinguishes between selected single nucleotide polymorphisms (SNP) strongly associated with crossbred performance and remaining SNP. For the selected SNPs, breed-specific allele affects were fitted whereas for the remaining SNP it was assumed that effects are the same across breeds (SEL-BOA model). We used data from three purebred populations; S, LR, and LW, and one commercial crossbred population (S x (LR x LW) or S x (LW x LR)). We selected SNPs that explained together either 5% or 10% of the total crossbred genetic variance for average daily gain in each breed of origin. The model was compared to the BOA model (allowing all SNP-alleles to have a different effect for crossbred performance depending upon the breed of origin) and a G model (all SNP-alleles having the same effect for crossbred performance across breeds). Across the models, the heritability for crossbred performance was very similar with values of 0.29-0.30. Across the breeds, the estimate of the r_{pc} increased by 21.5% with the BOA compared to the G model. With the SEL-BOA models, in general, the r_{pc} for the selected SNPs was larger than for the non-selected SNPs. For breed LR, the r_{pc} for non-selected SNP and selected SNP estimated with the SEL-BOA 5% and SEL-BOA 10% were very different compared to the r_{pc} estimated with the G or BOA model. For breeds S and LW, there was not a big discrepancy for the r_{pc} estimated with the SEL-BOA models and with the G or BOA model. Differences of prediction accuracies between models were small, but there was a tendency that the SEL-BOA model performed better than the other models. We conclude that the BOA model calculates more accurate breeding values of purebred animals for crossbred performance than the G model when r_{pc} differs (≈10%) between the G and the BOA model. Superiority of the SEL-BOA model compared to the BOA model was only observed for scenario SEL-BOA 10% and when r_{pc} for the non-selected and selected SNP differed both (≈20%) from the r_{pc} estimated by the G or BOA model.

Key words: origin of alleles, crossbred, genomic prediction, finisher, pig

5.1 Introduction

The breeding goal of pig breeding programs is commonly to select purebred animals for improved performance of their crossbred descendants. It has been shown that using crossbred information, in addition to commonly used purebred information, improves the accuracy of selection. The benefit was observed using crossbred phenotypes either with pedigree (Wei and Van der Steen, 1991) and even more pronounced with crossbred genomic information (Xiang et al., 2017; Sewell et al., 2018). The most common genetic markers used for genomic selection are single nucleotide polymorphisms (SNPs), i.e. bi-allelic markers. For crossbred animals, as their genome is a mosaic of genomic regions inherited from the different parental breeds, depending from which breed a SNP-allele was inherited from, it might have different effects. These different allele effects can arise because: (1) quantitative trait loci (QTL) may be in linkage disequilibrium with different single nucleotide polymorphisms (SNP) depending from which parental breed the QTL was inherited (Lopes, 2016), (2) partly different quantitative trait nucleotides (QTN) could be underlying a QTL in different parental breeds, while the common QTN may have different minor allele frequencies (MAF) in the parental breeds, with the extreme case where it is not segregating in one or more breeds (Wientjes et al., 2015), (3) epistatic interactions may be differ between parental breeds (Mackay, 2014). In most previous studies using crossbred genomic information potential differences in SNP-allele effects due to the breed of origin were ignored (e.g., Hidalgo et al., 2015; Veroneze et al., 2015; Sewell et al., 2018). A model that accounts for breed of origin of alleles (BOA model), has been proposed by Dekkers (2007), Ibáñez-Escriche et al. (2009) and Christensen et al. (2014). The BOA model was expected to be beneficial when using commercial crossbred genomic information for estimation of breeding values of purebred pigs for crossbred performance. The observed benefits of the BOA model, however, were limited to traits with low genetic correlation between purebred and crossbred performance (r_{pc}) and for crossbred populations that originated from distantly-related breeds, as was shown in studies with simulated two-way (Ibáñez-Escriche et al., 2009; Esfandyari et al., 2015) and three-way crossbred data (Ibáñez-Escriche et al., 2009) and in studies with real two-way (Xiang et al., 2016) and three-way crossbred data (Sevillano et al., 2017).

The BOA model allows all SNP-alleles to have a different effect for crossbred performance depending upon the breed of origin. In a recent study, Sevillano et al. (2018a) confirmed that the effect of haplotypes strongly associated with crossbred

performance are different depending upon from which population they originate. It was also shown, however, that the majority of the genomic regions are not or only weakly associated with crossbred performance. We hypothesized that targeting genomic regions strongly associated with crossbred performance and differentiating their SNP-allele effects according to their breed of origin, might improve prediction models for crossbred performance. Therefore, the objective of this study was to develop a model that accounts for breed-specific allele effects only for SNPs strongly associated with crossbred performance, and for the rest of the SNPs assumes that effects are the same across breeds. Thus, the model had one across-breed component, and a breed-specific component for each breed of origin. The performance of this model, in terms of estimated variances for the different model components and overall prediction accuracy, was tested using combined information from both purebred and three-way commercial crossbred pigs for average daily gain. The model was compared to the BOA model (allowing all SNP-alleles to have a different effect for crossbred performance depending upon the breed of origin) and a G model (all SNP-alleles having the same effect for crossbred performance across breeds).

5.2 Methods

5.2.1 Data

The data consisted of three purebred pig populations; Synthetic boar (S), Landrace (LR), and Large White (LW), and one commercial crossbred population (S x (LR x LW) or S x (LW x LR)). All pigs were genotyped using one of the three following SNP panels: Illumina PorcineSNP60.v2 BeadChip (60K.v2), Illumina PorcineSNP60 BeadChip (60K), or Illumina PorcineSNP10 BeadChip (10K). Pigs genotyped with the 60K or 10K chips were imputed to the 60K.v2 panel using FImpute Version 2.2 software (Sargolzaei et al., 2014) with default parameter settings and using pedigree information. The imputation strategy was similar to Sevillano et al. (2016), where each of the three purebred populations, LR, LW, and S, were imputed in two steps: (1) pigs genotyped with the 10K chip were imputed to 60K, and (2) all pigs with 60K data (imputed or genotyped) were imputed to 60K.v2. For the commercial crossbred population, imputation was done in a single step, commercial crossbred pigs genotyped with the 10K chip were directly imputed to 60K.v2, because all ancestors were genotyped or already imputed to 60K.v2.

Purebred pigs were located in nucleus farms while crossbred pigs were located in experimental farms representative of commercial production conditions.

Phenotypes for average daily gain (ADG) were measured in most of the purebred and commercial crossbred pigs. ADG for purebred pigs was calculated as the difference of on-test body weight at an average age of 60 days and off-test body weight at an average age of 173 days divided by the number of days. ADG for commercial crossbred pigs was calculated as the difference of on-test body weight at an average age of 70 days of age and body weight at end of the finishing period, which was on average 120 kg, divided by the number of days.

The numbers of available genotypes and phenotypes were 7575, 3288 and 12 794 for purebred population S, LR and LW respectively, and 2816 for the commercial crossbred population. For all pigs, four generations of pedigree information were included for analysis.

5.2.2 Proposed model

The proposed model considers breed-specific effects only for SNP strongly associated with performance in crossbred, and for the remaining SNP assumes that effects are the same across breeds. To build this model, we first needed to determine the breed of origin of alleles in crossbred pigs and secondly, determine which SNP are strongly associated with crossbred ADG. In this section, we will firstly introduce the proposed model, followed by a subsection "Inference of the breed of origin of alleles" where we explain how we determined the breed of origin of alleles in crossbred pigs, and we finish with a subsection "Targeting SNP" where we explain how we determine which are the SNP strongly associated with ADG performance in crossbred pigs. Hereafter, we will refer to the SNP strongly associated with crossbred performance as "selected SNP" and to the remaining SNP as "non-selected SNP".

5.2.2.1 The model

To model breed-specific effects for SNPs strongly associated with crossbred performance and across-breed effects for all other SNPs, the following 4-trait animal model was fitted (SEL-BOA model):

$$\mathbf{y_S} = \mathbf{X_S b_S} + \mathbf{W_S u_S} + \mathbf{Z_S a_S} + \mathbf{e_S},$$

$$\mathbf{y_{LR}} = \mathbf{X_{LR} b_{LR}} + \mathbf{W_{LR} u_{LR}} + \mathbf{Z_{LR} a_{LR}} + \mathbf{e_{LR}},$$

$$\mathbf{y_{LW}} = \mathbf{X_{LW} b_{LW}} + \mathbf{W_{LW} u_{LW}} + \mathbf{Z_{LW} a_{LW}} + \mathbf{e_{LW}},$$

$$\mathbf{y_{CB}} = \mathbf{X_{CB} b_{CB}} + \mathbf{W_{CB} u_{CB}} + \mathbf{Z_{CB}} \mathbf{g}_{\mathbf{CB}}^{(\mathbf{S})} + \mathbf{Z_{CB}} \mathbf{g}_{\mathbf{CB}}^{(\mathbf{LR})} + \mathbf{Z_{CB}} \mathbf{g}_{\mathbf{CB}}^{(\mathbf{LW})} + \mathbf{Z_{CB} a_{CB}} + \mathbf{e_{CB}},$$

where $\mathbf{y_S}$, $\mathbf{y_{LR}}$, $\mathbf{y_{LW}}$, and $\mathbf{y_{CB}}$ are the vectors of the phenotypes for S, LR, LW, and commercial crossbred pigs, respectively; $\mathbf{b_S}$, $\mathbf{b_{LR}}$, $\mathbf{b_{LW}}$, $\mathbf{b_{CB}}$ represent the vectors of fixed effects farm*breed*sex and birth weight as covariable and $\mathbf{X_S}$, $\mathbf{X_{LR}}$, $\mathbf{X_{LW}}$, $\mathbf{X_{CB}}$ are the respective incidence matrices relating pig phenotypes to fixed effects; $\mathbf{u_S}$, $\mathbf{u_{LR}}$, $\mathbf{u_{LW}}$, $\mathbf{u_{CB}}$ represent the vectors of random common litter effects, and $\mathbf{W_S}$, $\mathbf{W_{LR}}$, $\mathbf{W_{LW}}$, $\mathbf{W_{CB}}$ are the respective incidence matrices relating pig phenotypes to litter effects; $\mathbf{a_S}$, $\mathbf{a_{LR}}$, $\mathbf{a_{LW}}$, are the vectors of additive genetic effects in PB, $\mathbf{g_{CB}^{(S)}}$, $\mathbf{g_{CB}^{(LR)}}$, $\mathbf{g_{CB}^{(LW)}}$ are the vectors of the additive genetic effect of PB gametes in commercial crossbreds due to the selected SNPs, $\mathbf{a_{CB}}$ is the vector of additive genetic effect in commercials crossbred considering only the non-selected SNPs, and $\mathbf{Z_S}$, $\mathbf{Z_{LR}}$, $\mathbf{Z_{LW}}$, $\mathbf{Z_{CB}}$ are the respective incidence matrices. Finally, $\mathbf{e_S}$, $\mathbf{e_{LR}}$, $\mathbf{e_{LW}}$, $\mathbf{e_{CB}}$ represent the vectors of random residual effects. The variance-covariance of the common litter effect and residual effect were:

$$\mathrm{Var}\begin{bmatrix}\mathbf{u_S}\\ \mathbf{u_{LR}}\\ \mathbf{u_{LW}}\\ \mathbf{u_{CB}}\end{bmatrix} = \begin{bmatrix} \mathrm{I}_S\sigma^2_{u_S} & 0 & 0 & 0\\ 0 & \mathrm{I}_{LR}\sigma^2_{u_{LR}} & 0 & 0\\ 0 & 0 & \mathrm{I}_{LW}\sigma^2_{u_{LW}} & 0\\ 0 & 0 & 0 & \mathrm{I}_{CB}\sigma^2_{u_{CB}}\end{bmatrix},$$

$$\text{and } \mathrm{Var}\begin{bmatrix}\mathbf{e_S}\\ \mathbf{e_{LR}}\\ \mathbf{e_{LW}}\\ \mathbf{e_{CB}}\end{bmatrix} = \begin{bmatrix} \mathrm{I}_S\sigma^2_{e_S} & 0 & 0 & 0\\ 0 & \mathrm{I}_{LR}\sigma^2_{e_{LR}} & 0 & 0\\ 0 & 0 & \mathrm{I}_{LW}\sigma^2_{e_{LW}} & 0\\ 0 & 0 & 0 & \mathrm{I}_{CB}\sigma^2_{e_{CB}}\end{bmatrix}.$$

The variance-covariance of additive genetic effect for breed S origin based on selected SNPs was:

$$\mathrm{Var}\begin{bmatrix}\mathbf{a_S}\\ \mathbf{a_{CB}^{(S)}}\\ \mathbf{g_S}\\ \mathbf{g_{CB}^{(S)}}\end{bmatrix} = \begin{bmatrix}\sigma^2_{a_S} & \sigma_{a_S,g_S}\\ \sigma_{g_S,a_S} & \sigma^2_{g_S}\end{bmatrix} \otimes \mathbf{G^{(S)}} = \begin{bmatrix}\sigma^2_{a_S} & \sigma_{a_S,g_S}\\ \sigma_{g_S,a_S} & \sigma^2_{g_S}\end{bmatrix} \otimes \begin{bmatrix}\mathbf{G_{S,S}} & \mathbf{G_{S,CB}^{(S)}}\\ \mathbf{G_{CB,S}^{(S)}} & \mathbf{G_{CB,CB}^{(S)}}\end{bmatrix},$$

where breed S pigs have additive effects based on selected SNPs, $\mathbf{a_S}$ for purebred performance and $\mathbf{a_{CB}^{(S)}}$ for crossbred performance. The commercial crossbred pigs have additive effects based on selected SNPs for the breed S gametes, $\mathbf{g_{CB}^{(S)}}$ for crossbred performance and $\mathbf{g_S}$ for purebred performance. This last effect, $\mathbf{g_S}$, is an artificial random vector that is added to be able to define the variance-covariance

of additive genetic effects with the above Kronecker product, but does not have practical relevance (Christensen et al., 2015). The matrix $\mathbf{G}^{(\mathbf{S})}$ is a breed-specific partial relationship matrix for breed S which contains four blocks, one within S pigs ($\mathbf{G}_{\mathbf{S,S}}$), two for S with commercial crossbred pigs ($\mathbf{G}^{(\mathbf{S})}_{\mathbf{S,CB}}$ and $\mathbf{G}^{(\mathbf{S})}_{\mathbf{CB,S}}$), and one within commercial crossbred pigs ($\mathbf{G}^{(\mathbf{S})}_{\mathbf{CB,CB}}$).

The variance-covariance structures for breeds LR and LW are defined similarly, and the three variance-covariance structures are assumed independent, i.e. no covariances are considered between S, LR, and LW effects (Christensen et al., 2015). There are six selected SNPs genetic variance components, one for purebred and one for crossbred performance for each breed of origin, and three covariance components, one for each breed of origin. To construct the three breed-specific partial relationship matrices, $\mathbf{G}^{(\mathbf{S})}$, $\mathbf{G}^{(\mathbf{LR})}$, and $\mathbf{G}^{(\mathbf{LW})}$, we used the breed of origin of phased alleles in commercial crossbred pigs. Then, the breed-specific partial relationship submatrices are defined as, e.g. breed S origin:

$$\mathbf{G}_{\mathbf{S,S}} = \left(\mathbf{M}^{\mathbf{S}}_{\mathbf{sel}} - 2\mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right)\left(\mathbf{M}^{\mathbf{S}}_{\mathbf{sel}} - 2\mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right)'(\mathrm{F}^{\mathrm{S}})^{-1},$$

$$\mathbf{G}^{(\mathbf{S})}_{\mathbf{S,CB}} = \left(\mathbf{M}^{\mathbf{S}}_{\mathbf{sel}} - 2\mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right)\left(\mathbf{M}^{\mathbf{CB(S)}}_{\mathbf{sel}} - \mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right)'(\mathrm{F}^{\mathrm{S}})^{-1}, \text{ and}$$

$$\mathbf{G}^{(\mathbf{S})}_{\mathbf{CB,CB}} = \left(\mathbf{M}^{\mathbf{CB(S)}}_{\mathbf{sel}} - \mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right)\left(\mathbf{M}^{\mathbf{CB(S)}}_{\mathbf{sel}} - \mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right)'(\mathrm{F}^{\mathrm{S}})^{-1},$$

where $\mathbf{M}^{\mathbf{S}}_{\mathbf{sel}}$ is a matrix containing breed-specific allele content of selected SNPs for purebred S pigs (coded as 0, 1, or 2). $\mathbf{M}^{\mathbf{CB(S)}}_{\mathbf{sel}}$ is a matrix containing breed S allele content of selected SNPs for commercial crossbred pigs (coded as 0, or 1), so that alleles not assigned to breed S as breed of origin were set to missing, meaning that they had an entry of zero in the centred matrix represented by $\left(\mathbf{M}^{\mathbf{CB(S)}}_{\mathbf{sel}} - \mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right)$ and therefore effectively did not contribute to the computed breed S partial relationship; $\mathbf{p}^{\mathbf{S}}$ is the vector of breed S specific frequencies of the counted allele (p_j^S), where p_j^S was calculated across S and commercial crossbred pigs by counting the occurrences of alleles originating from the S breed and coded as 1, divided by the total number of S alleles in the S and commercial crossbred pigs on locus j. Finally, the scaling factor was defined as $\mathrm{F}^S = \sum_j 2p_j^S(1 - p_j^S)$. The breed-specific partial relationship submatrices $\mathbf{G}^{(\mathbf{LR})}$ and $\mathbf{G}^{(\mathbf{LW})}$ are defined similarly to $\mathbf{G}^{(\mathbf{S})}$. However, the entries of the $\mathbf{M}^{\mathbf{CB(LR)}}_{\mathbf{sel}}$ matrix containing the breed LR allele content for commercial crossbred pigs are set to a missing value if the origin of the allele

corresponds to the other maternal line, and effectively does not contribute to the breed-specific partial relationship matrix for LR.

For additive genetic effects in commercial crossbred pigs based on non-selected SNP we did not model breed-specific allele effects and therefore this was defined by one vector, $\mathbf{a_{CB}}$. The variance-covariance matrix of non-selected SNP genetic effect was:

$$\mathrm{Var}\begin{bmatrix}\mathbf{a_S}\\ \mathbf{a_{LR}}\\ \mathbf{a_{LW}}\\ \mathbf{a_{CB}}\end{bmatrix} = \begin{bmatrix}\sigma^2_{a_S} & \sigma_{a_S,a_{LR}} & \sigma_{a_S,a_{LW}} & \sigma_{a_S,a_{CB}}\\ \sigma_{a_S,a_{LR}} & \sigma^2_{a_{LR}} & \sigma_{a_{LR},a_{LW}} & \sigma_{a_{LR},a_{CB}}\\ \sigma_{a_S,a_{LW}} & \sigma_{a_{LR},a_{LW}} & \sigma^2_{a_{LW}} & \sigma_{a_{LW},a_{CB}}\\ \sigma_{a_S,a_{CB}} & \sigma_{a_{LR},a_{CB}} & \sigma_{a_{LW},a_{CB}} & \sigma^2_{a_{CB}}\end{bmatrix} \otimes \mathbf{G_{non-sel}}.$$

The genomic relationship matrix ($\mathbf{G_{non-sel|sel}}$) was constructed using the first method in VanRaden (2008):

$$\mathbf{G_{non-sel}} = (\mathbf{M_{non-sel}} - 2\mathbf{1p}')(\mathbf{M_{non-sel}} - 2\mathbf{1p}')'/\mathrm{F}^{-1},$$

where $\mathbf{M_{non-sel}}$ is a matrix containing non-selected SNP genotypes for each pig (coded as 0, 1, or 2), $\mathbf{p}$ is the vector of the frequencies of the counted allele (p_j) calculated across the entire genotyped population, and the scaling factor was defined as $\mathrm{F} = \sum_j 2p_j(1-p_j)$.

