Genomic selection for improved crossbred performance

Marcos Soares Lopes

Thesis committee

Promotor
Prof. Dr Johan A.M. van Arendonk
Professor of Animal Breeding and Genetics
Wageningen University

Co-promotors
Dr Henk Bovenhuis
Associate professor, Animal Breeding and Genomics Centre
Wageningen University

Dr John W.M. Bastiaansen
Researcher, Animal Breeding and Genomics Centre
Wageningen University

Dr Egbert F. Knol
Head of research
Topigs Norsvin Research Center, Beuningen

Other members
Prof. Dr Bas J. Zwaan, Wageningen University
Prof. Dr Flavio S. Schenkel, University of Guelph, Canada
Prof. Dr Jörn Bennewitz, University of Hohenheim, Stuttgart, Germany
Dr Pieter Knap, Genus-PIC, Rabenkirchen-Fauluck, Germany

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

Genomic selection for improved crossbred performance

Marcos Soares Lopes

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 19 February 2016
at 4 p.m. in the Aula.

Lopes, M.S.
Genomic selection for improved crossbred performance,
202 pages.

PhD thesis, Wageningen University, Wageningen, NL (2016)
With references, with summary in English

ISBN 978-94-6257-631-5

Abstract

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

With the implementation of genomic selection in pig breeding, the genetic progress in purebred populations is expected to increase up to 55% compared to traditional selection based on pedigree information. However, as most of the animals in the pork production system are crossbreds, the increase in genetic progress in purebreds will only be observed on production farms if this progress is expressed in the performance of crossbreds. The main aim of this thesis was to evaluate different models based on genomic information which can be applied to improve performance of crossbred animals. Another aim was to gain insight into genetic architecture of (complex) traits and to investigate how selection history has influenced haplotype patterns of current commercial pigs. This thesis shows that by going beyond traditional genomic selection models, phenotypes can be predicted more accurately. Therefore, these improved models should be considered to improve crossbred performance. Further, this thesis provides important insights into the genetic architecture of the evaluated (complex) traits and also shows evidence that human-driven introgression and selection have shaped the genome of current commercial pig breeds. The research presented in this thesis was performed using data from pigs and the discussion on the practical application of results was focused on pig breeding. The results are relevant for all livestock species where crossbreeding is applied.

Contents

1

General introduction

1.1 From ancient pigs to today's pork

From domestication of ancient pigs to present day, pig farming has undergone a true metamorphosis. One of the key factors that have enabled the more recent and fast developments of pig farming was the establishment of the commercial breeding companies, as they are known today. With breeding companies, the trait recording system has been improved and the selection methods have been enhanced. Selection of the best animals to be the parents of the next generation has moved from only selecting animals with the best phenotypes to selection based on breeding values estimated using sophisticated statistical methods. As part of further developments, recently, estimation of breeding values has progressed from being based on pedigree information to genomic information, or a combination of both. Despite these developments, one thing has not changed since the establishment of breeding companies: the focus has always been mainly on genetic progress in purebred lines. However, most of the pigs in today's pork industry (which generates the pork that arrives at the table of consumers) are crossbred animals and not all progress observed in the purebred lines is transferred to the crossbred animals. Therefore, focusing on improvement of crossbred performance might be an effective way of making all these breeding developments more visible to production farms and potentially to consumers as well. Once the focus of genetic improvement shifts to crossbred performance, a next big change in pig breeding may be necessary: to go beyond additive effects (i.e. breeding values) by accounting for other effects that may contribute to the improvement of crossbred performance.

1.2 Ancient pigs

Current pig breeds originate from the Eurasian wild boar (*Sus scrofa*), which started to be domesticated about 9,000 years ago (Giuffra *et al.* 2000). The domestication process, as shown by mitochondrial DNA studies (Giuffra *et al.* 2000; Larson *et al.* 2005; Megens *et al.* 2008), resulted in two independent groups of breeds: the European and the Asian breeds. However, in the 18th and 19th centuries, Asian pigs were imported to Europe and multiple crosses between European and Asian breeds were made for combining beneficial traits of both groups. This extensive intercross between European and Asian breeds resulted in the gene pool that is found in today's commercial pig breeds (Amills *et al.* 2010).

1.3 Searching for the best pigs

The introduction of Asian pigs in European pig populations coincides with the emergence of the industrial revolution (White 2011). In this period, agriculture activities in northern Europe were intensified as well as trading between Europe and Asia. In this scenario, European farmers realized that it would be beneficial to their pig production to combine characteristics of Asian pigs, such as higher levels of backfat thickness and litter size, with characteristics of European pigs, such as longer body (Jones 1998; White 2011). Therefore, Asian pigs were introduced in European pig populations which contributed to the transformation of the European pigs from a household animal into an industrial product (White 2011).

1.4 Breeding companies and crossbreeding

Since the start of domestication of ancient pigs, animals have been selected to produce offspring that fit the purpose of production. With the industrial revolution, the selection of animals aiming at better offspring performance was intensified. However, only in the 1960s, commercial pig breeding companies arose in the pig industry as a specialized business (Merks 2000). Commercial pig breeding programs, as they are known today, started-up after the work of (Smith 1964) and (Moav & Hill 1966) who showed the benefits of having specialized sire and dam lines. With the specialized purebred sire and dam lines and the establishment of breeding programs, breeders had the opportunity to protect their breeding stock. In addition, they also started crossing these pure lines to produce a better production animal (a crossbred), which is able to capitalize on heterosis and the complementarity of breeds (Visscher *et al.* 2000).

With the specialization of sire and dam lines, the pyramid breeding structure that is still applied in current pig breeding programs was established (See Figure 1.1 for more detail). Since then, instead of producing their own replacement gilts and boars, commercial farms started buying replacement animals from multiplier farms. The specialization of parental lines and also of the farms (being divided in nucleus, multiplier and production farms), together with better phenotype recording and artificial insemination techniques, rapidly increased the genetic gain of production traits, especially in sire lines (Merks 2000; Merks *et al.* 2012). A more pronounced increase in genetic gain of reproduction traits, such as litter size, was observed a few decades later (in the 1990s) when the best linear unbiased prediction (BLUP) via mixed model equations (Henderson 1975) was introduced in pig breeding (Merks 2000).

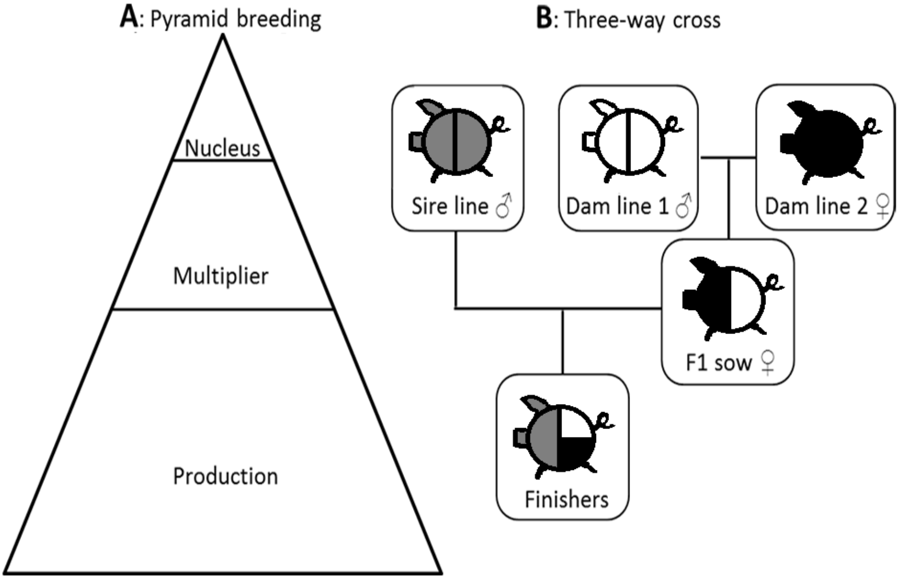

Figure 1.1 Pyramid breeding (A) and three-way cross (B) scheme. Nucleus farms: contain purebred animals (sire and dam lines) which are kept at high health and management level and are sold or provide semen to multipliers farms. Multiplier farms: breed and multiply purebred animals from different lines and cross these lines to produce crossbred animals (F1 sows, two-way cross) that are sold to production farms. Production farms: cross the F1 sows with purebred sires to produce the final crossbred animals (Finishers, three-way cross), which represent the vast majority of animals in the full pig production pyramid.

With BLUP, phenotypic information and family relationships (pedigree information) were included in the prediction of breeding values. Breeding values (i.e. additive genetic effects, Falconer & Mackay 1996) measure the potential of an animal as a parent and are, therefore, widely used as a decision tool for selecting the (purebred) parents of the next generation. However, when it comes to crossbred animals, farmers are interested in the performance rather than breeding values of their animals. The performance of a crossbred animal is influenced by additive effects. However, other effects, such as dominance, imprinting and breed-specific effects, may also have an influence on crossbred performance. If the evaluated trait is influenced by these other effects, the genetic correlation between purebred and crossbred performance will not be 1, which means that the improvements observed in the purebreds may not be translated to improvements in the crossbreds (Dekkers 2007; Hidalgo *et al.* 2015). Therefore, when selecting purebred lines for crossbred performance, the potential use of dominance, imprinting and breed-specific effects could be considered an alternative to improve the efficiency of selection.

Dominance effects are non-additive effects resulting from the interaction between alleles at the same locus. In livestock and plant breeding, dominance effects are of interest because dominance has been suggested as one of the genetic mechanisms explaining heterosis (Davenport 1908; Shull 1908; Bruce 1910; Xiao *et al.* 1995; Visscher *et al.* 2000). The possible benefit of heterosis in the performance of crossbreds is one of the major reasons for applying crossbreeding in pigs. Dominance is also relevant for breeding programs because when dominance variation increases the genetic correlation between purebred and crossbred performance decreases (Wei *et al.* 1991), which will have a direct implication on selection of purebreds for crossbred performance. Results from pedigree-based studies have shown that the ratio between the amount of dominance and additive variance for several traits in pigs ranged from 11% to 78% (Culbertson *et al.* 1998; Misztal *et al.* 1998).

Genomic imprinting is an epigenetic phenomenon in which only the paternally- or maternally-inherited allele is expressed rather than both alleles from homologous chromosomes (Ferguson-Smith 2011). The difference between the two possible heterozygotes (CG and GC, being the first letter the allele inherited from the sire) characterizes imprinting at a phenotypic level (Hager *et al.* 2009). Imprinting effects are expected to have an important contribution to the fetal development and might also affect immediate postnatal growth (Moore & Haig 1991; Reik & Walter 2001). Genes with maternal expression tend to divide all resources between the progeny and maximize the reproductive performance. On the other hand, paternally-expressed genes are expected to promote growth of the offspring and increase demand on maternal resources, characterizing a conflict between maternal and paternal interest (Thomsen *et al.* 2004). Hundreds of imprinted genes (e.g. *IGF2*, *DIO3* and *NOEY2*) have been identified in mammals (http://geneimprint.com/site/genes-by-species), however, the contribution of imprinting to the total genetic variance of economically-important traits is still unknown.

Breed-specific effects are observed when the same allele from a given marker has a different effect on the crossbred phenotype depending on its breed origin. Breed-specific effects are expected to occur when 1) the linkage disequilibrium between the markers and the QTL differ between breeds and 2) when the allele frequency of the QTL is different across breeds, which leads to different allele substitution effect across breeds (Ibanez-Escriche *et al.* 2009). This difference in allele substitution effect may reduce the genetic correlation between purebred and crossbred population, which is detrimental for selecting purebreds for crossbred performance.

Dominance, imprinting and breed-specific effects have not been extensively exploited in pig breeding yet because it is difficult (or not possible at all) to obtain sufficient data to accurately estimate them using pedigree information (De Vries *et al.* 1994; Ibanez-Escriche *et al.* 2009; Vitezica *et al.* 2013). Therefore, the focus of breeders has been on purebred selection and additive genetic effects, even though the final product has been a crossbred animal since the emergence of pig breeding programs. With the development of genomic tools, breeders now have the opportunity to go beyond additive genetics effects and exploit other effects that may contribute to the improvement of crossbred performance.

1.5 Genomics era

Genomic information was introduced in pig breeding in the early 1990s with the DNA test for halothane sensitivity (Fujii *et al*. 1991). Since then, fast developments of genomics offered the opportunity for further application of genomics in breeding programs. The first attempts consisted of application of marker-assisted selection (MAS) using Quantitative Traits Loci (QTL) identified in linkage-based studies. The true benefits of genomics, however, became more pronounced only after the development of dense Single Nucleotide Polymorphisms (SNPs) panels. With SNPs, the identification of QTL moved from linkage-based studies to Genome-Wide Association Studies (GWAS), and MAS was replaced by genomic selection. A brief description of these methods based on genomic information is given below.

1.5.1 QTL mapping and MAS

QTL mapping was pioneered by Sax (1923) working on seed-coat variation in beans. In pigs, the first QTL mapping study was performed by Anderson *et al*. (1994), who reported QTL regions for small intestine length, backfat thickness and fat deposition. Thereafter, thousands of QTL for a large variety of traits have been identified (PigQTLdb, http://www.animalgenome.org/QTLdb/pig.html; Hu *et al*. 2013).

Up to 2010, most QTL mapping studies (linkage-based studies) in pigs was performed using microsatellite markers, which are characterized as LE markers (Dekkers 2004). LE markers are expected to be in linkage equilibrium (LE) with the causative mutation (i.e. there is a random association between the alleles of the markers and QTL). Therefore, to be able to perform QTL mapping using microsatellites, it is required to create linkage disequilibrium (non-random association between the alleles of the markers and the QTL) at least within family or by crossing divergent populations (Hayes & Goddard 2003). Although these linkage studies resulted in the identification of a large number of QTL, their practical application in MAS did not achieve the expected gain in genetic progress (Jonas & de Koning 2015). The reason for this lack of success was mainly the difficulty in identifying reliable markers linked to the QTL due to the structure of the evaluated populations and the low availability of markers (Dekkers 2004; Heffner *et al.* 2009). The use of experimental crosses between divergent populations resulted in the identification of QTL that had different allele frequency between breeds (often alternative QTL alleles were fixed in different breeds) and, therefore, these QTL could not be used for selection within line (Dekkers 2004). Due to the low availability of microsatellites, the precise location of QTL was difficult to be pointed out. Most QTL identified using microsatellites resulted in large confidence intervals, sometimes covering almost the entire chromosome (e.g. Nagamine *et al.* 2003).

In 2009, the first SNP chip for pigs with about 60,000 markers became available (Ramos et al. 2009), opening up new opportunities for a more precise QTL mapping. Using SNPs, the association analysis can be performed in the same population where the potential identified QTL will be used for selection purposes. Differently from microsatellites, some SNPs are linkage disequilibrium (LD) markers (Dekkers 2004). The LD markers are expected to be in LD with the causative mutation, even in purebred populations. Therefore, the use of experimental crosses to create linkage phase for within-breed QTL mapping is not needed because LD between some of the SNPs and the QTL is expected to already exist.

With all its benefits compared to microsatellites, SNPs have been widely used in GWAS. With GWAS, the power to detect new QTL has increased and the confidence intervals of QTL identified in previous studies have narrowed down. However, GWAS findings have not been extensively exploited for selection purposes (under marker-assisted approach). This is because when GWAS started to be performed, the focus of animal breeding was already on the use of all markers (without any pre-selection) in genomic selection approaches (Meuwissen *et al.* 2001). Also, GWAS has mainly focused on the identification of genetic variants with additive effects, neglecting the possibility of identifying QTL with other genetic effects, such as dominance.

1.5.2 Genomic selection

Genomic selection consists of the estimation of genomic breeding values using a large number of markers (typically SNPs) spread across the whole genome (Meuwissen *et al.* 2001). While in MAS only a few markers associated with a QTL are accounted for, in genomic selection all markers are included in the genomic evaluations (without pre-selection based on significance and location of the QTL).