The SEL-BOA model was implemented in the MiXBLUP software (Ten Napel et al., 2016). To estimate de variance components we used the same SEL-BOA model in the MTG2 software (Lee and Van der Werf, 2016).

5.2.2.2 Inference of the breed of origin of alleles

To infer the breed of origin of alleles in crossbred pigs we used the BOA approach developed by Vandenplas et al. (2016) using the parameter settings recommended by Sevillano et al. (2016). The BOA approach consisted of three steps : (1) Phasing the haplotypes of both purebred and commercial crossbred pigs with AlphaPhase1.1 software (Hickey et al., 2011). Phasing was performed using pedigree, and using nine combinations of haplotype length and each combination was run both considering “Offset” and “NotOffset” modes, the “Offset” mode shifts the start of the cores to halfway along the first core, creating 50% overlaps between cores. These settings allowed each allele to be considered 18 times through different haplotypes of variable length. (2) Determining the unique haplotypes among the purebred pigs. For assigning a breed of origin to a

haplotype, at least 80% of its copies were required to be observed in a specific breed. (3) Assigning the breed of origin for each allele carried on the haplotypes of commercial crossbred pigs based on the knowledge of the breed of origin of the haplotypes, on the zygosity (i.e., homozygosity or heterozygosity) of the locus, and on the breed composition of the crossbred. Alleles that were not assigned a breed of origin were set to missing. SNPs for which the paternal or maternal allele were assigned a breed of origin in less than 90% of the cases were removed. Commercial crossbred pigs with assigned breed of origin for less than 90% of their genome were removed. If an allele was observed less than 5 times in one of the three breed of origin in the purebred populations or in the commercial crossbred population, the corresponding SNP was also removed from the final set of SNPs. The final SNP set for subsequent analyses consisted of 41,529 SNPs. All populations were analyzed with the same set of SNPs.

5.2.2.3 Targeting SNPs

Estimates for breed-specific SNP allele substitution effects were obtained from Sevillano et al. (2018a) where they used a genomic BLUP with breed-specific partial relationship matrices (BOA model) (Sevillano et al., 2017). With this approach, genomic estimated breeding values (GEBV) for crossbred performance were calculated, and afterwards converted to SNP-allele effects by breed of origin. The BOA model allows all SNPs to have breed-specific alleles. Therefore it is similar to the SEL-BOA, however, for each breed the BOA-model only has the breed-specific component. GEBV of purebred pigs for crossbred performance were then converted to SNP-allele effects ($\hat{\mathbf{a}}_{\mathbf{CB}}$), e.g. for breed S:

$$\hat{\mathbf{a}}_{\mathbf{CB(S)}} = \mathbf{W}^{\mathbf{CB(S)}}\hat{\mathbf{b}}_{\mathbf{CB(S)}},$$

where $\mathbf{W}^{\mathbf{CB(S)}}$ contains centered genotypes and $\hat{\mathbf{b}}_{\mathbf{CB}}$ are allele substitution effects, which can be obtained respectively by:

$$\mathbf{W}^{\mathbf{CB(S)}} = \left(\mathbf{M}^{\mathbf{CB(S)}} - \mathbf{1}\mathbf{p}^{\mathbf{S}\prime}\right) \text{ and}$$

$$\hat{\mathbf{b}}_{\mathbf{CB(S)}} = \mathbf{W}^{\mathbf{CB(S)}\prime}(\mathbf{W}^{\mathbf{CB(S)}}\mathbf{W}^{\mathbf{CB(S)}\prime})^{-1}\hat{\mathbf{a}}_{\mathbf{CB(S)}} = (\mathrm{F}^{\mathrm{S}})^{-1}\mathbf{W}^{\mathbf{CB(S)}\prime}(\mathbf{G}^{(\mathbf{S})}_{\mathbf{CB,CB}})^{-\mathbf{1}}\hat{\mathbf{a}}_{\mathbf{CB(S)}}.$$

SNP-allele effects for crossbred performance of the other purebred populations were calculated similarly.

Afterwards, Sevillano et al. (2018a) calculated the proportion of variance explained by a group of SNPs in nonrandom association, called LD blocks (see Sevillano et al.,

2018a for details on how LD blocks were built). In a GBLUP model, all SNPs are considered simultaneously in the model, therefore, the effect of a QTL is likely distributed across all SNPs that have a nonrandom association with the QTL. For this reason, it is recommended to calculate the proportion of variance explained by a group of SNPs in nonrandom association instead of reporting effects of single SNPs (Lopes, 2016). LD blocks were built per breed of origin, therefore, nonrandom association between alleles at two loci was tested in the commercial crossbred population between all pair of loci coming from the same breed of origin. Percentage of genetic variance explained by the *i*-th LD block was calculated as in Wang et al. (2014):

$$\frac{\mathrm{Var}_{(a_i)}}{\sigma_a^2} \times \frac{x_n}{n} \times 100\% = \frac{\mathrm{Var}(\sum_{j=1}^{n} \mathbf{z}_j \hat{\alpha}_j)}{\sigma_a^2} \times \frac{x_n}{n} \times 100\%,$$

where a_i is the genetic value of the *i*-th LD block, σ_a^2 is the total genetic variance, $\mathbf{z}_j$ is a vector of gene content of the *j*-th SNP for all purebred individuals of the same breed, $\hat{\alpha}_j$ is the estimated effect of the *j*-th SNP within the *i*-th LD block that contains *n* SNPs, and $x_n$is the mean number of SNPs across LD blocks. The factor $\frac{x_n}{n}$ adjust explained variances for the number of SNPs included in the LD block.

For selecting SNP to be consider to have breed-specific allele effects, we took the top LD blocks that explained together at the most either 5 or 10 % of the total genetic variance in each breed of origin. Selected LD blocks per breed of origin were merged in one group and all the SNPs in each of the selected LD blocks were then classified as selected SNPs so their effects would be estimated in the SEL-BOA model as breed-specific. The non-selected SNPs were assumed to have the same effect across the three breeds of origin, as outlined before. The SEL-BOA model was then ran twice, considering 5 and 10% of all SNPs as selected SNPs (SEL-BOA 5% and SEL-BOA 10% models).

5.2.3 Cross-validation

5.2.3.1 Comparison of models

For comparison to the SEL-BOA model, we also calculated GEBV of purebred pigs for crossbred performance using the BOA model (allowing all SNP-alleles to have a different effect for crossbred performance depending upon the breed of origin) and a G model (all SNP-alleles having the same effect for crossbred performance across breeds).

5.2.3.2 Training set

The accuracy of GEBV of purebred pigs for crossbred performance from all models was evaluated as the average accuracy obtained from 4-fold cross-validation. Because of different degrees of relationship between purebreds and commercial crossbred pigs, each of the four populations were first divided into four mutually exclusive clusters, using the K-means clustering method applied to a dissimilarity matrix computed from elements of the **G** matrix (Saatchi et al., 2011). The commercial crossbred pigs were not evenly distributed across the four clusters, therefore the clusters were reorganized to contain each more or less ¼ of the commercial crossbred pigs with the closest relationship (i.e. highest average relationship) based on the **G** matrix. Then, within each breed, each of the four crossbred clusters was assigned to one of the four purebred clusters with the closest relationship (i.e. highest average relationship) based on the **G** matrix to form a fold. Therefore, each fold contains one purebred cluster and one crossbred cluster. This way, for each breed, we obtained four folds to be included in the cross-validation.

In each training analysis, the data excluded phenotypes of purebred and commercial crossbred pigs from one fold to train on the remaining three folds to predict GEBV for crossbred performance of the excluded purebred pigs (validation set). This resulted in every purebred pig having GEBV for crossbred performance that were obtained without using performance of the most closely-related commercial crossbred pigs for training. Thus, the information coming from the most closely-related commercial crossbred pigs could be used for validation. The number of pigs in the validation and training sets for each of the folds of the cross-validation are in Table 5.1.

5.2.3.3 Validation set

For the purebred pigs used for the validation, some sort of phenotype is needed to be able to compute the prediction accuracy. Purebred pigs cannot have an own performance for crossbred performance. In our data they did not have large offspring groups, needed to compute average offspring performance as an accurate phenotype. Therefore, we calculated deregressed proofs (DRP) for purebred pigs within the validation sets to validate the predictions of our models. For this, first we obtained estimated breeding values (EBV) from the 4-trait model with a pedigree-based relationship matrix. This resulted in an EBV for crossbred performance for each purebred pig. The EBV were estimated based on performance of the commercial crossbred pigs assigned to each of the validation

folds (Table 5.1). Within each validation fold, the EBV of purebred pigs for crossbred performance were then deregressed according to Calus et al. (2016). The deregression involved removal of all effects of relatives in the same validation set, and correction for regression to the mean, to obtain a more accurate estimate of the expected phenotype. In addition, a weighting factor (*w*) was estimated for each DRP value based on the reliability of the calculated DRP. These *w* are the effective record contributions (Přibyl et al., 2013), and reflect the amount of information in the DRP contributed by the animal's crossbred relatives, correcting for any information of the crossbred relatives of other purebred animals that contributed to its EBV before deregression.

Table 5.1 Cross-validation strategy for performance of average daily gain in crossbred

Fold	Training				Validation			
	S	LR	LW	CB	S	LR	LW	CB
1	5365	2624	9061	2112	2183	665	3738	704
2	5771	2329	8194	2117	1777	960	4605	699
3	6017	2188	10327	2109	1531	1101	2471	707
4	5491	2726	10815	2110	2057	562	1980	706

Numbers of individuals for Synthetic boar (S), Landrace (LR), Large White (LW), and three-way crossbred (CB) pigs.

5.2.3.4 Predictive ability

Accuracies of all models were calculated as the weighted correlation between the DRP and the GEBV of purebred pigs for crossbred performance, where the weighting factor *w* was used to account for differences in the amount of available information on relatives to estimate DRP. The standard error (SE) of the correlations were approximated as $(1-\mathrm{r}^2)/\sqrt{\mathrm{N}}$, were r is the estimated correlation of the model, and N is the number of validation animals (Stuart and Ord, 1994).

5.3 Results

5.3.1 Targeted SNPs

We selected the top LD blocks that explained together either 5% or 10% of the total crossbred genetic variance for ADG in each breed of origin using the BOA model that treats all SNPs in the same way. For the 5% scenario, for breed S origin there were 18 LD blocks which included in total 428 SNPs; for breed LR origin there were 41 LD blocks which included in total 661 SNPs, and for breed LW origin there were 26 LD blocks which included in total 524 SNPs. These three groups of selected LD

blocks per breed of origin were merged in one group resulting in 1498 SNPs classified as selected SNPs. These selected SNPs represent 3.6% of the whole SNP panel. The numbers of selected SNPs by breed of origin and the overlap between them are illustrated in Figure 5.1A. For the 10% scenario, for breed S origin, there were 66 LD blocks which included in total 1554 SNPs; for breed LR origin, there were 109 LD blocks which included in total 1512 SNPs, and for breed LW origin, there were 73 LD blocks which included in total 1131 SNPs. These three groups of selected LD blocks per breed of origin were merged in one group resulting in 3809 SNPs classified as selected SNPs. These selected SNPs represent 9.2% of the whole SNP panel. The numbers of selected SNPs by breed of origin and the overlap between them are illustrated in Figure 5.1B.

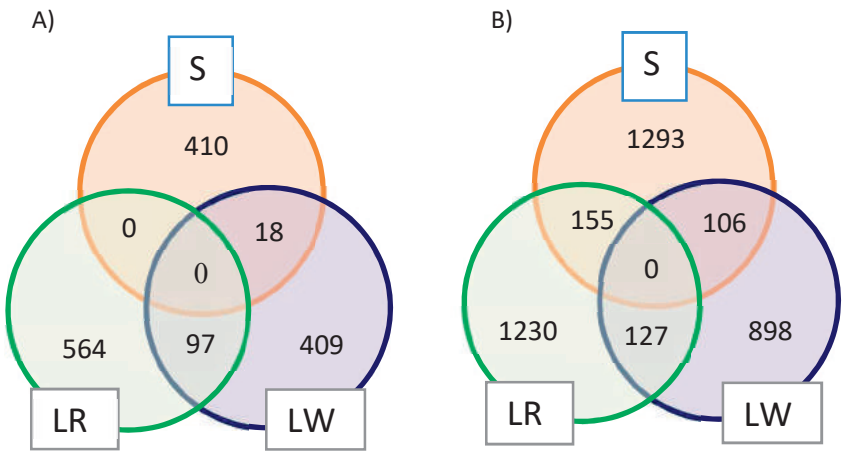

Figure 5.1 Numbers of selected SNPs by breed of origin and the overlap between them. (A) for scenario 5% and (B) for scenario 10%.

5.3.2 Variance components, heritabilities, and genetic correlations

Estimated variance components for ADG using the G, BOA, SEL-BOA 5% and SEL-BOA 10% models are in Table 5.2. The standard errors of the estimated variance components in Table 5.2 are provided in Table S5.1 [see Supplementary material, Additional file S5.1]. In the SEL-BOA 5% model, the selected SNPs explained 39%,

43%, and 40% of the total crossbred genetic variance for S, LR, and LW, respectively. And for the SEL-BOA 10% model, the selected SNPs explained 77%, 75%, and 79% of the total crossbred genetic variance for S, LR, and LW, respectively.

Comparing purebred variance components across models, additive genetic variances were larger when estimated with the G model and smaller when estimated with the BOA model, while with the SEL-BOA models they were in between, but in general, estimates were similar across models. Likewise, heritability estimates were similar across models, around 0.17, 0.23, and 0.22 for S, LR, and LW, respectively. For the SEL-BOA, in this comparison the considered additive variance was obtained as the sum of the variance explained by the selected and non-selected SNPs.

Comparing crossbred variance components across models, additive genetic variances were very similar across G (2284), BOA (2285) and SEL-BOA 5% (2280) models, while the SEL-BOA 10% model had a slightly larger additive variance (2349). For the BOA and SEL-BOA models, in this comparison the considered additive variance was obtained as the weighted sum of the variance explained by the selected and non-selected SNPs, using weights of 0.50 for the paternal breed, and 0.25 for the maternal breeds. Crossbred heritabilities were similar across models (0.29-0.30).

Comparing crossbred genetic variance components by breed of origin, we observed similar estimates independent of the model used for breed S origin, however, for breed LR and LW origin, the estimates differed largely according to the model. The genetic correlations between performance of purebred and crossbred pigs (r_{pc}) estimated with the G model did not differs largely to the r_{pc} estimated with the BOA model for breed S origin, but a larger difference was observed for the maternal breeds LR and LW, however, differences between the models were within the range of the standard errors. With the SEL-BOA models, the r_{pc} for the selected SNPs was larger than for the non-selected SNPs, except for breed S origin when calculated with the SEL-BOA 5% which was zero. For breed LR origin, the estimate of the r_{pc} for the selected SNPs was larger than unity, for further analysis we fixed the value to 0.99. Similar, for breed LR origin, the r_{pc} for the non-selected SNPs calculated with the SEL-BOA 10%, had a value lower than zero and large SE (±0.31), for further analysis we fixed it to zero. For LR breed, the r_{pc} for non-selected SNP and selected SNP estimated with the SEL-BOA 5% and SEL-BOA 10% were very

different compared to the r_{pc} estimated with the G or BOA model. For S and LW breeds, there was not a big discrepancy for the r_{pc} estimates across models, except for the r_{pc} of zero estimated for the selected-SNPs with SEL-BOA 5% model. Crossbred heritability estimates for the SEL-BOA 5% were higher for the non-selected SNPs (0.17) than for the selected SNPs (0.12). Conversely, crossbred heritability estimates for SEL-BOA 10% were lower for the non-selected SNPs (0.07) than for the selected SNPs (0.23).

Table 5.2 Additive genetic variance (σ_a^2), litter variance (σ_u^2), residual variance (σ_e^2), and heritabilities for each breed for purebred (PB) and crossbred (CB) performance, and genetic correlation between PB and CB performance (r_{PC}), estimated using the G[a], BOA[b], and SEL-BOA[c] models.

		Non-sel	Sel			Non-sel	Sel	Non-sel	Sel			Non-sel	Sel	Non-sel	Sel
G	S	2686		3205	9271	0.18		2284		1404	4168	0.29		0.66	
	LR	2005		2501	4085	0.23								0.44	
	LW	2320		2278	5664	0.23								0.49	
BOA	S	2212		3204	9282	0.15		2224		1304	4118	0.30*		0.73	
	LR	1912		2503	4076	0.23		1806						0.56	
	LW	2135		2276	5669	0.21		2883						0.62	
SEL-BOA	S	2064	410	3230	9310	0.14	0.03	1357	866	1363	4154	0.17	0.12*	0.73	0.00
5%	LR	1301	583	2529	4134	0.15	0.07		1041					0.08	1.10
	LW	1904	325	2306	5676	0.19	0.03		920					0.51	0.75
SEL-BOA	S	1596	798	3200	9349	0.11	0.05	543	1777	1370	4088	0.07	0.23*	0.59	0.85
10%	LR	897	1003	2513	4142	0.10	0.12		1631					-0.25	1.14
	LW	1648	592	2307	5674	0.16	0.06		2039					0.61	0.81

S = Synthetic boar, LR = Landrace (LR), LW = Large White (LW).

[a]G model, model for across-breed effects for all SNPs

[b]BOA model, model for breed-specific effects for all SNPs.

[c]SEL-BOA model, model with breed-specific effects for SNPs strongly associated with crossbred performance and across-breed effects for all other SNPs. SEL-BOA (5%) and SEL-BOA (10%) considering top 5% or top 10% of the SNPs associated with crossbred performance as strongly associated with crossbred performance, respectively.

*SEL-BOA model has two estimates for $\boldsymbol{\sigma^2_{a_{PB}}}$, $\boldsymbol{\sigma^2_{a_{PB}}}$, and $\mathbf{r_{pc}}$, one for the across-breed component (Non-sel) and the other for breed-specific component (Sel)

*$(0.5\sigma^2_{a_S} + 0.25\sigma^2_{a_{LR}} + 0.25\sigma^2_{a_{LW}})/(0.5\sigma^2_{a_S} + 0.25\sigma^2_{a_{LR}} + 0.25\sigma^2_{a_{LW}} + \sigma^2_{u_{CB}} + \sigma^2_{e_{CB}})$

5.3.3 Predictive ability for breeding values

Accuracies of the four models for GEBV of purebred pigs for crossbred performance for ADG are in Table 5.3. In general the differences between the models were small, but there was a tendency that the SEL-BOA model performed better than the other models. For the paternal breed S and maternal breed LR, highest prediction accuracies were obtained with the SEL-BOA 10% model, followed by similar accuracies obtained with SEL-BOA 5%, BOA and G model. For the maternal breed LW, similar accuracies were obtained with the four models.