In the last decade, different methods for genomic selection have been developed, including the extensive Bayesian alphabet (Gianola *et al.* 2009; Habier *et al.* 2011). For practical application, however, the most used method in pig breeding is the so-called "single-step" genomic evaluation (Legarra *et al.* 2009; Misztal et al. 2009; Christensen & Lund 2010), which can be considered as an extension of the genomic BLUP (GBLUP) method. Using the Bayesian methods, firstly, SNP effects are estimated using data on animals that were genotyped and phenotyped. In a second step, these SNP effects are used to estimate the breeding values of animals that were genotyped but have no phenotypes (normally young candidates for selection). In GBLUP, the genomic information is used to account for family relationships via the realized genomic relationship matrix (VanRaden 2008) instead of the pedigree-based average relationship matrix (**A** matrix) used in traditional BLUP. Using either Bayesian methods or GBLUP, only genotyped animals can be included in the genomic evaluations and, therefore, extra steps are required for combining information from genotyped and non-genotyped animals. The single-step, however, uses an **H** matrix that combines genomic-based relationships between genotyped animals with pedigree-based relationships with non-genotyped animals. Being able to deal with both genotyped and non-genotyped animals at once is the greatest advantage of single-step over GBLUP and the Bayesian methods.

Although estimation of breeding values has progressed from being based on pedigree information to genomic information, or a combination of both (single-step), the focus of breeders has been on additive effects (i.e. breeding values). However, if the goal is to predict performance rather than breeding values of an animal, models that go beyond additive effects (e.g. that account for dominance, imprinting and breed-specific effects) might be more effective. Further, applying either GBLUP or single-step, it is assumed that all SNPs explain the same amount of variance of the trait (infinitesimal model, Goddard 2009). However, it has been shown that quantitative traits are controlled by a finite number of genes (Hayes & Goddard 2001), which means that some SNPs might explain more variance than others. Therefore, accounting for the genetic architecture of complex traits by using SNPs identified in GWAS might be an interesting approach for enhancing traditional genomic selection methods.

1.6 This thesis

The main goal of this thesis was to evaluate different strategies based on genomic information that could be used to improve performance of crossbred animals. More general, the aim was to evaluate strategies that could be applied to increase production efficiency of production farms by improving performance of crossbred pigs. Another aim was to gain insight into the genetic architecture of the evaluated (complex) traits and to investigate how selection history has influenced haplotype patterns of current commercial pigs.

In **chapter 2**, an SNP regression approach was applied to estimate the contribution of additive, dominance, and imprinting effects to the total genetic variation. The SNP regression method was validated in simulated data and applied to three traits in three purebred pig populations. In **chapter 3**, the accuracy of prediction with a model that accounts for only additive effects was compared with results from a model that accounts for both additive and dominance effects simultaneously using lifetime daily gain records from three purebred pig populations. In **chapter 4**, a genome-wide search for additive and dominance effects on number of teats was performed to find the sources of dominance variance and to investigate the importance of dominance using a high-density SNP panel in a Landrace-based population of pigs. In **chapter 5**, the utility of GWAS findings for the prediction of phenotypes was investigated. Individual SNPs were incorporated in the traditional methods (BLUP and GBLUP) resulting in marker-assisted BLUP (MA-BLUP) and marker-assisted GBLUP (MA-GBLUP). In **chapter 6**, the existence of breed-specific effects was investigated. First, the separate contribution of each purebred line to the genetic variance of a trait observed on (two-way) crossbred animals was estimated. Second, the prediction accuracy of crossbred performance was compared using a traditional model and a model that accounts for breed-specific effects. In **chapter 7**, the hypothesis that the introgression landscape in commercial breeds is shaped mostly by artificial selection was tested by searching for an association between introgressed Asian haplotypes and commercial traits. In the general discussion (**chapter 8**), the results of this thesis were put in a broader perspective. The main focus of that chapter was on the practical application of the results of this thesis in pig breeding with the specific aim to improve crossbred performance. The future perspectives of pig breeding were also discussed.

1.7 References

Amills, M., Clop, A., Ramírez, O., Pérez-Enciso, M. (2010) Origin and genetic diversity of pig breeds. *Encyclopedia of life sciences*. John Wiley & Sons, Chichester.

Andersson, L., Haley, C.S., Ellegren, H., Knott, S.A., Johansson, M., Andersson, K., Andersson-Eklund, L., Edfors-Lilja, I., Fredholm, M., Hansson, I. (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science,* **263,** 1771-1774.

Bruce, A.B. (1910) The Mendelian theory of heredity and the augmentation of vigor. *Science,* **32,** 627-628.

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

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

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

De Vries, A., Kerr, R., Tier, B., Long, T. (1994) Gametic imprinting effects on rate and composition of pig growth. *Theor. Appl. Genet.,* **88,** 1037-1042.

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

Dekkers, J.C. (2004) Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. *J. Anim. Sci.,* **82,** E313-E328.

Falconer, D.S., Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics,* 4th edn. Longmans Green, Harlow.

Ferguson-Smith, A.C. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. *Nat. Rev. Genet.,* **12,** 565-575.

Fujii, J., Otsu, K., Zorzato, F., De Leon, S., Khanna, V.K., Weiler, J.E., O'Brien, P.J., MacLennan, D.H. (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science,* **253,** 448-451.

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

Giuffra, E., Kijas, J., Amarger, V., Carlborg, Ö., Jeon, J.-T., Andersson, L. (2000) The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics,* **154,** 1785-1791.

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

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

Hager, R., Cheverud, J.M., Wolf, J.B. (2009) Relative contribution of additive, dominance, and imprinting effects to phenotypic variation in body size and growth between divergent selection lines of mice. *Evolution,* **63,** 1118-1128.

Hayes, B., Goddard, M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.,* **33,** 209-230.

Hayes, B., Goddard, M.E. (2003) Evaluation of marker assisted selection in pig enterprises. *Livest. Prod. Sci.,* **81,** 197-211.

Heffner, E.L., Sorrells, M.E., Jannink, J.-L. (2009) Genomic selection for crop improvement. *Crop. Sci.,* **49,** 1-12.

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

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

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

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

Jonas, E., de Koning, D.-J. (2015) Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. *Front. Genet.,* **6**.

Jones, G. (1998) Genetic aspects of domestication, common breeds and their origin. *The genetics of the pig* (eds M. Rothschild & A. Ruvinsky), pp. 17-50. CAB International, Oxon.

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

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

Megens, H.-J., Crooijmans, R., San Cristobal, M., Hui, X., Li, N., Groenen, M. (2008) Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genet. Sel. Evol.,* **40,** 103-128.

Merks, J., Mathur, P., Knol, E. (2012) New phenotypes for new breeding goals in pigs. *Animal,* **6,** 535-543.

Merks, J.W. (2000) One century of genetic changes in pigs and the future needs. *BSAS occasional publication***,** 8-19.

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

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

Misztal, I., Varona, L., Culbertson, M., Bertrand, J.K., Mabry, J., Lawlor, T.J., Van Tassel, C.P., Gengler, N. (1998) Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. *Biotechnologie, agronomie, société et environnement,* **2,** 227-233.

Moav, R., Hill, W. (1966) Specialised sire and dam lines. *Anim. Prod.,* **8,** 375-390.

Moore, T., Haig, D. (1991) Genomic imprinting in mammalian development: a parental tug-of-war. *Trends Genet.,* **7,** 45-49.

Nagamine, Y., Haley, C.S., Sewalem, A., Visscher, P.M. (2003) Quantitative trait loci variation for growth and obesity between and within lines of pigs (Sus scrofa). *Genetics,* **164,** 629-635.

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

Reik, W., Walter, J. (2001) Genomic imprinting: parental influence on the genome. *Nat. Rev. Genet.,* **2,** 21-32.

Sax, K. (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. *Genetics,* **8,** 552.

Shull, G.H. (1908) The composition of a field of maize. *J. Hered.***,** 296-301.

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

Thomsen, H., Lee, H., Rothschild, M., Malek, M., Dekkers, J. (2004) Characterization of quantitative trait loci for growth and meat quality in a cross between commercial breeds of swine. *J. Anim. Sci.,* **82,** 2213-2228.

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

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

Vitezica, Z.G., Varona, L., Legarra, A. (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics,* **195,** 1223-1230.

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

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

2

Estimation of additive, dominance, and imprinting genetic variance using genomic data

Marcos S Lopes[1,2], John WM Bastiaansen[2], Luc Janss[3], Egbert F Knol[1], Henk Bovenhuis[2]

[1] Topigs Norsvin Research Center, 6640 AA, Beuningen, the Netherlands; [2] Wageningen University, Animal Breeding and Genomics Centre, 6700 AH, Wageningen, the Netherlands; [3] Aarhus University, Centre for Quantitative Genetics and Genomics, DK-8830, Tjele, Denmark

G3: Genes | Genomes | Genetics (2015) 5(12):2629-2637

Abstract

Traditionally, exploration of genetic variance in humans, plants, and livestock species has mostly been limited to the use of additive effects estimated using pedigree data. However, with the development of dense panels of SNPs (Single Nucleotide Polymorphisms), the exploration of genetic variation of complex traits is moving from quantifying the resemblance between family members to the dissection of genetic variation at individual loci. With SNPs we were able to quantify the contribution of additive, dominance, and imprinting variance to the total genetic variance using an SNP regression method. The method was validated in simulated data and applied to three traits (number of teats, backfat and lifetime daily gain) in three purebred pig populations. In simulated data, the estimates of additive, dominance, and imprinting variance were very close to the simulated values. In real data, dominance effects account for a substantial proportion of the total genetic variance (up to 44%) for these traits in these populations. The contribution of imprinting to the total phenotypic variance of the evaluated traits was relatively small (1-3%). Our results indicate a strong relationship between additive variance explained per chromosome and chromosome length, which has been previously described for other traits in other species. We also show that a similar linear relationship exists for dominance and imprinting variance. These novel results improve our understanding of the genetic architecture of the evaluated traits and shows promise to apply the SNP regression method to other traits and species, including human diseases.

Key words: SNP, variance component, phenotype prediction, pigs

2.1 Introduction

Traditionally, exploration of genetic variance in humans, plants, and livestock species has mostly been limited to the use of additive effects estimated using pedigree data. In this context, the role of genetics in complex traits has been quantified as heritability, e.g. the proportion of the total phenotypic variance explained by additive genetic variance (Visscher & Goddard 2014). However, the estimation of heritability using additive models does not only capture additive gene action but can potentially also capture part of the dominance effects and epistatic interactions (Falconer & Mackay 1996; Hill *et al.* 2008). In addition, traditional additive models ignore imprinting effects, which are also expected to contribute to the genetic architecture and evolution of complex traits (Cheng *et al.* 2013; Lawson *et al.* 2013). Therefore, the proportion of phenotypic variation that is explained by all genetic effects and how much of the total genetic variation is actually due to additive effects is still unclear in modern genetics (Vinkhuyzen *et al.* 2013).

One of the main limitations to better understanding the genetic architecture of complex traits is that typically the data structure does not allow simultaneous estimation of additive, dominance, and imprinting variance (De Vries *et al.* 1994; Vitezica *et al.* 2013). Further, imprinting effects might be confounded with common litter or maternal effects (Tier & Meyer 2004; Wolf & Cheverud 2012). With the development of dense panels of SNPs (Single Nucleotide Polymorphisms), the exploration of genetic variation of complex traits is moving from the quantification of the resemblance between family members to the dissection of genetic variation at individual loci (Vinkhuyzen *et al.* 2013). Because these genetic effects can be estimated, simultaneously, we can now aim to quantify the contribution of additive, dominance and imprinting variance to the total genetic variance.

Dominance effects are of great interest for both plant and livestock breeding because dominance has been suggested as one of the genetic mechanisms explaining heterosis (Davenport 1908; Shull 1908; Bruce 1910; Xiao *et al.* 1995; Visscher *et al.* 2000; Charlesworth & Willis 2009; Shen *et al.* 2014). Only recently, however, with the development of molecular genetics, attempts have been made to quantify and exploit the proportion of genetic variance due to dominance effects in plants (Muñoz *et al.* 2014) and livestock (Toro & Varona 2010; Su *et al.* 2012; Vitezica *et al.* 2013; Zeng *et al.* 2013; Da *et al.* 2014; Nishio & Satoh 2014; Sun *et al.* 2014). Regarding imprinting, hundreds of imprinted genes (e.g. *IGF*2, *DIO*3 and *NOEY*2) have been identified in mammals (http://geneimprint.com/site/genes-by-species), however, the fraction of the total genetic variance due to imprinting effects has not yet been investigated using genomic data.

Moving beyond additive effects, e.g. accounting for dominance and imprinting effects in addition to additive effects, may not only improve our understanding of the genetic architecture of complex traits but also improve the prediction of phenotypes (Lee *et al.* 2008; de Los Campos *et al.* 2010). This could be beneficial, for example, in predicting disease risk in humans (Wray *et al.* 2007) or for establishing mating strategies in plant or animal breeding aimed at maximizing the phenotypic performance of the (crossbred) offspring (Toro & Varona 2010; Muñoz *et al.* 2014). The objective of this study was to estimate the contribution of additive, dominance, and imprinting effects to the total genetic variation using an SNP regression approach. The method was validated in simulated data and applied to three traits in three purebred pig populations.

2.2 Material and methods

2.2.1 Genomic data

A total of 2,013 Landrace, 2,402 Large White and 1,384 Pietrain animals were genotyped using the Illumina Porcine SNP60 Beadchip (Ramos *et al.* 2009). SNPs with call rate <0.95, minor allele frequency <0.05, strong deviation from Hardy-Weinberg equilibrium (χ^2>600), GenCall<0.15, unmapped SNPs and SNPs located on sex chromosomes, according to the Sscrofa10.2 assembly of the reference genome (Groenen *et al.* 2012), were excluded from the data set. After quality control, 34,912 SNPs for Landrace, 36,578 SNPs for Large White and 38,116 SNPs for Pietrain out of the initial 64,232 SNPs were kept for phasing procedures. All animals had a frequency of missing genotypes <0.05, therefore, no animals were excluded due to high frequency of missing genotypes.

Phasing and imputation of missing genotypes were performed within each line using AlphaImpute (Hickey *et al.* 2011), which combines genomic and pedigree information to determine the parental origin of alleles. The pedigree depth used in this analysis was up to 5 generations (between genotyped animals).

For each SNP of each individual, AlphaImpute (Hickey *et al.* 2011) generates two probabilities: P_1 being the probability that a specific allele was received from its father, say an allele G of a G/C SNP, and P_2 the probability that the same allele was received from its mother. Considering a heterozygous animal (GC) where the G allele was inherited, with certainty, from its father (and therefore a C allele from its mother), the probabilities would be P_1= 1 and P_2= 0. To obtain the regressors that allow the estimation of additive (*regA*), dominance (*regD*) and imprinting (*regI*) genetic variance, the following transformation of these probabilities was applied: *regA*= [($P_1 + P_2$) – 1], *regD*= (|$P_1 – P_2$|) and *regI*= ($P_1 – P_2$). Thus, the genotypes (GG, GC, CG, CC) were recoded as (-1, 0, 0, 1), (0, 1, 1, 0) and (0, -1, 1, 0) to evaluate additive, dominance and imprinting variances, respectively. To ensure accurate phasing, only animals which had both parents or at least one parent and one sib genotyped were used in further steps. Due to these restrictions, 1,538 Landrace, 1,595 Large White, and 1,272 Pietrain animals were available for the estimation of variance components.

2.2.2 Simulation

To verify whether our data structure and statistical model allow disentangling additive, dominance, and imprinting effects, we simulated for the Landrace population a trait with additive, dominance and imprinting effects, with mean 0, total genetic variance equal to 0.30 and total phenotypic variance equal to 1. Real genotypes of this population were used in the simulation procedures. A total of 15 SNPs with minor allele frequency between 0.45 and 0.50 were randomly selected to have an effect on the trait (QTLs): five with only additive effects, five with only dominance effects and five with only imprinting effects. These SNPs were located on different chromosomes that were also randomly selected. Each QTL accounted for 2% of the total phenotypic variance. Therefore, the additive, dominance, and imprinting heritabilities were 10% each. The genetic variance (V_G) of a single QTL was defined as described by De Koning *et al.* (2002):

$$V_G = V_a + V_d + V_i$$

$$V_a = 2pq[a + d(p-q)]^2$$

$$V_d = (2pqd)^2$$

$$V_i = 2pqi^2$$

where V_a, V_d and V_i are, respectively, the additive, dominance, and imprinting variances and *a*, *d* and *i* are, respectively, the additive, dominance, and imprinting effects of a given QTL with allele frequencies *p* and *q*. As each QTL was simulated to either have an additive, dominance or imprinting effect, *a* of each additive QTL, *d* of each dominant QTL and *i* of each imprinted QTL could be defined as: $a = \sqrt{(0.02/2pq)}$, $d = \sqrt{0.02}/2pq$, and $i = \sqrt{(0.02/2pq)}$. The simulated phenotype of the j^{th} animal then becomes:

$$phenotype_j = \sum_{s=1}^{n=15} (regA_{js}a_s + regD_{js}d_s + regI_{js}i_s) + e_j$$

where *n* is the number of QTL (SNPs) affecting the trait, *e* is a random environmental component sampled from a random distribution with variance equal to 0.70 and *regA*, *regD* and *regI* are defined as described in the genomic data section above. We generated 10 replicates of this simulation and SNPs that met the selection criteria were allowed to have an effect in only one of the replicates.