Table 5.3 Accuracies* of G[a], BOA[b], and SEL-BOA[c] models calculated for estimating breeding values of purebred pigs for crossbred performance for each of the four folds of cross-validation and average weighting factor (*w*) of the calculated DRP per validation fold

Folds	*w*	G	BOA	SEL-BOA 5%	SEL-BOA 10%
Synthetic boar					
1	0.15	0.132	0.118	0.116	0.129
2	0.18	0.024	0.030	0.026	0.038
3	0.23	0.119	0.119	0.112	0.145
4	0.09	0.092	0.092	0.092	0.100
Mean		*0.092*	*0.090*	*0.086*	*0.103*
Landrace					
1	0.26	0.140	0.177	0.110	0.152
2	0.25	0.166	0.167	0.153	0.204
3	0.21	0.111	0.105	0.183	0.175
4	0.24	0.172	0.163	0.173	0.202
Mean		*0.147*	*0.153*	*0.155*	*0.183*
Large White					
1	0.20	0.150	0.163	0.158	0.149
2	0.16	0.138	0.133	0.133	0.135
3	0.11	0.123	0.135	0.118	0.114
4	0.21	0.149	0.160	0.154	0.155
Mean		*0.140*	*0.148*	*0.141*	*0.134*

*Accuracies measured as weighted correlation between DRP and EBVs

**Approximate standard errors computed as $(\mathbf{1}-\mathbf{r^2})/\sqrt{\mathbf{N}}$, were equal to 0.035 to 0.036 for the mean accuracies across the folds and for all methods

[a]G model, model for across-breed effects for all SNPs. [b]BOA model, model for breed-specific effects for all SNPs. [c]SEL-BOA model, model with breed-specific effects for SNPs strongly associated with crossbred performance and across-breed effects for all other SNPs. SEL-BOA 5% and SEL-BOA 10% considering top 5% or top 10% of the SNPs associated with crossbred performance as strongly associated with crossbred performance, respectively.

5.4 Discussion

The objective of this study was to develop a model that accounts for breed-specific allele effects only for SNPs strongly associated with crossbred performance, and for the rest of the SNPs assumes that effects are the same across breeds. To construct the relationship matrices for this SEL-BOA model, we selected SNPs that explained together at the most either 5 or 10 % of the total genetic variance in each breed of origin using the BOA model. In the SEL-BOA 10% model the selected SNP actually explained 77%, 75%, and 79% of the total additive genetic variance for S, LR, and LW, respectively. This shows that the SEL-BOA model was really able to attribute much more genetic variance to the selected SNP than the original BOA model, where all SNP were treated similarly in the model. These high percentages of explained variance left little crossbred additive genetic variance to be explained by the non-selected SNPs, so we did not pursue any scenarios that selected even more SNPs.

Across the models, the heritability for crossbred performance was very similar. However, the models using breed of origin of alleles (BOA, SEL-BOA 5%, and SEL-BOA 10%) showed that estimates of crossbred additive genetic variance differed between the three breeds. This suggests that the G model, on average, obtains the correct heritability, even if the contribution to the crossbred variance of the different breeds varies. In theory, the crossbred additive variance components estimated with the BOA model comprises the variance observed in crossbred pigs due only to the alleles coming from the analyzed breed. This implies that the breed-specific r_{pc} values estimated with the BOA model are effectively correlations of effects on purebred and crossbred performance of alleles originating from the same breed, while the G model estimates one r_{pc} value considering effects of alleles originating from all breeds involved. Therefore, r_{pc} are expected to be higher when calculated with the BOA model rather than the G model, and this is also what we observed in our estimates. For breed S, estimated crossbred genetic variance and r_{pc} were very similar between the G and BOA model, and no benefit for calculating GEBV of S purebred animals for crossbred performance was observed using the BOA model. However a benefit was observed for breeds LR and LW that showed larger differences in their estimates of crossbred genetic variance and r_{pc} between the G and BOA model. Similar results were found by Sevillano et al. (2017) who used similar but smaller data sets.

With the SEL-BOA models, crossbred genetic variation is modelled separately for non-selected and selected SNP, where for the last component breed of origin specific effects are estimated. This has potentially two advantages, arising from

having separate variance components for the selected and non-selected SNPs. Firstly, the model is able to assign more variance to SNPs with a strong association to the trait than the G and BOA models, and less to the non-selected SNPs. Secondly, it can differentiate the r_{pc} values for the two categories of SNPs. Differences in variance estimates alone are not sufficient to cause a difference in accuracy, the benefit of the SEL-BOA model comes when r_{pc} estimates are also different. For instance, for breeds LR and LW, the crossbred genetic variance estimated for non-selected and selected SNP estimated with the SEL-BOA 5% and SEL-BOA 10% were very different compared to the crossbred genetic variance estimated with the G or BOA model. However, for LW, there were not large differences across the estimates of r_{pc}, subsequently, no benefit of the SEL-BOA models were observed. Conversely, for LR, the r_{pc} for non-selected and selected SNP estimated with the SEL-BOA 5% and SEL-BOA 10% were very different compared to the r_{pc} estimated with the G or BOA model. The estimated r_{pc} for the selected SNPs was greater than one, and we assumed a value of 0.99 in the subsequent analyses, meaning that their estimated effects are similar for purebred and crossbred performance. On the other hand, the r_{pc} for the non-selected SNPs was below zero, and we assumed a value of zero in the subsequent analyses. This means that their estimated effects for purebred and crossbred performance are totally different, and using crossbred information is needed for estimating effects for crossbred performance as it cannot be derived from purebred information. As a result, SEL-BOA models were more accurate for calculating GEBV of LR purebred animals for crossbred performance than the BOA or G models.

For breed S, similar to breed LW, accuracies for calculating GEBV of S purebred animals for crossbred performance were similar between the SEL-BOA models and the other models. For these breeds, there was not a big discrepancy for the r_{pc} estimates, except for the r_{pc} estimated for the selected-SNPs with SEL-BOA 5% model. In this case, however, the impact might not be so high because the selected SNPs only represented 39% of the crossbred genetic variance, therefore the main genetic variance was due to the non-selected SNPs that have an r_{pc} that was close to the estimates of the BOA and G models. Therefore, for these cases r_{pc} might have been more precisely estimated for each group of alleles with the SEL-BOA models, such as the slightly observed benefit with the SEL-BOA 5% suggests. The differences were however small, which may in part be because the SEL-BOA models actually may have had lower power than the G model because of the larger number of effects fitted. In general this is a problem that is faced by all models

using the concept of breed of origin of alleles (Ibánẽz-Escriche et al., 2009; Vandenplas et al., 2017).

Although with the SEL-BOA 5% the selected SNPs explained 39%, 43%, and 40% of the total crossbred genetic variance for S, LR, and LW, respectively, this model performed similar to the G model for S and LW. For LR, allowing the 1498 selected SNPs to have a different effect rather than effects estimated combining the other breeds S and LW, improved accuracy. An important question is why LR did seem to benefit from using the SEL-BOA model, while S and LW did not. It is good to note that the S breed was created as a combination of Large White and Pietrain, which suggests that the S and LW breed, a Large White based dam line, are somehow related. On the other hand LR is a Landrace based dam line and LR pigs have undergone a different selection pressure that may have shaped their genomic architecture differently, possibly resulting partly in different haplotypes, and different haplotypes frequencies for the haplotypes that are in common with the other breeds (Egbert Knol, personal communication). In a previous study, Sevillano et al. (2018a) observed that the explained genetic variance of haplotypes associated to the MC4R gene, which has a missense mutation with a known effect on ADG (Kim et al., 2000), was considerably lower for the LR and also this breed showed the lowest allele frequency of the mutation compared to breeds S and LW. This seems to confirm that the LR breed indeed is quite different from the S and LW breeds. Similar to the MC4R, other regions coming from the LR breed might also show different genetic variances compared to S and LW, providing a possible explanation why this breed shows some benefit when some SNP effects are estimated separately by breed of origin in the SEL-BOA 5%. With the SEL-BOA 10%, the benefit for LR breed is even larger. With the SEL-BOA 10% model the benefit of the BOA model is obtained while reducing possible disadvantages due to calculating three times as many effects, because breed of origin specific effects are estimated for fewer SNPs.

5.5 Conclusions

The BOA model was more accurate for calculating GEBV of purebred animals for crossbred performance than the G model when estimated crossbred genetic variances and r_{pc} differed largely between the G model and the BOA model. Superiority of the SEL-BOA model compared to the BOA model was only observed for the SEL-BOA model 10% when r_{pc} for the non-selected SNP and selected SNP differed strongly from the r_{pc} estimated by the BOA model.

Chapter 6

General discussion

6.1 Introduction

The breeding goal for pig-breeding companies is to improve the performance of crossbred animals. To realize this breeding goal, it is important to estimate, as correctly as possible, the potential of purebred animals to produce crossbred offspring with superior performances, i.e. their estimated breeding value (EBV) for crossbred performance. Nowadays, EBVs are usually calculated using genomic information, which results in genomic estimated breeding values (GEBV). GEBV can be computed by multiplying the allele substitution effects of single-nucleotide polymorphisms (SNP) with the animal's genotype. These SNP allele effects are estimated using a reference population consisting of animals with both phenotypes and genomic information. Including the genomic information on crossbred pigs in the reference population allows for the estimation of SNP allele effects, while accounting for the crossbred genetic background, which might improve the accuracy of the GEBV. At the start of this project, the impact of using crossbred genomic information in the reference population in order to predict the crossbred performance of purebred animals was largely unknown. In this thesis, I have studied how to handle crossbred genomic information in prediction models, and, more precisely, I have proposed two approaches: 1) the BOA model – a model that estimates SNP allele effects for crossbred performance, specifically for each breed of origin that occurs in the crossbreds, and 2) the SEL-BOA model – a model that estimates SNP allele effects for crossbred performance based on their breed of origin, only for SNPs that are strongly associated with crossbred performance, while for the remaining SNPs, the model assumes identical allele effects, irrespective of origin. In this last chapter of the thesis, I will start by discussing the relevance of using crossbred information in reference populations. Subsequently, I will discuss approaches to determine the breed of origin of alleles, which is an important prerequisite for implementing prediction models that consider SNP alleles effects to be breed-specific. Furthermore, I synthesize the most important results of the two proposed prediction models. I then propose further improvements to these models and discuss the challenges for their practical implementation in breeding programs. Finally, I propose an alternative model for handling crossbred genomic information. In each section, I will discuss the current state of knowledge, explore how the work described in this thesis contributes to the current knowledge, identify remaining gaps in the knowledge, and then suggest avenues for future research.

6.2 Relevance of crossbred information for breeding programs

In the late 1980s, it was already recognized that the information collected on purebred animals in nucleus farms did not fully reflect the performance of crossbred animals in commercial circumstances (Merks, 1989). Genetic correlation between the performance of purebred and crossbred animals (r_{pc}) is an indicator of the correlation between what breeders do in isolation and what farmers need in their commercial farm. An r_{pc} lower to unity might be caused by differences between purebred and crossbred animals in environmental and additive genetic variance. In order to take into account these differences and deal directly with a r_{pc} lower to unity, a selection strategy called combined crossbred and purebred selection (CCPS) was proposed (Wei and van der Werf, 1994). In this strategy, the nucleus purebred population can be directly selected for performance in commercial circumstances without actually being housed there. Having purebred populations housed in commercial environments is not in the interest of the breeders, as the commercial circumstances where crossbred animals need to perform are not homogeneous, as I will discuss in the next paragraph. This will mean that many nucleus purebred populations will be needed.

Since commercial farms are located all over the world, commercial conditions can vary, due to different climates, feed, and diseases. As an example, figure 6.1 presents the growth curves of three-way commercial crossbred gilts housed in two different commercial farms, and on one farm, the gilts were fed three different diets. This graph shows how performance can differ between farms, but also within farms. If sires had offspring in these four environments, the EBV of the sires for crossbred performance might change according to where the data of their offspring, which is used for evaluation, comes from, i.e. in which farm and under which feed regimen. Differences in EBV can cause a re-ranking of the sires. Such re-ranking is known to be caused by an interaction between the genetics of an animal and the environment (GxE). In this example, the possible ranking difference in the farm using different diets could be caused by the different feeding, however, the causes of the possible ranking difference between the two farms are unknown. My study Sevillano et al. (2018b), which evaluated the effect of feeding two different diets to genetically similar growing-finishing full-sib gilts and boars, revealed that a cereals-alternative ingredients diet improved the ratio of protein to lipid deposition, compared to a corn-soybean meal diet. In a further study, Godinho et al. (2018) proved that this difference in lipid deposition was partially explained by a

genotype-by-feed interaction, a special type of GxE, which caused a re-ranking of genotypes and heterogeneity of genetic variance. The presence of GxE across commercial farms will not be further discussed here, as it goes beyond the scope of this thesis. However, it is important to realize that commercial circumstances where crossbred animals need to perform are not homogeneous, and, in global breeding programs, the commercial farms used for data collection should reflect the average commercial farm. Alternatively, for local breeding programs, the data should come from specific commercial farms, in order to rank the sires according to their EBV in local circumstances.

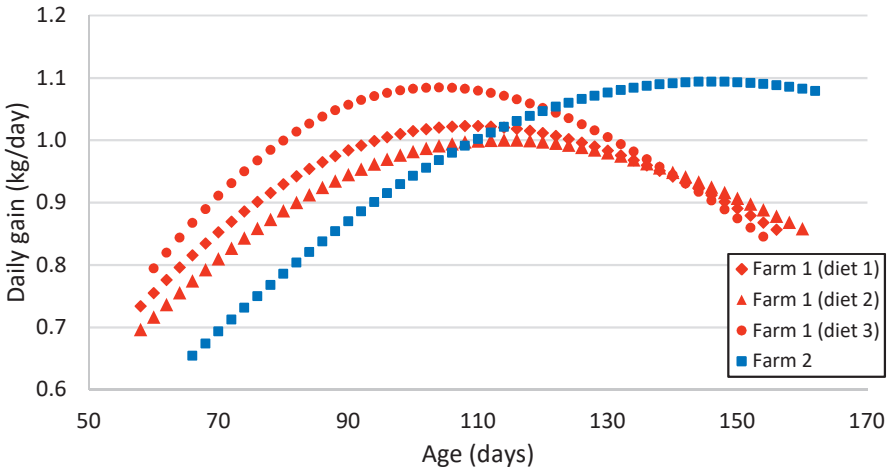

Figure 6.1. Growth curve of three-way commercial crossbred gilts housed on two different farms, i.e. Farm 1 (red) and Farm 2 (blue ■). Farm 1 fed three different diets i.e. diet 1 (red ♦), diet 2 (red ▲), and diet 3 (red ●). There were 402 gilts on Farm 1 and 447 gilts on Farm 2. On Farm 1, there were 209, 117, and 76 gilts on diet 1, diet 2, and diet 3, respectively.

Historically, most pig-breeding companies have evolved in the following way: in the 1960s, when crossbreeding systems were widely adopted, selection was solely based on purebred phenotypic information. Since the 1990s, selection has included crossbred phenotypic information (CCPS), and currently, genomic selection includes purebred phenotypic, genomic information, and crossbred phenotypic information. Moving from selection being based solely on purebred phenotypic information to CCPS was shown to improve genetic response for two-way crossbred animals, using the selection index theory (Wei and van der Werf, 1994). Even higher responses were observed using genomic information. In purebred

animals, adding genomic information was shown to improve genetic response in purebred animals (Forni et al., 2010; Knol et al., 2016; Lopes et al., 2018). Following these developments in selection strategies, it can be expected that including crossbred genomic information could lead to further improvements. As genomic selection approaches rely directly, or indirectly, on the estimation of SNP allele substitution effects, including crossbred animals in the reference population means that SNP allele substitution effects are being estimated for commercial conditions.

6.2.1 Relevance of crossbred genomic information

Wei and van der Werf (1994) and Bijma and Van Arendonk (1998) identified the potential of using crossbred information for traits presenting a r_{pc} lower than 0.8. Two studies using two-way crossbred animals showed that training with crossbred genomic information, instead of purebred genomic information, improved the prediction of the performance of crossbred individuals when r_{pc} was smaller than 0.8 (Esfandyari et al., 2015; Hidalgo, 2015). However, breeding companies are interested in the outcome of using crossbred genomic information to predict the EBV of purebred individuals for crossbred performance, as that is their breeding goal. So far, there is only one study that evaluates the outcome of training with genomic information from crossbred pigs on the EBV of purebred individuals for crossbred performance (Sewell et al., 2018). This study was performed with a small dataset of three-way crossbred pigs. They observed that the accuracy of the EBV of purebred individuals for crossbred performance in average daily gain (ADG), loin depth (LD), and back fat thickness (BF) increased by 0.1 when including crossbred genomic information in the reference population, in comparison to when only purebred genomic information was included. Since training with crossbred pigs is expected to result in better accuracy, and because genotyping an individual is still becoming cheaper, I expect that, in the near future, genomic selection in pig-breeding companies will include genomic information from crossbred pigs. Still, results using a larger dataset to show the added value of using crossbred genomic information in the reference population are required.

In order to fill this gap, I included an analysis using a large dataset from a commercial sire line, the same synthetic boar line from which data was used for the analyses in **chapters 2-5**, and their three-way crossbred offspring. This dataset gives the opportunity to evaluate the added value of using crossbred information in the reference population with greater power because of the number of animals. There were 593 sires with phenotypic and genomic information, 47,181 purebreds, and 53,625 commercial crossbred offspring with phenotypic information. The

accuracy of the EBV of purebred individuals for crossbred performance for ADG, average daily feed intake (ADFI), BF, and LD were calculated using different information sources in the reference population: 1) only purebred phenotypes, 2) purebred and crossbred phenotypes, 3) purebred phenotypes, genotypes, and crossbred phenotypes, and 4) purebred and crossbred phenotypes and genotypes. To test the predictive ability of the models, I masked the phenotypes from 142 youngest sires and the phenotypes of their 15,137 commercial crossbred offspring. The number of individuals in the reference and validation sets are presented in Table 6.1. For the 142 validation sires, I calculated the average of the individual crossbred offspring deviation (IODs). On average, the validation sires had 107 crossbred offspring. EBVs and IODs were estimated using a multi-trait model on MiXBLUP software (Ten Napel et al., 2016), and the single-step approach (Aguilar et al., 2010) was used to combine the data across both genotyped and non-genotyped individuals.

Table 6.1 Number of phenotypes and genotypes available in the reference and validation sets.

	Reference		Validation	
Population*	Phenotype	Genotype	Phenotype	Genotype
Sires	451	451	142	142
PB_{off}	47,181	6594	0	0
CB_{off}	38,488	2996	15,137	1190

*PB_{off} = Purebred offspring, CB_{off} = Commercial crossbred offspring. To improve estimations, I also included 1,099,397 phenotypes and 46,730 genotypes of other dam lines used, in order to produce the crossbred offspring with the evaluated sire line, as well as other sire lines that were kept in the same farms as the evaluated sire line, and their respective crossbred offspring (both in the reference and validation set).

The prediction accuracy of the four reference populations is shown in Figure 6.2. When the information for training only comes from purebred phenotypes (reference population 1), the accuracy of the EBV of purebred pigs for crossbred performance is affected by the trait's r_{pc}. The r_{pc} of all traits were relatively high and comparable (0.75 to 0.88). After including commercial crossbred phenotypic information (reference population 2), a slight increase of EBV accuracies (3-7%) was observed for ADFI, BF, and LD, but not for ADG, which actually had a decrease in accuracy (-5%). To better understand why including crossbred phenotypes had such a low impact, I also calculated the accuracy in a reference population with only

crossbred phenotypes, and the accuracies were 0.10, 0.36, 0.27, and 0.21 for ADG, ADFI, BF, and LD respectively. Even though crossbred phenotypes in this data are useful for predicting ADFI, BF, and LD, for some reason, this is not reflected when using reference population 2. Most of the crossbred data came from CCPS farms, meaning that crossbred animals were housed in the same environment as purebred animals. Therefore, adding crossbred phenotypes did not add information about the environmental background. This may be why I did not see a bigger positive impact when using crossbred phenotypes, beyond the benefit of using a larger training population per se. After including purebred genomic information in the reference population, prediction accuracy increased drastically (31-62%) for ADFI, BF, and LD, but only 17% for ADG. The accuracy increased another 2-19% after including commercial crossbred genomic information. Therefore, including genomic information, either from purebred or crossbred animals, seems to always be beneficial for prediction accuracy. As accuracies are defined for crossbred performance, I also tested a reference population where phenotypes were only collected in crossbred animals, keeping genomic information in both populations. This reference population is interesting, largely because more animals per generation were available to phenotype. However, the accuracy decreased to the level of the reference population 3, except for ADFI. To conclude, including genomic information from both purebred and crossbred was beneficial. In certain traits, it is possible to phenotype only crossbred animals without losing prediction accuracy for ADFI. However, in other traits, it is necessary to have phenotypes in both purebred and crossbred animals, in order to achieve the highest prediction accuracy.

Assuming the same selection intensities and generation intervals in all reference populations, the relative improvement in accuracy reflects the expected change in genetic improvement. Accuracies for the EBV of purebreds for crossbred performance are currently in the magnitude of 0.5, which still leaves room for improvement. This was the starting point of the research in this thesis; how can we improve the prediction model to better estimate crossbred performance in purebred animals? Adding genomic information from crossbreds improved accuracy, as it allowed us to take into account genetic differences in purebred and commercial crossbred performance. However, I expect that genomic information from crossbreds can be better utilized, as there is information in the crossbred genome that is unused by the genomic prediction models that are currently available. In this thesis, I investigated a model where the effects of SNP alleles in crossbred animals are specific to the parental breed of origin (Dekkers, 2007).

Before the discussion of this model, I will discuss approaches to determine the breed of origin of alleles, which is an important prerequisite for implementing prediction models that consider SNP alleles effects to be breed-specific.

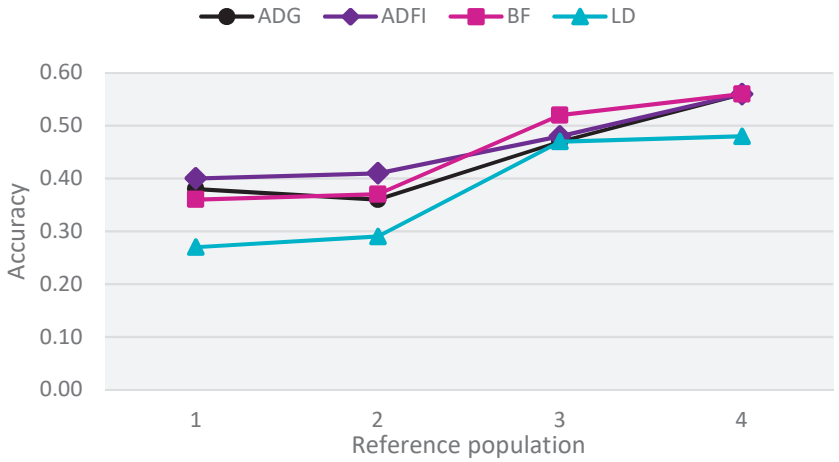

Figure 6.2. Predicted accuracy of the EBV of purebred animals for crossbred performance of average daily gain (ADG), average daily feed intake (ADFI), back fat thickness (BF), and loin depth (LD), using different reference populations: 1) purebred phenotypes, 2) purebred and crossbred phenotypes, 3) purebred phenotypes, genotypes, and crossbred phenotypes, and 4) purebred and crossbred phenotypes and genotypes.

6.3 Determination of the breed of origin of alleles

In this section, I will discuss the determination of the breed of origin of alleles. Accuracy in determining the breed of origin of alleles is expected to impact the prediction accuracy of models that account for breed-specific SNP allele effects. I will mainly discuss the approach used in **chapter 2**.

In order to estimate origin dependent allele effects, it is necessary to know the origin of all alleles of crossbred individuals. To assign the breed of origin of alleles in crossbred pigs, I used the approach developed by Vandenplas et al (2016). This approach, hereafter called the BOA approach, correctly assigns the breed of origin for ~96 % of the SNPs, and incorrectly for ~1%. It is not able to assign a breed of origin for ~3% of the SNPs of three-way crossbred pigs, simulated to originate from three distantly-related, purebred-based populations, such as the three purebred-based populations used in this thesis. Other methods that infer local ancestry in admixed populations that originate from two or more populations also exist, for

instance PCADMIX (Brisbin, 2010), SABER+ (Johnson et al., 2011), LAMP-LD (Baran et al., 2012), CHROMOPAINTER (Lawson et al., 2012), ALLOY (Rodriguez et al., 2013), and MULTIMIX (Churchhouse and Marchini, 2013). There are a few more methods that are only suitable for two-way admixture, e.g. Price et al. (2009). These methods for inferring local ancestry in admixed populations differ in their modelling assumptions, computation time, the data to which they may be applied (e.g. two-way admixture or more), and the parameters that must be specified by the user. These methods were developed for human studies in which ancestral populations are usually defined at a continental level (e.g. European, West African, Native American), which are genetically diverse populations with a Weir and Cockerham's F_{ST} (Weir and Cockerham, 1998) between them of around 0.10-0.16 (Bhatia et al., 2013). For pigs, purebred-based populations may include different lines of the same breed, or a cross of several breeds (i.e. a synthetic breed), which are still distantly related populations, with a F_{ST} of around 0.12-0.17 (**chapter 2**).

From the approaches mentioned above, MULTIMIX (Churchhouse and Marchini, 2013), which models the linkage disequilibrium background, had a good performance in assigning the ancestry to phased (±98%) and unphased data (±97%) from simulated admixed humans. The three-way admixed human population simulated by Churchhouse and Marchini (2013) and my three-way crossbred pig population differed in important aspects, which forces me to be cautious about a direct comparison of percentages of assignment between the approaches used for each population. The patterns of linkage disequilibrium in both populations might be different, and the breakdown of linkage disequilibrium might be higher in the human population because of its admixture dynamics (Bryc et al., 2010). Nonetheless, the F_{ST} between each of the three base populations were similar in the human study (0.10-0.16) (Bhatia et al., 2013) and my study (0.12-0.17) (**Chapter 2**). For MULTIMIX, the number of incorrectly classified SNPs depends on the extent of admixture, therefore, I would expect that assigning the origin of alleles would be easier for the crossbred pigs than for the admixed humans, and that the ~1% higher percentage of correct assignments of MULTIMIX, compared to BOA, could be slightly larger if tested on three-way crossbred pigs. However, the MULTIMIX approach requires input parameters of the number of SNPs per window, for which optimal values are unknown for the three-way crossbred population. Therefore, this approach needs to be tested beforehand by using simulated data of three-way crossbred pigs, in order to establish these input parameters before using it on real data. The other approaches mentioned previously achieved assignments lower than 87% for admixed humans (Baran et al., 2012; Rodriguez et al., 2013), which is

lower than MULTIMIX. This makes them unappealing to be tested in the three-way crossbred pig population used in this thesis, at least not before the MULTIMIX approach.

Some of the approaches for inferring local ancestry in human populations, as already mentioned with regards to the MULTIMIX approach, make use of information such as levels of linkage disequilibrium between subsets of SNPs in ancestral populations or recombination rates. The BOA approach does not use this information directly. Including this information could be useful to increase the percentage of assigned alleles to a breed of origin using the BOA approach. Phasing within the BOA approach is done by AlphaPhase1.1 software (Hickey et al., 2011) that implements a long-range phasing (LRP), which uses information from both related and seemingly unrelated individuals, by invoking the concept of surrogate parents (Kong et al., 2008). For any given locus, the failure to identify surrogate parents of an individual results in the locus being unphased. If a locus is unphased, it is impossible to assign a breed of origin to the alleles. LRP ignores the fact that recombination rates vary across the genome, meaning that the optimal block length also varies across the genome. Therefore, a long block built with a LRP might not be observed in the parental populations, probably because that block is actually a recombined haplotype. In **chapter 4**, I built blocks based on linkage disequilibrium, which were, at the most, 115 SNPs long. The LRP blocks built in **chapter 2** using the BOA approach were at least 450 SNPs long. As with LRP blocks, lengths are chosen arbitrarily, so it was possible that LRP blocks contained recombined haplotypes. Consequently, considering the linkage disequilibrium structure in unphased genomic regions from the BOA approach, allowing individual blocks to vary in size might increase the chances of identifying surrogate parents. Therefore, we might be able to phase that region. On the other hand, the approaches for human populations do not make use of the knowledge that each crossbred individual originates from a well-defined crossbreeding scheme. BOA was optimized for two-way and three-way crossbred schemes. For instance, for three-way crossbred pigs, both the expectation for the percentage of global ancestry (i.e. 0.5 for the paternal S breed, and 0.25 for each of the maternal breeds LR and LW) and the fact that these base populations are up to two or three generations behind, is known beforehand. There appear to be many available approaches to assign the breed of origin of alleles, but for the type of crossbred animals used in this thesis, the BOA approach still seems to be the best option.

6.4 Models for breed-specific allele effects

In this thesis, I investigated a model where the effects of SNP alleles in crossbred animals were specific to the parental breed of origin (BOA model) (Dekkers, 2007). Theoretically, the superiority of this model could be associated to the fact that it takes into account the differences between parental breeds, concerning linkage disequilibrium between SNPs and QTL, allele frequencies of both SNPs and QTL, and epistatic interactions. In practice, this model has limited benefits when tested using the real data of three-way crossbred pigs (**chapter 3**), because assuming that all SNPs are breed-of-origin dependent was too strong a supposition, which may not apply to all SNPs. Later, I verified that SNP-allele effects are breed-of-origin dependent for genomic regions strongly associated with crossbred performance, however, the majority of the genomic regions are not, or are only weakly, associated with crossbred performance (**chapter 4**). In order to overcome this, I developed a model that accounts for breed-specific allele effects, only for SNPs strongly associated with crossbred performance, while for the remaining SNPs, the model assumes identical effects irrespective of their origin (SEL-BOA model) (**chapter 5**). In this thesis, all the models implementing breed-specific SNP-allele effects were compared to a model where all SNP-alleles have the same effect across breeds (G model). In this section, I will synthesize the most important findings of using genomic prediction models that account for breed of origin of alleles in crossbred animals, and I will discuss potential improvements of the model, as well as its potential implementation in breeding programs.

The results in this thesis show that the BOA model, which estimates allele effects for crossbred performance of all SNPs based on their breed of origin, has limited benefit when tested using real data of three-way crossbred pigs (**chapter 3**). The positive impact of the BOA model was observed for average daily gain (ADG), and only for the maternal breeds, i.e. Landrace and Large White (**chapters 3 and 5**). Since the BOA model allows all SNPs to have allele effects estimated according to the breed of origin, this model may suffer from the need to estimate three times more effects using the same amount of data. Therefore, if alleles from a specific SNP have the same effect, independent of the breed of origin, the expected effects are the same, but each is estimated with less information. Similarly, Ibáñẽz-Escriche et al. (2009) concluded that the BOA model requires more information to be competitive against a model where SNP allele effects are considered to be the same across breeds. Results from the literature and from **chapter 3** suggest that, in cases where traits have a low r_{pc}, the use of the BOA model has some benefits. However, in **chapter 5**, it became clear that, when estimates of r_{pc} differ largely

between the BOA model and the G model, rather than a low r_{pc}, the use of the BOA model has some benefits. These results imply that the breed-specific r_{pc} values estimated with the BOA model are, effectively, correlated to the effects on purebred and crossbred performance of alleles originating from the same breed, while the G model estimates one r_{pc} value considering the effects of alleles originating from all breeds involved. Figure 6.3 represents the accuracy of GEBV for different traits and breeds calculated with the BOA model and the G model, and compared to the difference of r_{pc} estimated using the BOA and G model for the same traits and breeds. When the difference between the r_{pc}, estimated using the BOA model and the G model (r_{pc} difference= r_{pc} BOA- r_{pc} G), was positive and deviated further from zero, I observed higher accuracy in the BOA model, compared to the G model. I expected r_{pc} to be higher when calculated with the BOA model rather than the G model, because the crossbred additive variance components estimated using the BOA model comprises the variation observed in crossbred pigs, only due to the alleles coming from the analyzed breed. This expectation was observed in 50% of the estimates. As for the other estimates, this was not always observed in most cases in my results. This is understandable, however, as standard errors, especially in the BOA model, are quite high (0.09-0.29).

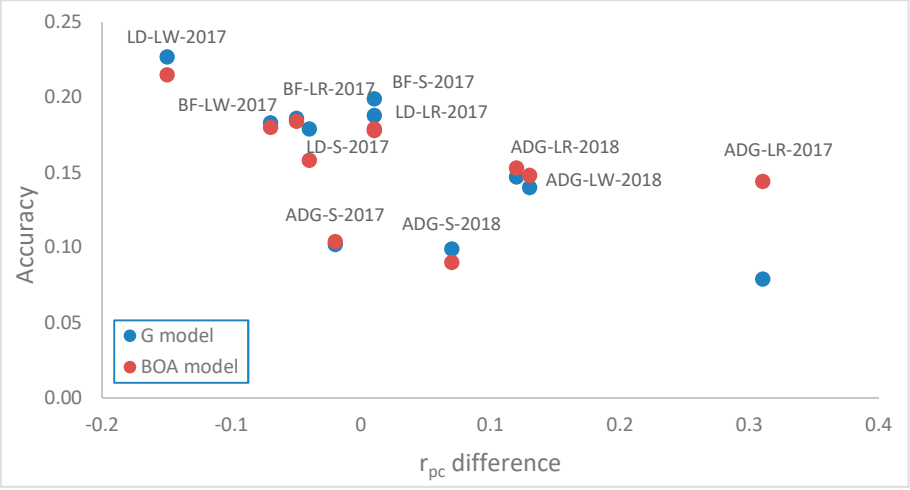

Figure 6.3. Accuracy of GEBV calculated with the G model (blue ●) or the BOA model (red ▲) on the y-axis, compared to the difference of r_{pc} estimated by the G and BOA model (r_{pc} difference= r_{pc} BOA- r_{pc} G) for different traits (loin depth (LD), back-fat thickness (BF), and average daily gain (ADG)) and different breeds (Synthetic sire (S), Landrace (LR), and Large White (LW)). Accuracies obtained from Chapter 3 (2017) and from Chapter 5 (2018).

With the SEL-BOA, I selected SNPs that, together, explained either 5% or 10% of the total crossbred genetic variance in each breed of origin. For those selected SNPs, the model considers breed-specific effects, and for the rest of the SNPs, the model assumes that effects are the same across breeds. By having separate variance components for the selected and non-selected SNPs, the model is able to assign more variance to the SNPs with a stronger association to the trait than the G and BOA models, and less to the non-selected SNPs. Also, this can differentiate the r_{pc} values into the two categories of SNPs. What I observed is that differences in variance estimates alone were not sufficient to cause a difference in accuracy, therefore, the benefit of the SEL-BOA model occurs when r_{pc} estimates are also different (**chapter 5**). As not all the SNPs show a different genetic variance according to their origin, differences in r_{pc} between the selected and non-selected SNPs indicate how different these SNPs behave, and making the distinction between these two groups is sometimes beneficial.

Results from the literature also suggest that, in cases where parental breeds are distantly related, the use of models that account for breed of origin have some benefits (Ibáñez-Escriche et al., 2009). From the three base populations, LR pigs underwent a different selection pressure that shaped their genomic architecture differently, as the results show regions with different genetic variances for certain traits, compared to S and LW (**chapter 4** and Egbert Knol, personal communication). Consequently, we observed that the LR breed immediately presented a benefit when some SNPs were already allowed to be estimated separately by breed of origin (SEL-BOA 5%).

Based on the accuracy estimated from the different models in **chapters 3** and **5**, I conclude that, in general, the differences between the models were small, but there was a tendency that the BOA model improves accuracies of GEBV for crossbred performance of traits and breeds that have a r_{pc} that is quite different when estimated with the BOA model instead of the G model. In addition to this, the SEL-BOA model tended to increase the benefit observed by the BOA model, if the r_{pc}, for the selected and non-selected SNP, differed strongly from the r_{pc} estimated by BOA model.

6.4.1 Further improvements to models for breed-specific allele effects

When genotypes originate from imputation, it has been shown that the accuracy of GEBV is linearly dependent on the imputation accuracy (Mulder et al., 2012). The

percentage of SNP alleles assigned a breed of origin are expected to have a similar impact on the prediction accuracies. For the three-way crossbred individuals, the assignment was around 95%. Therefore, if the accuracy is also linearly dependent on the assignment percentage, the prediction accuracy of our model will be at a maximum of 94% of the maximum achievable value, assuming 1% of the SNPs are incorrectly assigned. In the same study (Mulder et al., 2012), it was observed that using predicted gene dosages, rather than the most likely genotype, in order to account for the uncertainty of imputation, leads to higher accuracy and reduces bias of the imputed genotypes and the subsequent GEBV. I recommend testing the same approach with BOA, as a way to take into account the uncertainty of the assignment. The BOA approach can be modified to represent the probability that each haplotype is unique to a pure breed, instead of only assigning a breed if 80% or more of the haplotype's copies were present in only one of the purebred populations. Subsequently, the breed of origin probability for each allele at each locus will be the average probability of all haplotypes carrying that locus. In the BOA partial relationship matrices, the breed of origin probabilities can be used to weigh each of the breed-specific allele content for crossbred pigs (coded as 0 or 1), in order to calculate a predicted breed-specific allele dosage that is allowed to keep all the information available for the model. As previously mentioned, improving the BOA approach will have a limited impact on accuracy. Even if the percentage of assignment is improved, and 100% of the SNPs are assigned a breed of origin without errors, I expect that its impact will be limited to a maximum increase of 6% in GEBV accuracy. In other words, assuming a GEBV accuracy of 0.20, this means that GEBV accuracy will have increased to only 0.21 (0.20×1.06). However, using breed-specific allele dosage might have an impact on GEBV accuracy *per se*. Therefore, I believe it is worth performing these modifications to the BOA approach and the BOA model, in order to test their impact on prediction accuracy.

6.4.2 Practical application

This thesis showed that accounting for the breed of origin of alleles in crossbred animals slightly improves prediction accuracy for breeds and traits that have an r_{pc} that is quite different when estimated with the BOA model instead of the G model. Moreover, for breeds and traits where the r_{pc} for the 'non-selected SNP' and the 'selected SNP' differed strongly from the r_{pc} estimated by BOA model, it is better to use the SEL-BOA model, rather than the BOA model. For other breeds or traits, I expect that the application of the BOA or the SEL-BOA model will not improve, nor impair, the predictions.