2.2.3 Phenotypes

The phenotypic data consisted of the traits number of teats, backfat, and lifetime daily gain, which corresponds to the average daily weight increase from birth to ~120 kg. The response variables used to estimate the genetic variances were phenotypes pre-adjusted for fixed effects instead of the original observations. The pre-adjustment was based on a larger dataset that included all contemporaneous animals of the genotyped animals, rather than just using the group of genotyped animals. Using this larger data set allowed us to account more accurately for contemporary group effects. The fixed effects estimate used for the pre-adjustment of the phenotypes were obtained fitting a single trait pedigree-based linear model using ASReml v3.0 (Gilmour *et al.* 2009). The model for number of teats consisted of sex and herd-year-season as fixed effects and an additive genetic effect and a residual as random effects. The model for backfat consisted of sex, herd-year-week and weight as fixed effects and an additive genetic effect, common litter effect and a residual as random effects. For lifetime daily gain, the model consisted of sex and herd-year-week as fixed effects and an additive genetic effect, common litter effect and a residual as random effects. For the Landrace population, the final data set consisted of 141,248 animals for number of teats, 36,413 animals for backfat and 37,071 animals for lifetime daily gain. For the Large White population, the final data set consisted of 156,065 animals for number of teats, 41,192 animals for backfat and 41,740 animals for lifetime daily gain. For the Pietrain population, the final data set consisted of 33,964 animals for backfat and 31,184 animals for lifetime daily gain. The trait number of teats was not recorded in the Pietrain population. Descriptive statistics of the phenotypes are shown in Table 2.1.

Table 2.1 Descriptive statistics

Dataset	**NT** (units)			**BF** (mm)			**DG** (g)		
	N	**μ**	**SD**	**N**	**μ**	**SD**	**N**	**μ**	**SD**
Landrace									
All	141,248	15.27	1.06	36,413	12.47	2.54	37,071	598.47	70.63
Genotyped	1,538	15.62	1.04	1,405	12.55	2.20	1,394	628.27	62.93
Large White									
All	156,065	15,08	1.05	41,192	12,38	2.49	41,740	632.30	71.78
Genotyped	1,595	15.40	0.98	1,453	12.20	2.37	1,468	649.86	69.09
Pietrain									
All				33,964	7.98	1.49	31,184	603.86	75.89
Genotyped				1,272	7.82	1.28	1,145	630.70	65.64

Number of animals with phenotypic information (N), mean (μ) and standard deviation (SD) of the traits number of teats (NT), backfat (BF) and average daily gain from birth to ~120 kg (DG).

2.2.4 Statistical analyses

Parameters were estimated using models with random regression on SNP genotypes. Single trait within-line analyses were performed with three different models implemented in the program BayZ (http://www.bayz.biz/), the same for both real and simulated data:

$$\boldsymbol{y} = \mathbf{1}\mu + (\mathbf{Lb}) + \mathbf{Aa} + \mathbf{e} \qquad \text{(MA model)}$$

$$\boldsymbol{y} = \mathbf{1}\mu + (\mathbf{Lb}) + \mathbf{Aa} + \mathbf{Dd} + \mathbf{e} \qquad \text{(MAD model)}$$

$$\boldsymbol{y} = \mathbf{1}\mu + (\mathbf{Lb}) + \mathbf{Aa} + \mathbf{Dd} + \mathbf{Ii} + \mathbf{e} \qquad \text{(MADI model)}$$

where **y** is a vector of pre-adjusted phenotypic observations; μ is the mean of the populations and **1** a vector of ones; **L** is the design matrix for the common litter effects (only used for backfat and lifetime daily gain); **b** is an unknown vector of common litter effects; **A**, **D** and **I** are design matrices with regressors for additive, dominance, and imprinting effects respectively; **a**, **d,** and **i** are unknown vectors of additive, dominance, and imprinting effects respectively, and **e** is a vector of residuals. The entries of the design matrices **A**, **D,** and **I** are regressors calculated from the observed phased probabilities of the marker genotypes (*regA*, *regD* and *regI*), as described in the genomic data section above.

Assumed distributions were: **a** ~ $N(\mathbf{0},\mathbf{I}\sigma_a^2)$, **d** ~ $N(\mathbf{0},\mathbf{I}\sigma_d^2)$, **i** ~ $N(\mathbf{0},\mathbf{I}\sigma_i^2)$, **b** ~ $N(\mathbf{0},\mathbf{I}\sigma_{\mathrm{L}}^2)$ and **e** ~ $N(\mathbf{0},\mathbf{I}\,\sigma_e^2)$, with σ_a^2, σ_d^2, σ_i^2 being the per-SNP variance for additive, dominance, and imprinting effects, and σ_L^2 and σ_e^2 the common litter and residual variance, respectively. The model was fitted using a Bayesian approach in the Bayz software package (http://www.bayz.biz/) as described by Krag *et al.* (2013) to estimate variance components and heritabilities in SNP-based models. The prior distributions for unknown variance parameters were set as unbounded uniform, which makes the Bayesian posterior distribution mathematically identical to the likelihood. The generated Monte Carlo chain starts with all regression parameters and other location parameters at zero, and all variance parameters at 1, and blocked Gibbs samplers are employed to facilitate mixing. Each model was run as a single chain with a length of 500,000 (real data) and 100,000 (simulated data) which was sampled each 100 iterations. The first 50,000 iterations of each run were regarded as burn-in period.

As the evaluated models were SNP-based models, they do not readily provide estimates of total explained variance. One way to obtain the total explained variance from a model term like **Aa** is to write *var*(**Aa**)=**AA**'σ_a^2, and compute or evaluate the expected average diagonal of **AA**' to provide the constant to scale the per-SNP explained variance to total explained variance. In this way total variance in the models can be expressed as $[(\sigma_{\mathbf{L}}^2)+\sigma_{\mathbf{Aa}}^2$ + $\sigma_e^2]$ for MA, $[(\sigma_{\mathbf{L}}^2)+\sigma_{\mathbf{Aa}}^2$ + $\sigma_{\mathbf{Dd}}^2$ + $\sigma_e^2]$ for MAD and $[(\sigma_{\mathrm{L}}^2)+\sigma_{\mathbf{Aa}}^2$ + $\sigma_{\mathbf{Dd}}^2$ + $\sigma_{\mathbf{Ii}}^2$ + $\sigma_e^2]$ for MADI, with $\sigma_{\mathbf{Aa}}^2$=**AA'**σ_a^2 (total additive variance), $\sigma_{\mathbf{Dd}}^2$=**DD'**σ_d^2 (total dominance variance), and $\sigma_{\mathbf{Ii}}^2$=**II'**σ_i^2 (total imprinting variance). Only for backfat and lifetime daily gain, σ_{L}^2 was included. Alternatively, the variance contributed by a random effect could be estimated by evaluating the sample variance of the entries of the vectors **Aa, Dd**, and **Ii** at each iteration of the Gibbs sampler (Sorensen *et al.* 2001). This has the advantage that the posterior standard deviations can also be obtained for total explained variance, and explained variances can be split easily by chromosome. The latter is done by computing var(**Aa**) per MCMC cycle only for the part of the covariates in **A** and matching regression parameters in **a** that belong to a particular chromosome. The narrow sense heritability was defined as $\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$ and the proportion of phenotypic variance explained by dominance and imprinting effects were defined as $\sigma_{\mathbf{Dd}}^2/\sigma_{\mathrm{P}}^2$ and $\sigma_{\mathbf{Ii}}^2/\sigma_{\mathrm{P}}^2$ respectively.