Currently, routine genetic evaluation in pig-breeding companies is done using the 'single-step' method (Aguilar et al., 2010; Christensen and Lund, 2010). As explained in **chapter 1**, the single-step method combines the pedigree-based relationship matrix with genomic information. In this way, the evaluation is able to deal with both genotyped and non-genotyped animals. For the BOA model to be used in routine evaluations, it needs to be implemented using the single-step method. This means that the partial relationship matrices used in the BOA model need to be adjusted to be compatible to the corresponding breed specific (partial) pedigree-based relationship matrices. Procedures to achieve this have been described by Christensen et al. (2014, 2015), and tested with real data from two-way crossbred pigs by Xiang et al. (2016). Therefore, practically speaking, this does not present an obstacle to implement the BOA model in routine genetic evaluation. However, the running time of the genetic evaluation can be problematic. Major pig-breeding companies deliver EBVs for clients on a daily basis. Including genomic information from crossbred pigs without taking into account breed of origin is estimated to increase delivery time by up to three days (Rob Bergsma, personal communication). From my experience, the BOA model takes about 0.5 times longer than GBLUP to run, therefore, the delivery time is expected to be more than 4 days when calculating EBVs with the BOA model, as it is currently implemented. Using genomic information from crossbred pigs using a BOA approach means that, before knowing if an individual will be selected or discarded, the tested animal needs to stay in the testing farm for at least three extra days. If we consider the fact that a breeding program has 7500 testing animals (6000 gilts and 1500 boars), and the cost per day for maintaining one animal in the farm is 1 euro, every extra day of this breeding program has a consequence of 7500 euros worth of extra cost. Moreover, the determination of the breed of origin is also time consuming, and every time new crossbred animals are added to the dataset, the analysis has to be redone. The first part of the BOA approach, which includes the phasing with different window sizes, is, specifically, the most time consuming, as the haplotype library has to be built. The way the BOA model is currently implemented is not very time efficient. I can imagine some strategies to reduce the running time. For instance, the haplotype library built with the first reference population can be used in subsequent analyses for the phasing of new crossbred animals that are added to the dataset, without the need to phase the reference population again (Hickey et al., 2010). Also, if there is room for computational improvements, these modifications can be implemented to reduce the running time.

Furthermore, it might be of interest to implement the SEL-BOA model, as it displayed a better performance than the BOA-model for some traits and breeds (**chapter 5**). The implementation of the SEL-BOA model faces the challenge of multi-trait analysis, as it is commonly performed in routine breeding value evaluations. In a multi-trait analysis, many traits are included in the model, and they are analyzed with the same relationship matrix or with the same partial relationship matrices. Therefore, in the SEL-BOA model, the partial relationship matrices with SNPs selected to have breed-specific allele effects will not be trait-specific, but the union of all SNPs will be associated to at least one of the traits. This means that all SNPs, including those not associated with the target trait, will be considered to have the same variance component and breed-specific effects for all traits. As more SNPs are allowed to have breed-specific allele effects, the benefit of making a distinction between non-selected SNPs and selected SNPs will be diluted, and the accuracy of the SEL-BOA model will decay. An option to minimize this would be to limit the number of SNPs selected per trait. One way to achieve this could be by increasing the number of phenotyped animals, or genome wide association studies (GWAS) with sequence data, as they lead to a more clear identification of QTLs (Meuwissen and Goddard, 2010; Bouwman et al., 2018). Therefore, the QTL regions could be narrowed down, and the markers identified in full linkage disequilibrium with the identified QTL can be used directly in genomic predictions models, where two (Brøndum et al., 2015) or more (SEL-BOA, **chapter 5**) relationship matrices are used, without adding too many false positives to the analysis. In this way, the number of SNPs selected from the SNP panel to have breed-specific allele effects is reduced to only the SNPs highly associated with one or more traits of interest. However, meaningful GWAS with sequence data is only feasible when more than a thousand animals are sequenced or imputed to a sequence (Daetwyler et al., 2014; Sahana et al., 2014). However, few causative mutations have been identified in livestock species, mainly in cattle (Grisart et al., 2004; Ibeagha-Awemu et al., 2016). In pigs, sequence data is still limited, but previously detected QTL regions have been refined by the use of high-density SNP chips with approximately 660,000 SNPs (Marcos Lopes, personal communication). However, even by knowing the causal mutation, the SEL-BOA model will only be beneficial if the causal mutation explains the different variance proportions according to its breed of origin.

Since models that estimate breed-specific allele effects in part, or all, of the SNPs are only beneficial for a few traits and breeds, and the benefit is not of a large magnitude, it may not be worth investing time to overcome the challenges listed in

this section, in order to make these models more efficient for routine genetic evaluations. However, if a larger crossbred reference population becomes available, the potential of these models might improve, as there will be enough information to more accurately estimate the effects of alleles for the three breeds of origin. As previously mentioned, routine breeding value evaluations are performed using a multi-trait analysis, which is expected to increase the accuracy of the GEBV by making use of information from genetically correlated traits, especially for traits with a small number of phenotypic records (Guo et al., 2014). As more data is collected, the need for a multi-trait analysis is less relevant, which simplifies the application of the SEL-BOA for routine evaluations.

To conclude, I expect that the amount of crossbred phenotypic and genomic data will be much larger in the near future. With more data, the BOA and SEL-BOA models can be re-tested, because the effects of alleles for the three breeds of origin will have more power to be estimated. For the SEL-BOA model, more data increases the possibilities of using only SNPs highly associated with the trait, and the dependency on multi-trait analyses could be relaxed. If a significant benefit is observed after re-testing the models, it would be worthwhile investing some time in making the BOA model, or the SEL-BOA model, more computationally efficient, so that it can be implemented in routine evaluations.

6.5 Further optimization of models using crossbred reference population

As discussed in the previous section, the proposed BOA model in **chapter 3** and the SEL-BOA model in **chapter 5** only displayed benefits in a few traits and breeds. The question: 'How can we improve the prediction model to better estimate crossbred performance in purebred animals?' is still open. An interesting single-step approach that uses the concept of meta-founder, together with a multi-trait model, might be a potential alternative to improve the prediction of crossbred performance (Christensen et al., 2015; Xiang et al., 2017). Meta-founders are pseudo-individuals that are included in the pedigree, as founders without known parents. These meta-founders are arbitrarily grouped, based on breeds or lines to represent different base populations. For instance, for the pig populations used in this thesis, we have three base populations (meta-founders): Synthetic sire, Landrace, and Large White purebred-based population. This approach assumes that the base animals are related, inbred within breeds, and related between breeds, due to the finite size of the base population and the connections between populations (Legarra et al.,

2015). Similar to the BOA model, this approach constructs additive relationships, assuming the allele substitution effects of SNPs will be different in different populations. However, it also allows the effects to be similar across populations, which is accounted for through the correlations between breeds. One advantage of the meta-founders approach, compared to the BOA model, is that the genomic relationship matrix is easier to construct, because tracing the breed of origin of alleles is not required. It also has more authority, as the estimation of SNP allele effects by population is done using all of the data, instead of only the alleles coming from the analyzed breed, as in the BOA model. Xiang et al. (2017) tested this single-step approach with two meta-founders, using data from Danish Landrace, Yorkshire, and two-way crossbred pigs. They found slightly better results than when using a single-step model using the breed of origin of alleles (Xiang et al., 2016). The benefit of the meta-founder approach over the breed of origin approach in three-way crossbreeding systems still needs to be tested using real data of three-way crossbred pigs.

6.6 Concluding remarks

The work in this thesis provides a better comprehension of the mosaic nature of the genome of crossbred pigs, in terms of estimated effects and explained variances by breed of origin. Moreover, this thesis provides useful tools and methods to analyze the complex genome of crossbred individuals, and it shows the impact of using the knowledge of breed origin of alleles in the prediction of breeding values for crossbred performance. In this thesis, I only used data from pigs, but the insights presented can also be applied to crossbred populations from other species. Moreover, in my general discussion, I corroborate the starting assumption that inspired this research: crossbred genomic information is relevant in the evaluation of purebred pigs for crossbred performance. With the aim to create a better model for the genetic background of the crossbred pig, I tested the impact of accounting for differences in the breed of origin of crossbred pig's alleles. Despite some SNPs effects being dependent on the breed of origin, using this information in the prediction model only slightly improved the accuracies for traits and breeds that have an r_{pc} that is quite different when estimated with the BOA model instead of the G model. Similarly, the SEL-BOA model increases the benefit already observed by the BOA model, if the r_{pc} for the non-selected and selected SNP differs strongly from the r_{pc} estimated by BOA model. The benefit of the models that account for breed of origin is due to the crossbred additive variance components being estimated by breed of origin. When variance components are

estimated by breed of origin, the variation comprises only the alleles coming from the analyzed breed, either considering all the SNPs (BOA model) or only SNPs strongly association to the trait (SEL-BOA model). The results in this thesis should motivate breeding programs to use genomic information from crossbred animals. However, the developed genomic models in this thesis still showed limited benefits, which does not make them suitable models to be implemented in breeding programs. Therefore, BOA and SEL-BOA models are not the way to go for handling crossbred genomic information aiming to increase the genetic progress at the production level.

Supplementary material

Chapter 2

Additional file S2.1. Allele assignment (%) per chromosome (Chr) to synthetic boar (S), Landrace (LR), or Large White (LW) when using pedigree information and a relaxation factor of 20%.

Chr	Paternal	Maternal			Total
	Line S	Line LR	Line LW	Total	
1	48.52	22.83	22.24	45.07	93.59
2	49.78	21.68	22.43	44.11	93.89
3	49.84	23.35	22.42	45.76	95.60
4	49.81	23.31	22.40	45.71	95.52
5	49.26	23.85	22.61	46.46	95.72
6	49.66	22.78	23.54	46.32	95.98
7	49.49	23.31	22.40	45.71	95.20
8	49.78	22.94	23.04	46.00	95.78
9	49.80	23.46	22.27	45.73	95.53
10	49.84	22.66	22.55	45.21	95.05
11	49.66	24.28	21.86	46.14	95.80
12	49.26	22.60	21.71	44.32	93.58
13	48.52	22.83	22.24	45.07	93.59
14	49.70	22.10	23.02	45.10	94.80
15	49.77	22.91	23.35	46.26	96.03
16	49.81	22.03	23.93	45.97	95.77
17	49.77	23.58	23.09	46.68	96.45
18	49.77	23.25	23.22	46.47	96.24
Total	49.56	22.99	22.68	45.67	95.23

Chapter 3

Additional file S3.1. Standard errors of additive genetic variance (σ^2_a), litter variance (σ^2_u), residual variance (σ^2_e), and heritabilities for each breed for PB and CB performance, and genetic correlations between purebred and CB pigs (r_{pc}), estimated for each trait using the BOA[a], G_A[b], and G_B[c] models.

Model	Breed	$\sigma^2_{a_{PB}}$	$\sigma^2_{u_{PB}}$	$\sigma^2_{e_{PB}}$	h^2_{PB}	$\sigma^2_{a_{CB}}$	$\sigma^2_{u_{CB}}$	$\sigma^2_{e_{CB}}$	h^2_{CB}	r_{pc}
Average daily gain										
BOA	S	374	330	307	0.03	509	170	201	0.07	0.13
	LR	309	252	221	0.03	882				0.18
	LW	196	144	137	0.02	811				0.18
G_A	S	453	327	305	0.03	323	152	209	0.05	0.13
	LR	347	253	222	0.03					0.17
	LW	214	143	137	0.02					0.12
G_B	S	372	327	305	0.03	357	152	209	0.05	0.12
	LR	316	252	222	0.03					0.16
	LW	197	146	137	0.02					0.11
Back fat thickness										
BOA	S	0.09	0.07	0.07	0.03	0.46	0.17	0.19	0.06	0.12
	LR	0.11	0.08	0.08	0.03	0.77				0.15
	LW	0.09	0.06	0.05	0.02	0.90				0.12
G_A	S	0.11	0.07	0.07	0.03	0.30	0.15	0.19	0.05	0.10
	LR	0.12	0.08	0.08	0.03					0.12
	LW	0.10	0.06	0.05	0.02					0.10
G_B	S	0.09	0.07	0.07	0.03	0.34	0.15	0.20	0.05	0.10
	LR	0.11	0.08	0.08	0.03					0.12
	LW	0.09	0.06	0.05	0.02					0.09
Loin depth										
BOA	S	0.99	0.66	0.53	0.03	2.97	0.93	1.38	0.07	0.12
	LR	0.65	0.44	0.41	0.03	3.72				0.29
	LW	0.45	0.27	0.22	0.02	4.83				0.17
G_A	S	1.19	0.65	0.53	0.03	1.67	0.57	1.42	0.04	0.12
	LR	0.72	0.43	0.41	0.03					0.17
	LW	0.48	0.27	0.22	0.02					0.12
G_B	S	0.98	0.65	0.53	0.03	1.77	0.29	1.42	0.04	0.11
	LR	0.66	0.43	0.41	0.03					0.16
	LW	0.44	0.27	0.22	0.02					0.11

S = Synthetic boar, LR = Landrace, LW = Large White, and CB = three-way crossbred pigs

[a]BOA model, model with breed-specific relationship matrices, [b]$\mathbf{G_A}$ model, model with genomic relationship matrix by allele frequencies obtained across the genotyped population, [c]$\mathbf{G_B}$ model, model with genomic relationship matrix by breed-specific allele frequencies

Chapter 4

Additional file S4.1. Proportion of genetic variance for back fat thickness explained by the top 10 LD blocks for purebred and crossbred performance by breed of origin.

S

Chromosome	# snp	Start position, bp	End position, bp	rank PB	rank CB	gVar PB, %	gVar CB, %
1	17	158940596	160210902	10	9	0.24	0.31
4	21	2020990	2923924	45	8	0.11	0.33
5	7	29291278	29807656	6	13	0.31	0.24
5	16	30079766	30783586	4	6	0.35	0.39
6	20	45162579	46159717	13	7	0.21	0.37
6	27	47770178	49883373	5	3	0.32	0.53
11	10	7556280	7940249	16	5	0.20	0.40
11	15	7959313	8,759,687	14	4	0.20	0.47
11	22	8940153	9859528	11	10	0.22	0.30
14	4	111029177	111497364	9	>	0.24	0.02
15	30	102308093	104508529	2	2	0.64	0.59
15	25	124052987	124718596	7	>	0.29	0.05
17	5	19434891	19639253	3	>	0.48	0.01
18	13	9589537	10036306	8	31	0.27	0.14
18	8	10,083,200	10555467	1	1	1.62	1.14
Total[1]						4.51	4.51

LR

Chromosome	# snp	Start position, bp	End position, bp	rank PB	rank CB	gVar PB, %	gVar CB, %
1	27	53883921	55119115	20	6	0.12	0.19
2	28	10790595	11683915	7	13	0.15	0.14
4	9	17843084	18032705	5	>	0.17	0.03
4	4	76050292	76120254	>	9	0.04	0.17
5	5	66721316	66841468	8	>	0.15	0.05
6	12	4410006	4812421	27	5	0.11	0.22
6	26	5671575	6181785	23	8	0.11	0.17
6	19	66750992	67959109	10	31	0.15	0.10
6	15	85959532	86401903	6	19	0.16	0.13
9	20	67988	611480	4	2	0.17	0.32
8	26	60556630	63983448	3	3	0.18	0.31
9	17	127384290	127871583	9	24	0.15	0.12
9	24	128346581	128912053	1	1	0.25	0.33
14	13	14247493	14909953	32	7	0.10	0.18
15	21	119257860	119820905	2	4	0.21	0.30
16	6	35614962	35917453	35	10	0.10	0.15
Total[1]						1.73	2.35

LW

Chromosome	# snp	Start position, bp	End position, bp	rank PB	rank CB	gVar PB, %	gVar CB, %
1	13	12517775	12779800	60	9	0.10	0.23
1	16	159238083	160210902	5	28	0.26	0.14
2	16	9821331	10521615	7	1	0.25	0.38
2	6	144583667	144699882	6	4	0.25	0.29
2	7	144841166	145072905	1	6	0.35	0.26
5	3	66211719	66307363	4	36	0.27	0.12
5	11	91894653	92261586	21	7	0.16	0.26
6	24	77495951	78368681	2	2	0.31	0.37
9	24	87025203	91501337	10	23	0.21	0.15
10	4	25784747	25911585	9	30	0.22	0.13
11	10	7556280	7940249	3	3	0.29	0.29
11	27	7959313	9,394,153	23	5	0.15	0.29
15	21	118193036	118772018	34	10	0.12	0.22
15	17	132454779	132854143	17	8	0.17	0.26
16	17	32932796	33391365	8	29	0.22	0.14
Total[1]						2.61	2.85

[1]Total measured only considering the top 10 blocks

> Ranking higher than 100.

gVar PB = percentage of genetic variance explained by a LD block for purebred performance.

gVar CB = percentage of genetic variance explained by a LD block for crossbred performance.

Additional file S4.2. Proportion of genetic variance for average daily gain explained by the top 10 LD blocks for purebred and crossbred performance by breed of origin.

S								LR								LW							
Chromosome	# snp	Start position, bp	End position, bp	rank PB	rank CB	gVar PB, %	gVar CB, %	Chr	# snp	Start position, bp	End position, bp	rank PB	rank CB	gVar PB, %	gVar CB, %	Chr	# snp	Start position, bp	End position, bp	rank PB	rank CB	gVar PB, %	gVar CB, %
1	22	22995675	23914711	8	22	0.19	0.13	1	30	24160760	25558722	9	2	0.16	0.21	1	37	148170231	150386584	2	1	0.44	0.52
1	17	50797783	51394966	1	2	0.67	0.71	2	9	127150044	127601163	4	53	0.18	0.08	1	6	152357354	153127120	4	8	0.32	0.21
1	32	51502130	52696117	2	1	0.62	0.87	3	25	10101710	11076493	74	10	0.07	0.14	1	16	154108314	155462235	1	2	0.63	0.37
1	13	52764397	53673007	65	10	0.10	0.19	3	10	95567908	95998761	10	>	0.15	0.05	1	16	159238083	160210902	3	34	0.33	0.13
1	11	53753531	54125627	4	4	0.46	0.34	4	34	123951734	125150793	2	50	0.23	0.08	1	13	160903291	162372950	9	7	0.21	0.22
1	17	158940596	160210902	60	5	0.10	0.33	8	9	128638260	128759315	8	>	0.16	0.04	3	9	89799364	90371246	45	6	0.11	0.22
6	14	27029375	28268828	34	9	0.12	0.20	9	13	21695319	22071228	71	8	0.08	0.15	5	10	21487563	21970939	10	>	0.18	0.07
9	46	76052645	78754343	7	6	0.20	0.25	9	16	85529908	86702813	7	>	0.16	0.06	7	34	27394424	28541774	24	9	0.14	0.21
12	2	2269903	2291228	6	>	0.21	0.03	10	4	41980415	42035062	34	1	0.10	0.22	7	16	37145252	37994461	21	5	0.14	0.22
12	50	2371834	4006859	87	3	0.08	0.48	10	12	59598915	60001020	19	9	0.12	0.15	8	7	11342418	11417428	8	22	0.21	0.14
15	25	124052987	124718596	5	>	0.23	0.01	10	19	60023220	60408029	69	3	0.08	0.18	10	8	36258440	36584543	7	99	0.23	0.07
15	20	132308830	132854143	10	>	0.18	0.02	12	28	16536280	17289226	1	4	0.37	0.17	15	12	8009455	8319700	32	10	0.12	0.19
16	36	41412816	44459651	18	8	0.16	0.20	13	28	28111755	29212714	>	6	0.03	0.16	15	17	132454779	132854143	6	3	0.23	0.30
17	11	11211527	11438330	46	7	0.11	0.22	15	15	78438274	79022296	6	16	0.18	0.13	18	20	53,234,165	53887250	5	4	0.23	0.27
17	1	19407489	19407489	9	>	0.19	0.00	18	6	9757116	9935903	3	5	0.20	0.17								
17	5	19434891	19639253	3	>	0.62	0.00	18	11	44962762	45188783	5	>	0.18	0.05								
								18	12	54261734	54696517	15	7	0.13	0.15								
Total[1]						3.57	3.80	Total[1]						1.81	1.71	Total[1]						3.00	2.71

[1]Total measured only considering the top 10 blocks

> Ranking higher than 100.

gVar PB = percentage of genetic variance explained by a LD block for purebred performance.

gVar CB = percentage of genetic variance explained by a LD block for crossbred performance.

Additional file S4.3. Proportion of genetic variance for residual feed intake explained by the top 10 LD blocks for purebred and crossbred performance by breed of origin.

	S							LR							LW							
Chromosome	# snp	Start position, bp	End position, bp	rank PB	rank CB	gVar PB, %	gVar CB, %	Chr	# snp	Start position, bp	End position, bp	rank PB	rank CB	gVar PB, %	gVar CB, %	Chr	# snp	Start position, bp	End position, bp	rank PB	rank CB	gVar PB, %	gVar CB, %
2	13	8862251	9150134	10	>	0.20	0.05	1	13	183924206	185042994	1		0.53		1	24	10362059	10911742	7	>	0.19	0.06
2	25	133590610	134496051	>	6	0.01	0.21	6	46	14539308	16224309	3		0.26		1	30	24499758	25948260	9	>	0.19	0.10
2	33	134538298	135282067	>	1	0.15	0.42	6	10	21728055	21946695	4		0.24		1	37	148170231	150386584	70	8	0.09	0.19
3	15	111556261	111940811	>	10	0.00	0.18	6	12	23955670	24925892	5		0.24		1	13	183719835	184802857	3	2	0.25	0.29
4	21	2020990	2923924	3	4	0.29	0.25	6	56	62235659	65483397	2		0.29		1	14	246897187	247304332	>	4	0.04	0.25
4	19	94185015	94967342	>	5	0.01	0.23	12	28	16536280	17289226	6		0.19		3	20	98977271	100368914	2	36	0.26	0.12
5	13	5149762	5344082	>	3	0.00	0.28	14	47	45827810	47826895	7		0.19		4	8	104606569	104793040	5	>	0.23	0.06
5	16	78361124	79093904	62	9	0.09	0.20	16	24	18490090	19483752	8		0.17		4	18	105565718	106658684	24	3	0.13	0.26
6	11	155346307	155602352	7	>	0.22	0.04	17	5	29056018	29246809	9		0.16		6	31	14945308	16264536	1	68	0.38	0.09
7	33	2203524	2849221	>	2	0.05	0.30	18	6	47607716	47709763	10		0.16		6	16	107231177	108437309	21	1	0.14	0.32
10	10	1810290	2029001	2	39	0.37	0.11									7	20	2582366	2954492	>	5	0.02	0.20
10	18	3253139	3720144	8	>	0.21	0.04									7	16	37145252	37994461	10	60	0.19	0.10
12	45	8846410	9767814	1	36	0.59	0.12									10	9	46083236	46267311	8	41	0.19	0.11
14	47	46219359	48033173	>	8	0.06	0.21									12	21	9792506	10162364	6	>	0.21	0.03
14	23	125343446	125929866	9	7	0.21	0.21									12	23	16484131	17116574	29	6	0.13	0.20
15	14	14560948	14903909	5	>	0.23	0.07									13	16	52757976	53401107	>	10	0.02	0.18
16	33	18133808	19357238	4	12	0.25	0.17									15	42	27445103	28658318	4	51	0.24	0.11
18	21	47239897	47709763	6	40	0.23	0.11									15	26	127657851	128198698	33	9	0.12	0.18
																17	14	46801041	47554481	>	7	0.02	0.19
Total[1]						2.80	2.50	Total[1]						2.42		Total[1]						2.33	2.28

[1]Total measured only considering the top 10 blocks

> Ranking higher than 100.

gVar PB = percentage of genetic variance explained by a LD block for purebred performance.

gVar CB = percentage of genetic variance explained by a LD block for crossbred performance.

Additional file S5.1. Standard errors of additive genetic variance (σ_a^2), litter variance (σ_u^2), residual variance (σ_e^2), and heritabilities for each breed for purebred (PB) and crossbred (CB) performance, and genetic correlation between PB and CB performance (r_{PC}), estimated using the G[a], BOA[b], and SEL-BOA[c] models.

Model	Breed	$\sigma^2_{a_{PB}}$*		$\sigma^2_{u_{PB}}$	$\sigma^2_{e_{PB}}$	h^2_{PB}*		$\sigma^2_{a_{CB}}$*		$\sigma^2_{u_{CB}}$	$\sigma^2_{e_{CB}}$	h^2_{CB}*		r_{pc}*	
		Non-sel	Sel			Non-sel	Sel	Non-sel	Sel			Non-sel	Sel	Non-sel	Sel
G	S	273.7		216.2	199.1	0.016		284.2		199.3	167.8	0.031		0.081	
	LR	253.9		215.9	178.9	0.026								0.118	
	LW	159.1		112.8	94.3	0.013								0.081	
BOA	S	226.9		216.2	199.1	0.014		375.8		200.5	183.3	0.30+		0.086	
	LR	241.3		216.2	178.7	0.025		646.0						0.177	
	LW	146.5		112.7	94.3	0.012		687.5						0.103	
SEL-BOA	S	259.3	122.5	216.6	199.3	0.016	0.008	253.5	253.1	194.1	166.6	0.03	0.03	0.114	0.002
5%	LR	233.2	146.9	215.7	177.6	0.026	0.017		359.5					0.177	0.126
	LW	149.3	73.8	113.3	94.2	0.013	0.007		350.0					0.109	0.161
SEL-BOA	S	251.8	167.6	215.5	199.6	0.016	0.011	217.9	340.6	190.9	163.3	0.020	0.035	0.208	0.101
10%	LR	226.6	195.4	214.2	176.3	0.026	0.022		453.6					0.305	0.116
	LW	146.1	98.8	113.3	94.1	0.013	0.009		496.6					0.188	0.111

S = Synthetic boar, LR = Landrace (LR), LW = Large White (LW).

[a]G model, model for across-breed effects for all SNPs

[b]BOA model, model for breed-specific effects for all SNPs.

[c]SEL-BOA model, model with breed-specific effects for SNPs strongly associated with crossbred performance and across-breed effects for all other SNPs. SEL-BOA (5%) and SEL-BOA (10%) considering top 5% or top 10% of the SNPs associated with crossbred performance as strongly associated with crossbred performance, respectively.

References

Aguilar, I., Misztal, I., Johnson, D., Legarra, A., Tsuruta, S., and Lawlor, T. (2010). Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score1. *Journal of dairy science* 93**,** 743-752.

Badke, Y.M., Bates, R.O., Ernst, C.W., Schwab, C., and Steibel, J.P. (2012). Estimation of linkage disequilibrium in four US pig breeds. *BMC genomics* 13**,** 24.

Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G., and Avila, P.C. (2012). Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* 28**,** 1359-1367.

Bastiaansen, J., Bovenhuis, H., Lopes, M., Silva, F., Megens, H., and Calus, M. (2014). "SNP Effects Depend on Genetic and Environmental Context", in: *10th World Congress on Genetics Applied to Livestock Production*. Vancouver, Canada.

Bhatia, G., Patterson, N.J., Sankararaman, S., and Price, A.L. (2013). Estimating and interpreting FST: the impact of rare variants. *Genome research* 23, 1514-1521.

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

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

Bolormaa, S., Pryce, J.E., Kemper, K.E., Hayes, B.J., Zhang, Y., Tier, B., Barendse, W., Reverter, A., and Goddard, M.E. (2013). Detection of quantitative trait loci in Bos indicus and Bos taurus cattle using genome-wide association studies. *Genetics Selection Evolution* 45**,** 43.

Bouwman, A.C., Daetwyler, H.D., Chamberlain, A.J., Ponce, C.H., Sargolzaei, M., Schenkel, F.S., Sahana, G., Govignon-Gion, A., Boitard, S., and Dolezal, M. (2018). Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nature genetics* 50**,** 362.

Brandt, H., and Täubert, H. (1998). Parameter estimates for purebred and crossbred performances in pigs. *Journal of Animal Breeding and Genetics* 115**,** 97-104.

Brisbin, A. (2010). Linkage analysis for categorical traits and ancestry assignment in admixed individuals. PhD thesis. Cornell University.

Brøndum, R.F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., and Lund, M.S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of dairy science* 98**,** 4107-4116.

Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C.D., and Ostrer, H. (2010). Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences***,** 107, 8954-8961.

Cai, W., Casey, D., and Dekkers, J. (2008). Selection response and genetic parameters for residual feed intake in Yorkshire swine1. *Journal of animal science* 86**,** 287-298.

Calus, M.P.L., De Roos, A., and Veerkamp, R. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics* 178**,** 553-561.

Calus, M.P.L., and Vandenplas, J. (2015). "Calc_grm – a program to compute pedigree, genomic, and combined relationship matrices.". ABGC, Wageningen UR Livestock Research.

Calus, M.P.L., Vandenplas, J., Ten Napel, J., and Veerkamp, R. (2016). Validation of simultaneous deregression of cow and bull breeding values and derivation of appropriate weights. *Journal of dairy science*.

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

Christensen, O.F., Madsen, P., Nielsen, B., and Su, G. (2014). Genomic evaluation of both purebred and crossbred performances. *Genetics Selection Evolution* 46**,** 23.

Christensen, O.F., Legarra, A., Lund, M.S., and Su, G. (2015). Genetic evaluation for three-way crossbreeding. *Genetics Selection Evolution* 47**,** 98.

Churchhouse, C., and Marchini, J. (2013). Multiway admixture deconvolution using phased or unphased ancestral panels. *Genetic epidemiology* 37**,** 1-12.

Cucchi, T., Hulme-Beaman, A., Yuan, J., and Dobney, K. (2011). Early Neolithic pig domestication at Jiahu, Henan Province, China: clues from molar shape analyses using geometric morphometric approaches. *Journal of Archaeological Science* 38**,** 11-22.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., and Grohs, C. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics* 46**,** 858.

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

De Roos, A., Hayes, B.J, and Goddard, M. (2009). Reliability of genomic predictions across multiple populations. *Genetics* 183**,** 1545-1553.

Dekkers, J.C.M. (2004). Commercial application of marker-and gene-assisted selection in livestock: Strategies and lessons. *Journal of animal science* 82 E-suppel, E313-328.

Dekkers, J.C.M. (2007). Marker-assisted selection for commercial crossbred performance. *Journal of Animal Science* 85**,** 2104-2114.

Diniz, D., Lopes, M., Broekhuijse, M., Lopes, P., Harlizius, B., Guimarães, S., Duijvesteijn, N., Knol, E., and Silva, F. (2014). A genome-wide association study reveals a novel candidate gene for sperm motility in pigs. *Animal reproduction science* 151**,** 201-207.

Do, D.N., Ostersen, T., Strathe, A.B., Mark, T., Jensen, J., and Kadarmideen, H.N. (2014). Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. *BMC Genetics* 15**,** 27.

Ervynck, A., Dobney, K., Hongo, H., and Meadow, R. (2001). Born Free? New Evidence for the Status of" Sus scrofa" at Neolithic Çayönü Tepesi (Southeastern Anatolia, Turkey). *Paléorient* 27, 47-73.

Esfandyari, H., Sørensen, A.C., and Bijma, P. (2015). A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genetics Selection Evolution* 47**,** 76.

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary bioinformatics* 1**,** 47-50.

Falconer, D.S., Mackay, T.F., and Frankham, R. (1996). Introduction to quantitative genetics (4th edn).

Fan, B., Onteru, S.K., and Rothschild, M.F. (2009). The GGT1 and IGFBP5 genes are associated with fat deposition traits in the pig (Brief Report). *Archives Animal Breeding* 52**,** 337-339.

Fan, B., Lkhagvadorj, S., Cai, W., Young, J., Smith, R., Dekkers, J., Huff-Lonergan, E., Lonergan, S., and Rothschild, M. (2010). Identification of genetic markers associated with residual feed intake and meat quality traits in the pig. *Meat science* 84**,** 645-650.

Fernando, R., and Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. *Genetics Selection Evolution* 21**,** 467.

Forni, S., Aguilar, I., Misztal, I., and Deeb, N. (2010). Genomic relationships and biases in the evaluation of sow litter size, in: *Proc 9th World Congress on Genetics Applied to Livestock Production.* Liepzig, Germany.

Gilbert, H., Billon, Y., Brossard, L., Faure, J., Gatellier, P., Gondret, F., Labussière E., Lebret B., Lefaucheur L., Le Floch N., and Louveau I. (2017). Divergent selection for residual feed intake in the growing pig. A*nimal* 11**,** 1427-1439.

Gilmour, A., Gogel, B., Cullis, B., and Thompson, R. (2009). ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK.

Gilmour, A., Gogel, B., Cullis, B., Welham, S., and Thompson, R. (2015). ASReml user guide release 4.1 structural specification. VSN International Ltd, Hemel Hempstead, UK.

Godinho, R., Bergsma, R., Silva, F., Sevillano, C.A., Knol, E., Lopes, M., Lopes, P., Bastiaansen, J., and Guimarães, S. (2018). Genetic correlations between feed efficiency traits, and growth performance and carcass traits in purebred and crossbred pigs. *Journal of animal science* 96, 817-829.

Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J.-J., Kvasz, A., Mni, M., Simon, P., Frere, J.-M., and Coppieters, W. (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences* 101**,** 2398-2403.

Groenen, M.A., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., and Megens, H.-J. (2012). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491**,** 393.

Grossi, D.A., Jafarikia, M., Brito, L.F., Buzanskas, M.E., Sargolzaei, M., and Schenkel, F.S. (2017). Genetic diversity, extent of linkage disequilibrium and persistence of gametic phase in Canadian pigs. *BMC genetics* 18**,** 6.

Gualdron Duarte, J., Bates, R., Ernst, C., Raney, N., Cantet, R., and Steibel, J. (2013). Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genetics* 14**,** 38.

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC genetics* 15**,** 30.

Hayes, B., Bowman, P., Chamberlain, A., and Goddard, M. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science* 92**,** 433-443.

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

Hickey, J.M., Kinghorn, B.P., Cleveland, M., Tier, B., and Van Der Werf, J.H. (2010). Recursive long range phasing and long haplotype library imputation: building a global haplotype library for Holstein cattle. *Proc 9th world congress on genetics applied to livestock production.* Leipzig, Germany.

Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N., and Van Der Werf, J.H. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genetics Selection Evolution* 43**,** 12.

Hidalgo, A.M. (2015a). *Exploiting genomic information on purebred and crossbred pigs.* PhD Thesis. Wageningen University.

Hidalgo, A.M., Bastiaansen, J., Lopes, M.S., Veroneze, R., Groenen, M., and De Koning, D.-J. (2015b). Accuracy of genomic prediction using deregressed breeding values estimated from purebred and crossbred offspring phenotypes in pigs. *Journal of animal science* 93**,** 3313-3321.

Hidalgo, A.M, Lopes, M., Harlizius, B., and Bastiaansen, J. (2016). Genome-wide association study reveals regions associated with gestation length in two pig populations. *Animal genetics* 47**,** 223-226.

Ibánëz-Escriche, N., Fernando, R.L., Toosi, A., and Dekkers, J.C. (2009). Genomic selection of purebreds for crossbred performance. *Genetics Selection Evolution* 41**,** 1.

Ibáñez-Escriche, N., Forni, S., Noguera, J.L., and Varona, L. (2014). Genomic information in pig breeding: Science meets industry needs. *Livestock Science* 166**,** 94-100.

Ibeagha-Awemu, E.M., Peters, S.O., Akwanji, K.A., Imumorin, I.G., and Zhao, X. (2016). High density genome wide genotyping-by-sequencing and association identifies common and low frequency SNPs, and novel candidate genes influencing cow milk traits. *Scientific Reports* 6**,** 31109.

Jannink, J.-L., Lorenz, A.J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics* 9**,** 166-177.

Johnson, N.A., Coram, M.A., Shriver, M.D., Romieu, I., Barsh, G.S., London, S.J., and Tang, H. (2011). Ancestral components of admixed genomes in a Mexican cohort. *PLoS genetics* 7**,** e1002410.

Jorjani, H., Klei, L., and Emanuelson, U. (2003). A simple method for weighted bending of genetic (co) variance matrices. *Journal of dairy science* 86**,** 677-679.

Kijas, J., and Andersson, L. (2001). A phylogenetic study of the origin of the domestic pig estimated from the near-complete mtDNA genome. *Journal of Molecular Evolution* 52**,** 302-308.

Kim, K.S., Larsen, N., Short, T., Plastow, G., and Rothschild, M.F. (2000). A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. *Mammalian genome* 11**,** 131-135.

Kinghorn, B.P., Hickey, J.M., and Van Der Werf, J.H. (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.* Leipzig, Germany.

Knap, P., and Wang, L. (2012). "Pig breeding for improved feed efficiency", in: *Feed efficiency in swine, J.F. Patience*. p. 167-181.

Knol, E.F., Nielsen, B., and Knap, P.W. (2016). Genomic selection in commercial pig breeding. *Animal Frontiers* 6**,** 15-22.

Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., and Rafnar, T. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics* 40**,** 1068.

Kong, A., Steinthorsdottir, V., Masson, G., Thorleifsson, G., Sulem, P., Besenbacher, S., Jonasdottir, A., Sigurdsson, A., Kristinsson, K.T., and Jonasdottir, A. (2009). Parental origin of sequence variants associated with complex diseases. *Nature* 462**,** 868.

Kumar, N., Mitra, A., Ganguly, I., Singh, R., Deb, S.M., Srivastava, S.K., and Sharma, A. (2005). Lack of association of brucellosis resistance with (GT) 13 microsatellite allele at 3'

UTR of Nramp1 gene in Indian zebu (Bos indicus) and crossbred (Bos indicus× Bos taurus) cattle. *Veterinary microbiology* 111**,** 139-143.

Larson, G., Dobney, K., Albarella, U., Fang, M., Matisoo-Smith, E., Robins, J., Lowden, S., Finlayson, H., Brand, T., and Willerslev, E. (2005). Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* 307**,** 1618-1621.

Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS genetics* 8**,** e1002453.

Lee, S.H., and Van Der Werf, J.H. (2016). MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. *Bioinformatics* 32**,** 1420-1422.

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I., and Misztal, I. (2015). Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics* 200, 455-468.

Lin, P., Hartz, S.M., Zhang, Z., Saccone, S.F., Wang, J., Tischfield, J.A., Edenberg, H.J., Kramer, J.R., M.Goate, A., Bierut, L.J., Rice, J.P., and For the Coga Collaborators Cogend Collaborators, G. (2010). A New Statistic to Evaluate Imputation Reliability. *PLoS ONE* 5**,** e9697.