The portioning of the genetic variance as described above has been defined as the “genotypic model” (Vitezica *et al.* 2013), which implies that $\sigma_{\mathbf{Aa}}^2$, $\sigma_{\mathbf{Dd}}^2$, and $\sigma_{\mathbf{Ii}}^2$ are the variance of the genotypic additive, dominance, and imprinting values, respectively. The genotypic model and the breeding (or classical) model are statistically equivalent (i.e. they lead to the same probability model). However, the parameters obtained with these models have different interpretations. In the genotypic model, the additive variance is the variance of additive effects (average difference between homozygotes), while in the breeding model, additive effects are functions of allele substitution effects. To make the variance estimates from the genotypic model comparable to the estimates of the breeding model, a transformation of these results was proposed by Vitezica *et al.* (2013). We have applied the transformation proposed by Vitezica *et al.* (2013) to the estimates from the MAD model and included the results in the supporting material (File 2.S1).

Finally, we evaluated whether the proportion of variance explained by a single chromosome was related to its physical length. The length of a given chromosome was defined as the distance between the first and the last SNP on this chromosome according to the *Sus scrofa* 10.2 assembly (Groenen *et al.* 2012). The relationship between variance explained and the physical length of the chromosome was expressed as the coefficient of determination (r^2) from the regression of variance explained on physical length. Variance explained by each chromosome individually was obtained based on the effects of SNPs on that chromosome. SNP effects were from the analyses where all SNPs from all chromosomes were evaluated simultaneously. Variance per chromosome was estimated for all three models evaluated (MA, MAD and MADI).

2.2.5 Model comparison

Models were compared using the Deviance Information Criterion (DIC, Spiegelhalter *et al.* 2002). DIC is widely used for Bayesian model comparison and is analogous to the Akaike Information Criterion (AIC, Akaike 1974). DIC combines a measure of model fit (the expected deviance) with a measure of model complexity (the effective number of parameters) over all iterations after burn-in. The model with lowest DIC is chosen as the best fitting model (Spiegelhalter *et al.* 2002).

2.3 Results

2.3.1 Simulation

Average narrow sense heritability over the 10 replicates of the simulated trait was estimated at 0.116, 0.092 and 0.098 using the MA, MAD and MADI models, respectively (Table 2.2). The average proportion of phenotypic variance explained by dominance effects was 0.111 using the MAD model and 0.097 using the MADI model. Using the MADI model, the average proportion of phenotypic variance explained by imprinting effects was 0.103.

The pairwise sampling correlation between additive, dominance, imprinting and error variance are shown on Table 2.S1. The average correlation between the different variances of the 10 replicates of the simulated trait ranged from -0.586 to 0.018. The strongest correlations (-0.586 to -0.186) were observed between the residual variance and the variance of the three components of genetic variance evaluated. The lowest correlations (approximately zero) were observed between imprinting and additive variance and between imprinting and dominance variance.

2.3.2 Real data

For the trait number of teats, the narrow sense heritability estimated using MA was 0.319 in the Landrace and 0.343 in the Large White population (Tables 2.3 and 2.4). Using both MAD and MADI, the narrow sense heritability was approximately the same in both populations (~0.306). The estimates of the proportion of phenotypic variance explained by dominance effects, however, showed a large difference between populations. The proportion of phenotypic variance explained by dominance effects for number of teats in the Landrace population (0.039) was a little bit more than one-third that of the Large White population (~0.100). The proportion of phenotypic variance explained by imprinting effects for number of teats was low in both populations, 0.015 in the Landrace and 0.010 in the Large White.

For the trait backfat, the narrow sense heritability estimated using MA was 0.520, 0.390, and 0.419 in the Landrace, Large White and Pietrain populations respectively. Additive heritabilities decreased to 0.469, 0.353, and 0.394 in the Landrace, Large White, and Pietrain populations, respectively, when the MA model was replaced by MADI (Tables 2.3-5). MAD and MADI resulted in almost the same estimates of narrow sense heritability and of the proportion of phenotypic variance explained by dominance effects for backfat in all populations and narrow sense heritability was always lower than the estimate based on MA. Similar to number of teats, the proportion of phenotypic variance explained by dominance effects for backfat was variable between populations with estimates of 0.102 in the Landrace, 0.146 in the Large White, and 0.064 in the Pietrain population. The proportion of phenotypic variance explained by imprinting effects for backfat was 0.017 in the Landrace, 0.029 in the Large White, and 0.020 in the Pietrain population.

Finally, for the trait lifetime daily gain, the narrow sense heritability estimated using MA was 0.267 in the Landrace, 0.241 in the Large White, and 0.314 in the Pietrain population (Tables 2.3-5). Again, the narrow sense heritability and the proportion of phenotypic variance explained by dominance effects were similar using either MAD or MADI in all populations and additive estimates were smaller than those from the MA model. The proportion of phenotypic variance explained by dominance effects estimated with MADI were higher for lifetime daily gain than for the other two traits in the Landrace (0.158) and Pietrain populations (0.199) and similar to dominance for backfat in the Large White population (0.130). The proportion of phenotypic variance explained by imprinting effects of lifetime daily gain was again low, as for the other 2 traits (<0.020 in all populations).

2.3.3 Model comparison

Based on the estimated DIC (Table 2.6), MAD and MADI presented a better fit to the data than MA for all traits in all populations, except for number of teats in the Landrace population (which was the trait with the lowest proportion of dominance variance in this study). The MADI model was slightly superior to MAD for backfat in the Landrace population, for all traits in the Large White population, and for lifetime daily gain in the Pietrain population.

2.3.4 Variance explained by individual chromosomes

After estimation of SNP effects using all SNPs simultaneously, we estimated the additive, dominance, and imprinting variance using the complete set of SNPs and also the variances per individual chromosomes. Estimates obtained from genome-wide SNPs were close to results from adding up the contributions of individual chromosomes. The largest difference was observed for number of teats in the Landrace population. The additive variance estimated using all SNPs was 0.33 for number of teats, while adding up the contributions of individual chromosomes resulted in an estimate of 0.39 for number of teats. The variance explained per chromosome for all traits in all populations using MADI is shown in File S2. We report results obtained using MADI. Very similar results were observed with all three models (MA, MAD, and MADI).

For all traits, the proportion of additive, dominance, and imprinting variance explained per chromosome showed a strong linear relationship with chromosome length (r^2 ranging from 0.84 to 0.94).

2.4 Discussion

2.4.1 Simulation

Analysis of simulated data with the MADI model resulted in estimates of additive, dominance, and imprinting variance that were very close to the simulated values (Table 2.2). However, the additive variance was overestimated when using the MA model and the dominance variance was overestimated with the MAD model. This is a sensible result as all models, except the MADI model, are under-parameterized. A model that allows a proper dissection of the variances should yield variance components that are uncorrelated (Hill 2010). To test this, we calculated the pairwise sampling correlations between the error, additive, dominance, and imprinting variance (Table 2.S1). The pairwise sampling correlations were moderate and mostly negative (Table 2.S1). Based on these simulation results we, therefore, concluded that data structure and the methodology will allow us to disentangle additive, dominance and imprinting variance, although this simulated scenario may not be representative of the genetic architecture of a real complex trait.

2.4.2 Real data

In all populations, we observed a reduction in the narrow sense heritability of all evaluated traits when dominance effects were accounted for (e.g. using MAD instead of MA). The smallest decrease in narrow sense heritability was observed for number of teats (4.2%) and the highest for lifetime daily gain (21.3%), both in the Landrace population (Table 2.3). The broad-sense heritability (sum of the heritabilities due to all genetic effects used in the model) of all evaluated traits increased in all three populations when dominance and imprinting effects were added to the model. The broad-sense heritability of lifetime daily gain was >30% higher when using MADI compared to using MA (Tables 2.3-5). A reduction of additive genetic variance and an increase in the broad-sense heritability was previously reported (Su *et al.* 2012) when non-additive genetic effects were included in the model to evaluate daily gain in pigs. For height in trees, the narrow sense heritability was found to reduce by 26% with the inclusion of non-additive genetic effects in the model (Muñoz *et al.* 2014). According to Muñoz *et al.* (2014) and Pante *et al.* (2002), such a reduction in the additive genetic variance should be expected when dominance effects are present because dominance effects also contribute to the additive genetic variance in the MA model. However, (Vitezica *et al.* 2013) reported that such a reduction in the additive variance should be seen as an underestimation, as a consequence of overestimating the dominance variance. These authors described that when dominance is fitted in genotypic models (such as the one applied in the current study and by Su *et al.* (2012) and Muñoz *et al.* (2014), the part of the dominance effect that contributes to the allele substitution effect is shifted to the dominance variance. Because of this shift, the estimates from genotypic models are not directly comparable to pedigree-based estimates. Therefore, with methodology applied in the current study, we could interpret the decrease in the narrow sense heritability from the MA model compared to the MAD model as the contribution of dominance effects to the additive genetic variance. If the aim is to estimate breeding values and dominance deviations (i.e. the traditional breeding model), the parameterization proposed by (Vitezica *et al.* 2013) should be applied.