Lopes, M.S. (2016). Genomic selection for improved crossbred performance. PhD thesis. Wageningen University.

Lopes, M.S., Bovenhuis, H., Hidalgo, A.M., Arendonk, J.A., Knol, E.F., and Bastiaansen, J.W. (2017). Genomic selection for crossbred performance accounting for breed-specific effects. *Genetics Selection Evolution* 49**,** 51.

Lopes, M.S., Hanenberg, E., Dunkelberger, J., Harlizius, B., and Knol, E. (2018). On the added value of genomics for accurate prediction in pig breeding: learning from historical data, in: *the 11th World Congress on Genetics Applied to Livestock Production.* Auckland, Australia.

Lourenco, D., Tsuruta, S., Fragomeni, B., Chen, C., Herring, W., and Misztal, I. (2016). Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. *Journal of animal science* 94**,** 909-919.

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

Mackay, T.F. (2014). Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nature Reviews Genetics* 15**,** 22.

Makgahlela, M., Strandén, I., Nielsen, U., Sillanpää, M., and Mäntysaari, E. (2013). The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. *Journal of dairy science* 96**,** 5364-5375.

Makgahlela, M., Strandén, I., Nielsen, U., Sillanpää, M., and Mäntysaari, E. (2014). Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. *Journal of dairy science* 97**,** 1117-1127.

Meidtner, K., Wermter, A.K., Hinney, A., Remschmidt, H., Hebebrand, J., and Fries, R. (2006). Association of the melanocortin 4 receptor with feed intake and daily gain in F2 Mangalitsa× Piétrain pigs. *Animal genetics* 37**,** 245-247.

Merks, J. (1989). Genotype× environment interactions in pig breeding programmes. VI. Genetic relations between performances in central test, on-farm test and commercial fattening. *Livestock Production Science* 22**,** 325-339.

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

Meuwissen, T.H.E., and Goddard, M.E. (2010). Accurate prediction of genetic values for complex traits by whole genome resequencing. *Genetics* 185, 623-631.

Moghaddar, N., Swan, A.A., and Van Der Werf, J.H. (2014). Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. *Genetics Selection Evolution* 46**,** 58.

Mulder, H., Calus, M., Druet, T., and Schrooten, C. (2012). Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. *Journal of dairy science* 95**,** 876-889.

Nakavisut, S., Crump, R., Suarez, M., and Graser, H. (2005). "Genetic correlations between the performance of purebred and crossbred pigs", in: *Proceedings of the Association for the Advancement of Animal Breeding and Genetics* 16, 99-102. Noosa Lakes, Australia.

Newman, S., Reverter, A., and Johnston, D. (2002). Purebred-crossbred performance and genetic evaluation of postweaning growth and carcass traits in× crosses in Australia. *Journal of Animal Science* 80**,** 1801-1808.

Onteru, S.K., Gorbach, D.M., Young, J.M., Garrick, D.J., Dekkers, J.C., and Rothschild, M.F. (2013). Whole genome association studies of residual feed intake and related traits in the pig. *PloS one* 8**,** e61756.

Přibyl, J., Madsen, P., Bauer, J., Přibylová, J., Šimečková, M., Vostrý, L., and Zavadilová, L. (2013). Contribution of domestic production records, Interbull estimated breeding values, and single nucleotide polymorphism genetic markers to the single-step genomic evaluation of milk production. *Journal of dairy science* 96**,** 1865-1873.

Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., and Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics* 5**,** e1000519.

Pryce, J.E., Bolormaa, S., Chamberlain, A.J., Bowman, P.J., Savin, K., Goddard, M.E., and Hayes, B.J. (2010). A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *Journal of Dairy Science* 93**,** 3331-3345.

Puig-Oliveras, A., Revilla, M., Castelló, A., Fernández, A.I., Folch, J.M., and Ballester, M. (2016). Expression-based GWAS identifies variants, gene interactions and key regulators affecting intramuscular fatty acid content and composition in porcine meat. *Scientific reports* 6**,** 31803.

Ramos, A.M., Crooijmans, R.P., Affara, N.A., Amaral, A.J., Archibald, A.L., Beever, J.E., Bendixen, C., Churcher, C., Clark, R., and Dehais, P. (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**,** e6524.

Rodriguez, J.M., Bercovici, S., Elmore, M., and Batzoglou, S. (2013). Ancestry inference in complex admixtures via variable-length Markov chain linkage models. *Journal of Computational Biology* 20**,** 199-211.

Saatchi, M., Mcclure, M.C., Mckay, S.D., Rolf, M.M., Kim, J., Decker, J.E., Taxis, T.M., Chapple, R.H., Ramey, H.R., and Northcutt, S.L. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution* 43**,** 1.

Sahana, G., Guldbrandtsen, B., Thomsen, B., Holm, L.E., Panitz, F., Brøndum, R.F., Bendixen, C., and Lund, M.S. (2014). Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. *Journal of Dairy Science* 97**,** 7258-7275.

Salih, D.A., Tripathi, G., Holding, C., Szestak, T.A., Gonzalez, M.I., Carter, E.J., Cobb, L.J., Eisemann, J.E., and Pell, J.M. (2004). Insulin-like growth factor-binding protein 5 (Igfbp5) compromises survival, growth, muscle development, and fertility in mice. *Proceedings of the National Academy of Sciences* 101**,** 4314-4319.

Sargolzaei, M., Chesnais, J.P., and Schenkel, F.S. (2014). A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* 15**,** 478.

Schrag, T.A., Möhring, J., Maurer, H.P., Dhillon, B.S., Melchinger, A.E., Piepho, H.-P., Sørensen, A.P., and Frisch, M. (2009). Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. *Theoretical and applied genetics* 118**,** 741-751.

Seeley, R.J., Drazen, D.L., and Clegg, D.J. (2004). The critical role of the melanocortin system in the control of energy balance. *Annual Review of Nutrition.* 24**,** 133-149.

Sevillano, C.A., Lopes, M.S., Harlizius, B., Hanenberg, E.H., Knol, E.F., and Bastiaansen, J.W. (2015). Genome-wide association study using deregressed breeding values for cryptorchidism and scrotal/inguinal hernia in two pig lines. *Genetics Selection Evolution* 47**,** 18.

Sevillano, C.A., Vandenplas, J., Bastiaansen, J.W., and Calus, M.P. (2016). Empirical determination of breed-of-origin of alleles in three-breed cross pigs. *Genetics Selection Evolution* 48**,** 55.

Sevillano, C.A., Vandenplas, J., Bastiaansen, J.W., Bergsma, R., and Calus, M.P. (2017). Genomic evaluation for a three-way crossbreeding system considering breed-of-origin of alleles. *Genetics Selection Evolution* 49**,** 75.

Sevillano C.A., ten Napel J., Guimarães S.E.F., Silva F.F. and Calus M.P.L. (2018a). Effects of alleles in crossbred pigs estimated for genomic prediction depend on their breed-of-origin. *BMC Genomics* 19,740.

Sevillano C.A., Nicolaiciuc C. V., Molist F., Pijlman J. and Bergsma R. (2018b). Effect of feeding cereals-alternative ingredients diets or corn-soybean meal diets on performance and carcass characteristics of growing-finishing gilts and boars. *Journal of Animal Science. doi*: 10.1093/jas/sky339

Sewell, A., Li, H., Schwab, C., Maltecca, C., and Tiezzi, F. (2018). "On the value of genotyping terminal crossbred pigs for nucleus genomic selection for carcass traits", in: *the 11th World Congress on Genetics Applied to Livestock Production.* Auckland, Australia.

Slatkin, M. (1994). Linkage disequilibrium in growing and stable populations. *Genetics* 137**,** 331-336.

Slatkin, M., and Excoffier, L. (1996). Testing for linkage disequilibrium in genotypic data using the Expectation-Maximization algorithm. *Heredity* 76**,** 377.

Smith, C. (1964). The use of specialised sire and dam lines in selection for meat production. *Animal Science* 6**,** 337-344.

Stuart, A., and Ord, K. (1994). *Kendall's Advanced Theory of Statistics.* London: Hodder Education.

Ten Napel, J., Calus, M.P., Lidauer, M., Stranden, I., Mäntysaari, E., Mulder, H., and Veerkamp, R. (2016). "MiXBLUP, the Mixed-model Best Linear Unbiased Prediction software for PCs for large genetic evaluation systems". Version 2.0 ed. Wageningen, the Netherlands.

Thorsell, A.-G., Lee, W.H., Persson, C., Siponen, M.I., Nilsson, M., Busam, R.D., Kotenyova, T., Schüler, H., and Lehtiö, L. (2011). Comparative structural analysis of lipid binding START domains. *PLoS One* 6**,** e19521.

Tortereau, F., Servin, B., Frantz, L., Megens, H.-J., Milan, D., Rohrer, G., Wiedmann, R., Beever, J., Archibald, A., Schook, L., and Groenen, M. (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics* 13**,** 586.

Van Grevenhof, I.E., and Van Der Werf, J.H. (2015). Design of reference populations for genomic selection in crossbreeding programs. *Genetics Selection Evolution* 47**,** 14.

Vandenplas, J., Calus, M.P., Sevillano, C.A., Windig, J.J., and Bastiaansen, J.W. (2016). Assigning breed origin to alleles in crossbred animals. *Genetics Selection Evolution* 48**,** 61.

Vandenplas, J., Windig, J.J., and Calus, M.P. (2017). Prediction of the reliability of genomic breeding values for crossbred performance. *Genetics Selection Evolution* 49**,** 43.

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

Ventura, R.V., Lu, D., Schenkel, F.S., Wang, Z., Li, C., and Miller, S.P. (2014). Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. *Journal of Animal Science* 92**,** 1433-1444.

Veroneze, R., Bastiaansen, J.W., Knol, E.F., Guimarães, S.E., Silva, F.F., Harlizius, B., Lopes, M.S., and Lopes, P.S. (2014). Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (Sus scrofa) populations. *BMC genetics* 15**,** 126.

Veroneze, R., Lopes, M.S., Hidalgo, A.M., Guimarães, S., Silva, F., Harlizius, B., Lopes, P., Knol, E., M. Van Arendonk, J., and Bastiaansen, J. (2015). Accuracy of genome-enabled prediction exploring purebred and crossbred pig populations. *Journal of animal science* 93**,** 4684-4691.

Visscher, P., Pong-Wong, R., Whittemore, C., and Haley, C. (2000). Impact of biotechnology on (cross) breeding programmes in pigs. *Livestock Production Science* 65**,** 57-70.

Wang, H., Misztal, I., Aguilar, I., Legarra, A., Fernando, R.L., Vitezica, Z., Okimoto, R., Wing, T., Hawken, R., and Muir, W.M. (2014). Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. *Frontiers in genetics* 5**,** 134.

Wang, Y.H., Bower, N., Reverter, A., Tan, S., De Jager, N., Wang, R., Mcwilliam, S., Cafe, L., Greenwood, P., and Lehnert, S. (2009). Gene expression patterns during intramuscular fat development in cattle. *Journal of Animal Science* 87**,** 119-130.

Wei, M., and Van Der Steen, H. (1991). Comparison of reciprocal recurrent selection with pure-line selection systems in animal breeding (A review), in: *Animal Breeding Abstracts*), 281-298.

Wei, M., and Van Der Werf, J.H. (1995). Genetic correlation and heritabilities for purebred and crossbred performance in poultry egg production traits. *Journal of animal science* 73**,** 2220-2226.

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

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38**,** 1358-1370.

White, S. (2011). From globalized pig breeds to capitalist pigs: a study in animal cultures and evolutionary history. *Environmental History* 16**,** 94-120.

Wientjes, Y.C., Veerkamp, R.F., Bijma, P., Bovenhuis, H., Schrooten, C., and Calus, M.P. (2015). Empirical and deterministic accuracies of across-population genomic prediction. *Genetics Selection Evolution* 47**,** 5.

Wientjes, Y.C., and Calus, M. (2017). BOARD INVITED REVIEW: The purebred-crossbred correlation in pigs: A review of theory, estimates, and implications. *Journal of Animal Science* 95**,** 3467-3478.

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J.E., O'sullivan, N.P., Preisinger, R., Habier, D., Fernando, R., and Garrick, D.J. (2011). Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genetics Selection Evolution* 43**,** 1.

Xiang, T., Nielsen, B., Su, G., Legarra, A., and Christensen, O.F. (2016). Application of single-step genomic evaluation for crossbred performance in pig. *Journal of animal science* 94**,** 936-948.

Xiang, T., Christensen, O.F., and Legarra, A. (2017). Genomic evaluation for crossbred performance in a single-step approach with metafounders. *Journal of Animal Science* 95**,** 1472-1480.

Yang, J., Benyamin, B., Mcevoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., and Visscher, P.M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature* 42**,** 565-569.

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

Summary

In pig-breeding programs, selection of genetically best animals to produce the next generation is routinely performed by genomic selection. Genomic selection uses information of many markers spread across the genome. For pig breeding the common density is 50k single-nucleotide polymorphism (SNP) markers spread across the genome. Genomic information is used to calculate genomic estimated breeding values by combining SNP allele effects with the animal's genotype. These SNP allele effects are estimated in a training population composed by individuals with both phenotypes and genomic information. Since, any pig-breeding program involves a crossbreeding scheme, the selection in purebreds aims to improve crossbred performance at commercial farms. Therefore, currently training populations include purebred individuals with both phenotypes and genomic information but also crossbred individuals with phenotypes so the crossbred environmental background is taken into account. Including crossbred genomic information in the training population is expected to be beneficial for better estimation of SNP allele effects. Importantly, in crossbred animals, the genome is a mosaic of genomic regions inherited from the different parental breeds. As a result, depending from which breed a genomic region was inherited from, SNP alleles might have different effects. The main goal of this thesis, therefore, was to use crossbred genomic information in the training population to estimate SNP alleles effects by breed of origin, and evaluate if this approach improves the accuracy for estimation of breeding values of purebred animals for crossbred performance. Throughout this thesis I used pigs that originated from a three-way crossbreeding design, as common in pig production. In this data two maternal breed pigs, Landrace (LR) and Large White (LW), were crossed to produce F1 (LR x LW or LW x LR) crossbred pigs, which in turn were crossed with paternal synthetic breed (S) pigs to produce three-way crossbred pigs (S x (LR x LW) or S x (LW x LR)).

Chapter 2 showed that breed of origin of alleles of three-way crossbred pigs can be derived empirically for around 93.7 % of the alleles without using pedigree information. Pedigree information is useful to reduce computation time and can slightly increase the percentage of assignments to 94.6 %. The approach tested used a long-range phasing method to relax the dependency on genotyped parents and available pedigree information. Purebred haplotypes that were derived from the long-range phasing method were assigned to a breed if 80% or more of the copies were present in only one of the purebred populations. Subsequently, if an haplotype was observed in crossbred animals crossbred, all the alleles present in that haplotype were assigned the haplotype's breed of origin. The relatively assignment percentage

allowed the use of models that implement breed-specific effects of SNP alleles in genomic prediction.

In **chapter 3**, I compared a model that accounts for breed-specific effects of SNP alleles (BOA model) to models in which effects of SNP alleles for crossbred performance were assumed to be the same across breeds, using either breed-specific allele frequencies (G_A model) or allele frequencies averaged across breeds (G_B model). The comparison was done by evaluating the accuracies of the models for estimating breeding values of purebred animals for crossbred performance for average daily gain, back fat thickness, and loin depth. Only for average daily gain for LR breed, differences between the models was significant, where BOA had the highest accuracy. Across all traits, models G_A and G_B yielded similar predictions. I concluded that using the BOA model was especially relevant for traits with a low genetic correlation between purebred and crossbred performance (r_{pc}), such as average daily gain in breed LR.

The BOA model allows allele effects of all SNPs to be breed-specific. This assumption might not hold for all the SNPs and possible reasons were studied in **chapter 4**. In this study, SNP-allele substitution effects were estimated for a commonly used SNP panel using the BOA model. Estimated breeding values for purebred and crossbred performance were converted to SNP-allele effects by breed-of-origin. Differences between purebred and crossbred, and between breeds-of-origin were evaluated by comparing percentage of variance explained by genomic regions for back fat thickness, average daily gain, and residual feed intake. I observed overlapping regions across purebred and crossbred performance explaining similar additive genetic variance. The number of overlapping regions was related to the trait r_{pc}. Moreover, I observed some overlapping regions across breeds of origin that explained relatively large proportions of genetic variance for crossbred performance; albeit that the actual proportion of variance deviated across breeds-of-origin. To illustrate underlying mechanisms, I evaluated the estimated effects across breeds of origin for haplotypes associated to a missense mutation in the MC4R gene with a known effect on back fat thickness and average daily gain, and for the mutation itself. Results confirmed that even if a causal locus has similar effects across breeds of origin, estimated effects and explained variance in its region using a commonly used SNP panel can strongly depend on the allele frequency of the underlying causal mutation observed in a specific breed.

I hypothesized that targeting regions strongly associated with crossbred performance and only differentiating their SNP-allele effects according to their breed of origin, might improve prediction models for crossbred performance. Therefore, in **chapter 5** I developed a model that estimates breed-specific effects for alleles of SNPs strongly associated with crossbred performance, and for the rest of the SNPs assumes that allele effects are the same across breeds (SEL-BOA model). I selected the SNPs most strongly associated with crossbred performance based on the results from **chapter 4** that explained together either 5% or 10% of the total crossbred genetic variance for average daily gain in each breed of origin. I compared the prediction accuracies of the SEL-BOA model to those from the G_B and BOA model. Differences of prediction accuracies between models was small. The BOA model predicted crossbred performance better than the G_B model when estimated crossbred genetic variances and r_{pc} by breed of origin differed largely between the G_B and the BOA model. Superiority of the SEL-BOA model compared to the BOA model was only observed for the scenario 10% and when r_{pc} for the selected and non-selected SNP differed strongly from the r_{pc} estimated by the BOA model.

Finally, in **chapter 6** I discussed about the relevance of commercial crossbred information for breeding programs, and I showed that adding genomic crossbred information in the training population, without considering breed-specific effects of SNP alleles, already improved prediction accuracies of breeding values of purebred pigs for crossbred performance. I further focused on the practical applications of models that account for breed-specific effects of SNP alleles, therefore I discussed the approach used in this thesis to determine the breed of origin of alleles as well as alternative approaches for inferring local ancestry in admixed populations. I presented some challenges that models implementing breed-specific effects of SNP alleles have to overcome before they can be adopted in routine genetic evaluation in pig-breeding companies. I also presented some ideas that might improve the prediction accuracies of the BOA and SEL-BOA model, as well as an alternative model for using crossbred genomic information different to the breed of origin approach. Since the BOA and the SEL-BOA model are beneficial only for a few traits and breeds, I concluded that for practical application it may not be worth to invest time to overcome these challenges and future research should focus on alternative models to better utilize crossbred genomic information.