Table 2.2 Estimated variance components and proportion of phenotypic variance (σ_{P}^2) explained by additive, dominance, and imprinting effects for the simulated data.

Model	Variance components				Variance explained		
	σ_e^2	$\sigma_{\mathbf{Aa}}^2$	$\sigma_{\mathbf{Dd}}^2$	$\sigma_{\mathbf{Ii}}^2$	$\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$ *	$\sigma_{\mathbf{Dd}}^2/\sigma_{\mathrm{P}}^2$	$\sigma_{\mathbf{Ii}}^2/\sigma_{\mathrm{P}}^2$
MA	0.869 ± 0.035	0.115 ± 0.036			0.116 ± 0.035		
MAD	0.789 ± 0.038	0.091 ± 0.035	0.110 ± 0.048		0.092 ± 0.035	0.111 ± 0.047	
MADI	0.698 ± 0.041	0.097 ± 0.033	0.097 ± 0.057	0.102 ± 0.028	0.098 ± 0.027	0.097 ± 0.061	0.103 ± 0.032
Simulated	0.700	0.100	0.100	0.100	0.100	0.100	0.100

σ_e^2: residual variance; $\sigma_{\mathbf{Aa}}^2$: total additive variance; $\sigma_{\mathbf{Dd}}^2$: total dominance variance; $\sigma_{\mathbf{Ii}}^2$: total imprinting variance. *$\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$: narrow sense heritability.

Table 2.3 Variance components and proportion of phenotypic variance (σ_{P}^2) explained by additive, dominance, and imprinting effects for the real data in the Landrace population.

Trait	Model	Variance components					Variance explained		
		σ_{e}^2	σ_{L}^2	$\sigma_{\mathbf{Aa}}^2$	$\sigma_{\mathbf{Dd}}^2$	$\sigma_{\mathbf{Ii}}^2$	$\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$ *	$\sigma_{\mathbf{Dd}}^2/\sigma_{\mathrm{P}}^2$	$\sigma_{\mathbf{Ii}}^2/\sigma_{\mathrm{P}}^2$
NT	**MA**	0.739 ± 0.042		0.346 ± 0.041			0.319 ± 0.035		
	MAD	0.714 ± 0.046		0.334 ± 0.045	0.042 ± 0.034		0.306 ± 0.037	0.039 ± 0.031	
	MADI	0.700 ± 0.046		0.334 ± 0.043	0.042 ± 0.036	0.016 ± 0.014	0.305 ± 0.036	0.039 ± 0.033	0.015 ± 0.012
BF	**MA**	1.232 ± 0.120	0.431 ± 0.115	1.803 ± 0.150			0.520 ± 0.035		
	MAD	1.082 ± 0.138	0.362 ± 0.117	1.604 ± 0.199	0.345 ± 0.172		0.472 ± 0.047	0.102 ± 0.052	
	MADI	1.031 ± 0.135	0.369 ± 0.119	1.596 ± 0.195	0.345 ± 0.166	0.057 ± 0.038	0.469 ± 0.046	0.102 ± 0.050	0.017 ± 0.011
DG	**MA**	1,435 ± 111	380 ± 106	662 ± 113			0.267 ± 0.043		
	MAD	1,213 ± 141	314 ± 105	525 ± 122	446 ± 185		0.210 ± 0.047	0.178 ± 0.071	
	MADI	1,185 ± 138	319 ± 104	554 ± 124	395 ± 179	47 ± 34	0.221 ± 0.048	0.158 ± 0.070	0.019 ± 0.014

σ_e^2: residual variance; σ_{L}^2: common litter variance; $\sigma_{\mathbf{Aa}}^2$: total additive variance; $\sigma_{\mathbf{Dd}}^2$: total dominance variance; $\sigma_{\mathbf{Ii}}^2$: total imprinting variance; *$\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$: narrow sense heritability.

Table 2.4 Variance components and proportion of phenotypic variance (σ_{P}^2) explained by additive, dominance, and imprinting effects of the real data in the Large White population.

Trait	Model	Variance components					Variance explained		
		σ_{e}^2	σ_{L}^2	$\sigma_{\mathbf{Aa}}^2$	$\sigma_{\mathbf{Dd}}^2$	$\sigma_{\mathbf{Ii}}^2$	$\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$ *	$\sigma_{\mathbf{Dd}}^2/\sigma_{\mathrm{P}}^2$	$\sigma_{\mathbf{Ii}}^2/\sigma_{\mathrm{P}}^2$
NT	**MA**	0.637 ± 0.032		0.333 ± 0.032			0.343 ± 0.029		
	MAD	0.578 ± 0.041		0.299 ± 0.037	0.094 ± 0.043		0.307 ± 0.034	0.097 ± 0.044	
	MADI	0.563 ± 0.040		0.300 ± 0.038	0.103 ± 0.043	0.010 ± 0.009	0.307 ± 0.035	0.105 ± 0.043	0.010 ± 0.009
BF	**MA**	1.086 ± 0.090	0.371 ± 0.084	0.931 ± 0.093			0.390 ± 0.034		
	MAD	0.927 ± 0.105	0.307 ± 0.085	0.848 ± 0.105	0.320 ± 0.125		0.353 ± 0.039	0.133 ± 0.051	
	MADI	0.834 ± 0.115	0.302 ± 0.084	0.852 ± 0.104	0.353 ± 0.134	0.071 ± 0.041	0.353 ± 0.039	0.146 ± 0.055	0.029 ± 0.017
DG	**MA**	1,867 ± 129	282 ± 114	682 ± 100			0.241 ± 0.033		
	MAD	1,715 ± 144	234 ± 108	590 ± 120	300 ± 157		0.208 ± 0.040	0.106 ± 0.055	
	MADI	1,659 ± 149	230 ± 108	557 ± 129	370 ± 178	34 ± 29	0.195 ± 0.043	0.130 ± 0.062	0.012 ± 0.010

σ_e^2: residual variance; σ_{L}^2: common litter variance; $\sigma_{\mathbf{Aa}}^2$: total additive variance; $\sigma_{\mathbf{Dd}}^2$: total dominance variance; $\sigma_{\mathbf{Ii}}^2$: total imprinting variance; *$\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$: narrow sense heritability.

Table 2.5 Variance components and proportion of phenotypic variance (σ_{P}^2) explained by additive, dominance, and imprinting effects for the real data in the Pietrain population.

Trait	Model	Variance components					Variance explained		
		σ_{e}^2	σ_{L}^2	$\sigma_{\mathbf{Aa}}^2$	$\sigma_{\mathbf{Dd}}^2$	$\sigma_{\mathbf{Ii}}^2$	$\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$ *	$\sigma_{\mathbf{Dd}}^2/\sigma_{\mathrm{P}}^2$	$\sigma_{\mathbf{Ii}}^2/\sigma_{\mathrm{P}}^2$
BF	**MA**	0.610 ± 0.050	0.096 ± 0.041	0.510 ± 0.054			0.419 ± 0.038		
	MAD	0.570 ± 0.056	0.082 ± 0.040	0.476 ± 0.064	0.085 ± 0.060		0.392 ± 0.046	0.070 ± 0.050	
	MADI	0.551 ± 0.060	0.084 ± 0.041	0.481 ± 0.062	0.078 ± 0.058	0.024 ± 0.019	0.394 ± 0.044	0.064 ± 0.047	0.020 ± 0.016
DG	**MA**	1,804 ± 143	223 ± 123	931 ± 130			0.314 ± 0.040		
	MAD	1,488 ± 163	148 ± 102	735 ± 150	584 ± 195		0.248 ± 0.047	0.198 ± 0.065	
	MADI	1,468 ± 172	154 ± 101	718 ± 169	591 ± 239	32 ± 31	0.242 ± 0.054	0.199 ± 0.080	0.011 ± 0.010

σ_e^2: residual variance; σ_{L}^2: common litter variance; $\sigma_{\mathbf{Aa}}^2$: total additive variance; $\sigma_{\mathbf{Dd}}^2$: total dominance variance; $\sigma_{\mathbf{Ii}}^2$: total imprinting variance; *$\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$: narrow sense heritability.