Training and supervision

The Basic Package	year	credits
WIAS Introduction Day	2014	0.3
Course on philosophy of science and/or ethics	2017	1.5
Course on essential skills	2015	1.2

Disciplinary Competences	year	credits
Preparing own PhD research proposal	2014	6
Programming and computer algorithms in animal breeding with focus on genomic selection and single-step GBLUP, Georgia, USA	2014	3.5
Genetic analysis using ASReml4.0, Wageningen	2014	1.5
Introduction to theory and implementation of Genomic Selection, Wageningen	2014	1.4
Gut health in pigs and poultry, Wageningen	2014	0.3
From sequence data to genomic prediction, Wroclaw, Poland	2015	1.0
WIAS Advanced Statistics Course Design of Experiments, Wageningen	2015	1.0
Genotype by environment interaction, uniformity and stability, Wageningen	2015	1.5
Short course on genomics tools, São Paulo, Brazil	2017	1.0
Design of Breeding Programs with Genomic Selection, Wageningen	2017	1.5

Professional Competences	year	credits
PhD Competence Assessment	2015	0.3
Techniques for Scientific Writing	2015	1.2
High impact writing in science	2015	1.3
Communication with the media and the general public	2015	1.0
Data management planning	2015	0.4
Brain Training	2016	0.3
The Final Touch: Writing the general introduction and discussion	2018	0.3

Presentation Skills	year	credits
European Association of Animal Production (Poster), Warsaw, Poland	2015	1.0
European Association of Animal Production (Oral), Belfast, North Ireland	2016	1.0
Latin American Association of Animal Production (Poster), Recife, Brazil	2016	1.0
Talking About Computing and Genomics (Invited speaker), Juiz de Fora, Brazil	2016	1.0
World Congress on Genetics Applied to Livestock Production (Poster), Auckland, New Zealand	2018	(1.0)
Teaching competences	year	Credits
Supervising Practical Genetics for Livestock Improvement	2015	3.0
Education and Training Total*		**33.5**

**One ECTS credit equals a studyload of approximately 28 hours*

Curriculum vitae

About the author

Claudia Alejandra Sevillano del Aguila was born in Lima, Peru on January 28th 1986. In 2002, she finished high school at Colegio Santísimo Nombre de Jesús, in Lima. Claudia wanted to study Veterinary, but when she learned that as an animal scientist you can also deal with nutrition and genetics of the animals, she decided to study Animal Sciences. Claudia started her BSc in Animal Husbandry at the Universidad Nacional Agraria La Molina (UNALM), Lima in 2003. She completed her BSc with honors in 2007.

Right after her bachelors, she went one year to the United States as part of an international exchange program where she worked for 9 months in a Smithfield's swine farm in North Carolina and studied for 3 months at the Ohio State University. When she returned to Lima in 2009, she started to work on her thesis to obtained the Engineering Degree at UNALM, which she also completed with honors. Enchanted by the abroad experience, she left Peru again in 2009 for one year trainingship in a swine farm in Denmark. Back in Peru in 2010, she started working as academic staff in the Faculty of Veterinarian Medicine and Animal Husbandry at the University Científica del Sur, Lima.

In 2011, Claudia was accepted to persue the double degree EURAMA Master's programme and awarded the NUFFIC scholarship. The first semester of the Master's programme she spent it at the Ecole d'Ingénieurs de Purpan, Toulouse, France and the second semester in Wageningen University and Research. Following her Master's advisor to perform a summer internship and being in the Netherlands, she was interested in having an experience at Topigs (now Topigs Norsvin) which she knew was a Dutch swine breeding company. The subject of this internship became her major thesis under the supervision of Dr. Lisbeth van der Weij and Saskia Bloemhof. Since then her relationship with Topigs Norsvin Research Center started and she continued working there as a researcher after completing her master's studies.

In 2014, she accepted the opportunity to combined her work with a PhD project at the Animal Breeding and Genomics Center, Wageningen University. As part of the PhD project collaboration, she spent one year (August 2016-August 2017) at the Universidade Federal de Viçosa, Minas Gerais, Brazil. The results of her PhD project are presented in this thesis entitled "Genomic evaluation considering the mosaic genome of the crossbred pig".

Peer-reviewed publications

Sevillano C.A., ten Napel J., Guimarães S.E.F., Silva F.F. and Calus M.P.L. 2018. Effects of alleles in crossbred pigs estimated for genomic prediction depend on their breed-of-origin. BMC Genomics 19:740. doi: 10.1186/s12864-018-5126-7

Godinho R.M., Bastiaansen J.W.M., **Sevillano C.A.**, Silva F.F., Guimarães S.E.F. and Bergsma R. 2018. Genotype by feed interaction for feed efficiency and growth performance traits in pigs. Journal of Animal Science 96:4125-35. doi: 10.1093/jas/sky304

Sevillano C.A., Nicolaiciuc C. V., Molist F., Pijlman J. and Bergsma R. 2018. Effect of feeding cereals-alternative ingredients diets or corn-soybean meal diets on performance and carcass characteristics of growing-finishing gilts and boars. Journal of Animal Science. doi: 10.1093/jas/sky339

Pamela I.O., Guimarães S.E.F., Verardo L.L., Azevedo A.L.S., Vandenplas J., Soares A.C.C., **Sevillano C.A.**, Veroneze R., Pires M.A., de Freitas C., Prata M.C.A., Furlong J., Verneque R.S., Panetto J.C.C., Carvalho W.A., Gobo D.O.R., da Silva M.V.G.B., Machado M.A. 2018. Genome wide association studies for tick resistance in Bos taurus x Bos indicus crossbred cattle: a deeper look into this intricate mechanism. Journal of Dairy Science. doi: 10.3168/jds.2017-14223

Godinho R.M., Bergsma R, Silva F.F., **Sevillano C.A.**, Knol E.F., Lopes M.S., Lopes P.S., Bastiaansen J.W.M. and Guimarães S.E.F. 2018. Genetic correlations between feed efficiency traits, and growth performance and carcass traits in purebred and crossbred pigs. Journal of Animal Science 96:817-29. doi: 10.1093/jas/skx011

Sevillano C.A., Vandenplas J., Bastiaansen J.W.M., Bergsma R. and Calus M.P. 2017. Genomic evaluation for a three-way crossbreeding system considering breed-of-origin of alleles. Genetics Selection Evolution 49:75. doi: 10.1186/s12711-017-0350-1

Sevillano C.A., Vandenplas J., Bastiaansen J.W.M. and Calus M.P. 2016. Empirical determination of breed-of-origin of alleles in three-breed cross pigs. Genetics Selection Evolution 48:55. doi: 10.1186/s12711-016-0234-9

Vandenplas J., Calus M.P., **Sevillano C.A.**, Windig J.J. and Bastiaansen J.W.M. 2016. Assigning breed origin to alleles in crossbred animals. Genetics Selection Evolution 48:61. doi: 10.1186/s12711-016-0240-y

Sevillano, C.A., Mulder H.A., Rashidi H., Mathur P.K. and Knol E.F. 2016. Genetic variation for farrowing rate in pigs in response to change in photoperiod and ambient temperature. Journal of Animal Science 94:3185-97. doi: 10.2527/jas.2015-9915

Sevillano C.A., Lopes M.S., Harlizius B, Hanenberg E.H., Knol E.F. and Bastiaansen J.W.M. 2015. Genome-wide association study using deregressed breeding values for cryptorchidism and scrotal/inguinal hernia in two pig lines. Genetics Selection Evolution 47:18. doi: 10.1186/s12711-015-0096-6

Conference proceedings

Sevillano C.A., Guimarães S.E.F., Silva F.F. and Calus M.P.L. 2018. Breed-specific genome-wide association study for purebred and crossbred performance. World Congress of Genetics Applied to Livestock Production, Auckland, New Zealand.

Godinho R.M., Bergsma R., **Sevillano C.A.**, Silva F.F., Guimarães S.E.F. and Bastiaansen J.W.M. 2018. Genotype by feed interaction in grower-finisher pigs fed different diets. World Congress of Genetics Applied to Livestock Production, Auckland, New Zealand.

Bastiaansen J.W.M. , Guimarães S.E.F., Ali B.M., **Sevillano C.A.**, Godinho R.M., Bergsma R., Lopes M.S., de Mey Y. and Calus M.P.L. 2018. LocalPork-breeding for local conditions. World Congress of Genetics Applied to Livestock Production, Auckland, New Zealand.

Godinho R.M., Bergsma R., **Sevillano C.A.**, Guimarães S.E.F., Silva F.F. and Bastiaansen J.W.M. 2017. Genetic correlation between the residual energy intake of purebred and crossbred pigs. European Association of Animal Production, Tallinn, Estonia.

Godinho R.M., Bergsma R., Silva F.F., **Sevillano C.A.**, Bastiaansen J.W.M. and Guimarães S.E.F. 2017. Genetic parameters for feed efficiency traits in growing-finishing purebred and crossbred pigs. Reunião Anual da Sociedade Brasileira de Zootecnia, Foz do Iguaçu, Brasil.

Godinho R.M., Bergsma R., Silva F.F., **Sevillano C.A.**, Bastiaansen J.W.M. and Guimarães S.E.F. 2017. Genotype by environment interaction for the residual feed intake of growing-finishing pigs fed either a corn-soy or a wheat-barley-by-products based diet. Simpósio Brasileiro de Melhoramento Animal, Riberão Preto, Brasil.

Sevillano C.A., Godinho R.M., Bastiaansen J.W.M., Vandenplas J., Bergsma R. and Calus M.P. 2016. Using phenotypes and genotypes of three-breed cross to

improve breeding value estimation of purebred animals. Reunião da Associação Latino-Americana de Produção Animal, Recife, Brasil.

Sevillano C.A., Vandenplas J., Bastiaansen J.W.M., Bergsma R. and Calus M.P. 2016. Using phenotypes of three-breed cross to improve breeding value estimation of purebred animals. European Association of Animal Production, Dublin, Ireland.

Harlizius B., Lopes M.S., Vandenplas J., **Sevillano C.A.** and Bastiaansen J.W.M. 2016. Genomic prediction of crossbred performance. Joint Annual Meeting, Salt Lake City, Utah, United States of America.

Sevillano C.A., Bastiaansen J.W.M., Vandenplas J., Bergsma R. and Calus M.P. 2015. Genomic prediction for feed efficiency in pigs based on crossbred performance. European Association of Animal Production, Warsaw, Poland.

Vandenplas J., Calus M.P., **Sevillano C.A.**, Windig J.J. and Bastiaansen J.W.M. 2015. Determination of the purebred origin of alleles in crossbred animals. European Association of Animal Production, Warsaw, Poland.

Sevillano C.A., Lopes M.S., Harlizius B. and Bastiaansen J.W.M. 2014. A genome-wide scan reveals novel loci associated with liability to scrotal and inguinal hernia in Large White pigs. World Congress of Genetics Applied to Livestock Production, Vancouver, Canada.

Sevillano C.A., Bloemhof S., van der Waaij E.H. and Knol E.F. 2013. Effect of heat stress during intrauterine development on subsequent litter size in sows. European Association of Animal Production, Nantes, France.

Acknowledgements

The first time I heard about the possibility of doing a PhD was from Egbert Knol when I had just finished my Masters degree and started to work as Research Assistant at Topigs Norsvin Research Center (TNRC) (Topigs Research Center at that time). Egbert proposed to me that I should pursue a PhD in Wageningen in collaboration with TNRC. My first answer was "No, I am not interested in pursuing an academic career" (now I laugh at myself), he simply replied "I will ask you again in 6 months" (thanks Egbert). When mentioned this to my mom, her reaction was "Are you crazy?! Even if you do not pursue an academic career, having a PhD degree is always a card up your sleeve" (thanks mom). I think you all can figure out what my answer was 6 months later.

So, on September 1st 2014 I started this scientific journey. I was not alone. I had a team with me and for that, I am extremely grateful to each one of them. Mario Calus, my co-promotor, he was always making sure (almost daily, at least weekly) that I was on the right path. Thank you Mario for being so accessible (and not only for work related issues), and to trust my work. I think I mentioned many times to people that I was really lucky to have you as a daily supervisor. Rob Bergsma, my supervisor from the TNRC side. Thank you for helping me with all the datasets, breeding value estimations and for being there for me during my struggles as with the SFR paper. Simone Guimaraes and Fabyano da Silva, my supervisors during my adventures in Brazil. Thank you Simone for making my transition to Brazil so much easier and for always making sure my experience at the Universidade federal de Viçosa was good. Thank you Fabyano for the fruitful scientific discussions. John Bastiaansen, the project leader. Thank you for always keeping track of my progress and for always explaining an overview of the journey status. I always enjoyed discussing with you (already back then when you supervised me during my minor Master's thesis), and maybe you haven't realized it yet, but I learned many things from you about what a professional career life is all about. Thank you for being so open in your discussions. In the last stretch of my PhD, Henk Bovenhuis joined the crew as my promotor. Thank you Henk for your support at this very last but most important part of the journey.

To succeed in this type of journey you need colleagues that enrich your daily routine, that give you support and that can create bonds that resemble a family. I was given the opportunity to have not just one, but two families: ABG and TNRC. Being an/a in-between/middle generation/s PhD student, I had the opportunity to meet many other PhD students who were on their own scientific journey. I want to thank you for making this experience an unforgettable one. I will start with the "old" PhD generation: André, Britt, Ewa, Hamed, Juanma, Juan, Juanma, Jovanna,

Katrijn, Gabriel, Gus, Yvonne, Mandy, Mathieu, Mirte, Nancy, Naomi, Robert, Sonia, Yogesh, my traveler girls Amabel, Cori and Merina (btw Iceland is still in the list!), and many others. Yvonneke: thanks for letting me hug you so much even though you hate it, surprisingly now you are the one proposing hugs. Juanma: thanks for your good heart, all the rides to Beuningen although most of the time I was sleeping, eres guay (o mejor dicho "chévere"). The "new" PhD generation: Biaty, Harmen, Ibrahim, Lim, Marieke, Pascal, Margot, Sanne, Shuwen, Zhou, and many others, thank you for everything. A special thanks to those who were my office mates: Yvonne, André, Marcos, Marzi, and Dianne. We had a nice office with plenty of discussion (nonsense and very important issues), devouring chocolate and surviving your last stretch of PhD! A big thanks to my second office mates: Mari and Rodrigo (and Maluma…always present), best office for finishing the PhD, en las buenas y en las malas siempre los Maluma babies, plus the great coffee …better if it's Ethiopian. Mari: thank you for your friendship, your insatiable ability to help people and your positivity…thank you for those lessons Maricucha. Rodrigo: who also started a similar scientific journey in 2014. We quickly understood that walking together was easier (and funnier) than alone, and since then we did the journey together and we finished together. I think we might have finished our PhDs earlier if we wouldn't have discussed so much about politics! Obrigada Pataxó, thank you for adopting me in Brazil and for allowing me to adopt you in all the places that you [mysteriously] always pop up, I believe the bond we built these last few years will last forever, amén Padre! Sonia and Zih-hua: although we shared an office only for a short period, it feels like it was for years, thank you for your good vibes chicas. My gratitude also extends to other ABG colleagues, Ada and Lisette: thanks for helping me with all the paperwork, I have always admired your efficiency. Jan ten Napel: thank you for all of your support with MiXBLUP and your fast responses to any of my concerns. Jérémie: thank you for all of your programming help, and support (almost a supervisor although not legally); merci pour tous les vendredis et dimanches de covoiturage.

The most profound gratitude goes to my TNRC family. First of all (and once again) I would like to thank Egbert. Egbert: you taught me many lessons during my scientific career (to trust in people's results, delegate, to share data and to collaborate, how to present results, etc.) Thank you for always keeping track of my progress, thank you for all of the brainstorming and above everything else thank you for trusting in me and giving me opportunities. Egiel and Arjan: you always had time to explain to me practical issues, you taught me a lot about the company and for that i'm so thankful. Barbara: it is always nice to chat with you, thank you for all

of your concerns but especially GRACIAS for your advice. Naomi and Jascha: nice to start a new job but even better if it is sharing office with these guys. Thanks guys for teaching me so much and for always surrounding me with a fun environment. Roos: always discussing many relevant and irrelevant things, thank you for sharing with me countless rides to Beuningen. Lisanne and Catalin: there were always nice chats in the Feeding Room, I think I should also include Chris here, as a guest office mate, he always brings cheer. My gratitude also extends to Pramod, Marleen, Hans, Dieuwke, Lianne, Diane and the crazy ladies from TA.

I have to give a special thanks to my paranymphs. Coralia: we met back in 2013 (or 2012?) during the GIL course, since then our friendship started or I should rather say our sisterhood (which includes some sister fights as well). We shared moments of happiness and sadness, to achievements, research ideas and scripts, and of course many trips, parties and dinners, gracias por todo Cori vales mil. Marcos: I can't recall the first time we met, but for sure it was before you were my minor thesis supervisor during my Masters. What can I say to you? A million thanks?…Nossa, até isso fica pouco! Thanks for always having the time and patience to teach me R, SQL, Plink, pulling down some datasets and sharing scripts. Thank you for all of the parties, dinners, trips and rides. Thank you for your friendship even if I am "chata pra caralho". Obrigada Marculito. You guys were like an oasis during this scientific journey, I am very glad I met you in this journey and now for standing next to me in the podium.

This journey happened mostly in Wageningen, there I met many people other than the scientific acquaintances and friends I had made. My deepest gratitude goes to my housemates, because in the end anywhere with you guys felt like home. Jagerskamp 66 or the Peruvian Embassy in Wageningen: thanks Ale and Mylu for adopting me in this house that felt like a little Peru, we had a really nice time living together, las quiero mucho. You girls left and new housemates came: Lore, Alvaro, Silvia. I feel really special for sharing the house with you guys because you all enriched my life. N9: thanks Abel, Daisy, Rick and Rio, although only for 3 months you made us feel at home since the first day. I spent 1 year in Brazil, there I had the opportunity to live at Moema's hostel, therefore I firstly need to thank Moema: this old lady was crazy, but she gained a piece of our hearts. André, Javier, Rafa, Yan, Vanessa, mijitas (Laury, Samy y Paola), Carlos, Kevin and Eduardo. I think only the crazy people lived in that house and that's why we had so much fun.

My deepest thanks goes to my parents Nati and Betel and my sister Cristina. Thank you for always being there for me despite our geographical distances. Padres, gracias por fomentar siempre la educación, gracias por incentivarme a alcanzar mis

sueños y sobre todo por abrir sus alas y dejarme volar. Sis, gracias por ser tan positiva y los buenos consejos que siempre reaniman. I want to express my gratitude to my family-in-law, thanks for opening your home to me, thank you for your hospitality and the nice weekends at the farm, no better way to distract the brain from work and to be away from the computer! Même si on doit travailler, le co-working à la ferme est toujours plus chouette qu'au bureau mais pas toujours efficient.

As always, my deepest love and thanks goes to Pierre, my husband, friend, and many times peer reviewer. Thank you for your love even when you have to explain Linux, Fortran and mathematics in general. Thank you for joining me in Brazil without second thoughts and then again in the Netherlands. Thank you for all of the fruitful discussions about my research, although I think this period will now be taken over by discussions on changing diapers. Ik hou van jou (as the real meaning of houden), comme t'aime bien dire.

There are many more people to thank for achieving my doctoral degree, many of whom I forgot, many of which will remain unknown for me. Thank you very much, veel bedankt, muito obrigada, merci beaucoup and muchas gracias.

This research is financially supported by the Netherlands Organisation for Scientific Research (NWO) through the LocalPork project W 08.250.102 in the Food and Business Global Challenges Program and by the Dutch Ministry of Economic Affairs (TKI Agri & Food project 12018) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin.

Study grant from Topigs Norsvin Research Center B.V. during the PhD programme is gratefully acknowledged.

Cover design by Claudia Sevillano and Katty Sandoval.

Printed by: DigiForce || ProefschriftMaken