Table 2.6 Deviance Information Criterion (DIC).

Population	Trait	MA	MAD	MADI
Landrace	NT	**1,418**	1,431	1,458
	BF	2,199	2,133	**2,093**
	DG	11,433	**11,325**	11,333
Large White	NT	1,225	1,206	**1,180**
	BF	2,174	2,051	**1,992**
	DG	12,894	12,861	**12,838**
Pietrain	BF	1,089	**1,051**	1,053
	DG	10,059	9,954	**9,952**

Numbers given in bold indicate the lowest DIC (best fit) obtained for each trait in each population.

The substantial higher values for the broad-sense heritability compared to the narrow-sense heritability indicate that dominance effects make an important contribution to the genetic variance of the evaluated traits and populations in this study, especially to the trait lifetime daily gain. For lifetime daily gain in the Pietrain population using MADI, we observed the highest proportion of phenotypic variance explained by dominance effects (0.198), with a ratio between dominance and additive variance of 0.82. In other lines, the ratios were also high. Using the same model, the ratio between dominance and additive variance was 0.71 in the Landrace and 0.66 in the Large White population. Moreover, our results also show that additive variance accounts for the largest fraction of the genetic variance. This is in agreement with a previous study that described that additive variance is expected to account for >50% (often about 100%) of the total genetic variance (Hill *et al.* 2008). However, the estimates of the additive variance of the traits in the populations here evaluated might still become smaller if epistatic interactions exist and would be included as a separate variance component. Although the role of epistatic interactions in the genetic architecture of complex traits has been investigated in different species (Le Rouzic *et al.* 2008; Su *et al.* 2012; Muñoz *et al.* 2014), we did not attempt to estimate epistatic variance because the power to identify these effects in segregating populations was expected to be low (Melchinger *et al.* 2007). To be able to detect epistatic interactions in outbred populations, loci with these effects should have a large effect and segregate with an intermediate frequency (Hill *et al.* 2008).

Individual loci that show the effect of imprinting have been identified for pigs, such as *IGF*2 (insulin-like growth factor 2 gene, Jeon *et al.* 1999; Nezer *et al.* 1999). The contribution of imprinting to the total genetic variance is however still unknown. No reports were found in the literature that attempt to quantify total imprinting variance using genomic data. In this study, the trait with the highest proportion of phenotypic variance explained by imprinting effects was backfat (0.017 in the Landrace, 0.029 in the Large White, and 0.020 in the Pietrain population), although the estimates were still quite low, in comparison to the amount of narrow sense heritability and the proportion of phenotypic variance explained by dominance effects. In mice, a gene expression QTL mapping study for body composition traits showed that imprinting QTLs accounted for only a limited amount of the phenotypic variance (<2.50%) for most traits (Cheng *et al.* 2013). In a pedigree-based study in pigs, it was shown that 5-7% of the phenotypic variance of backfat and 1-4% of growth rate was explained by paternal imprinting (De Vries *et al.* 1994). That study also showed maternal imprinting to account for 2-5% and 3-4% of the phenotypic variance of backfat and growth rate, respectively. Although our genomic estimates are lower than the pedigree-based estimates described by De Vries *et al.* (1994), the two results agree that imprinting effects are more important for backfat than for growth traits. The amount of phenotypic variance of number of teats due to imprinting effects has not yet been described in the literature. However, two imprinted QTLs have been reported on chromosomes 2 and 12 (Hirooka *et al.* 2001). These two QTL explained 1.3 and 2.2% of the phenotypic variance of number of teats, while in our study the proportion of phenotypic variance explained by imprinting effects of number of teats in both populations was ≤1.5%. The larger imprinting variances found by Hirooka *et al.* (2001) may in part be explained by the design, an experimental F2 population, analysed in their QTL study.

Due to the low proportion of phenotypic variance explained by imprinting, the relevance of estimating imprinting effects may be low when the aim is to predict the phenotypes number of teats, backfat and lifetime daily gain in the evaluated populations. However, this study shows that, when present, dominance and imprinting variance can be detected and estimated with an SNP regression model. Using pedigree-based analysis this would typically not be feasible, for different reasons. Firstly, the estimation of dominance variance using pedigree data requires data from large full-sib families (Vitezica *et al.* 2013), which is often not available in humans and livestock species. Secondly, pedigree-based methods have difficulties in disentangling imprinting from maternal and permanent environmental effects (Tier & Meyer 2004; Wolf & Cheverud 2012). Thirdly, pedigree-based analysis often overestimates additive variance (Vinkhuyzen *et al.* 2013) and underestimates dominance variance (Muñoz *et al.* 2014). Although the use of genome-wide markers, compared to pedigree data, has been described as a more precise alternative to partition the genetic variance (Lee *et al.* 2008;

Vinkhuyzen *et al.* 2013; Muñoz *et al.* 2014), it also has its pitfalls. If the causal variants are not in linkage disequilibrium with the SNPs used for the estimation of the variance components, their contribution to the variance will not be captured. The proportion of the variance explained by the SNPs is, therefore, likely to be underestimated (Vinkhuyzen *et al.* 2013). This phenomenon has been described as “the case of the missing heritability” (Maher 2008). Our genomic estimates of the additive genetic variance (Tables 2.3-5) were on average 28% lower than pedigree-based estimates that were obtained using the same data accounting only for additive effects. The pedigree-based heritability of number of teats was 0.340 in the Landrace and 0.420 in the Large White population; the pedigree-based heritability of backfat was 0.668 in the Landrace, 0.490 in the Large White and 0.513 in the Pietrain population; and the pedigree-based heritability of lifetime daily gain was 0.401 in the Landrace, 0.300 in the Large White and 0.474 in the Pietrain population (data not shown). Although these differences between the genomic and pedigree estimates are considerable, it is difficult to say if they are more likely due to an overestimation with pedigree, or due to an underestimation with genomics. Nevertheless, using genomic data to estimate additive, dominance, and imprinting variances allows us to not only better understand the genetic architecture of the evaluated traits, but it might also improve the prediction of phenotypes compared to pedigree-based methods.

In recent studies, the inclusion of dominance effects in genomic evaluations of livestock has been reported to increase the accuracy and decrease the bias of estimated breeding values (Toro & Varona 2010; Su *et al.* 2012). In addition, using dominance in genomic evaluations is expected to result in greater cumulative response to selection of purebred animals for crossbred performance than additive models, especially in the presence of overdominance and when retraining is not performed at each generation (Zeng *et al.* 2013). Even when purely additive effects were evaluated, the inclusion of dominance in the genomic evaluations did not decrease the accuracy of prediction (Toro & Varona 2010; Su *et al.* 2012; Zeng *et al.* 2013). In plants, simultaneously accounting for additive and non-additive effects was more stable and yielded higher predictive ability of the mean phenotype than models that only account for additive effects (Muñoz *et al.* 2014). Also in mice, the prediction of phenotypes of complex traits using a model with additive and dominance effects has proven to be feasible and accurate (Lee *et al.* 2008). Therefore, combining additive, dominance, and imprinting under a genomic prediction scope opens new perspectives for the optimization of animal and plant breeding programs aiming for an improved prediction of crossbred performance, and also for identification of individuals that are at a risk for a given disease.

2.4.3 Variance explained per individual chromosome

The strong linear relationship between chromosome length and proportion of variance explained per chromosome in our study was in line with the strong relationship between additive variance explained per chromosome and chromosome length previously described in humans (Yang *et al.* 2011) and in chickens (Abdollahi-Arpanahi *et al.* 2014). Here we also showed that the same applies for dominance and imprinting variance. This indicates that the additive, dominance, and imprinting variance of number of teats, backfat and lifetime daily gain in these populations is explained by many genes located throughout the genome, rather than by a few mutations with large effects.

The relationship between variance explained and chromosome length for number of teats in the Large White population, backfat in the Landrace population, and lifetime daily gain in the Pietrain population using MADI is illustrated in Figure 2.1. Although our results show that the variance of all three genetic effects have a strong relationship with chromosome length, the r^2 for the additive variance was lower than the r^2 for dominance and imprinting variance, especially in the Landrace and Large White populations. In addition, in the Pietrain population, the r^2 for dominance and imprinting variance (Figure 2.1C) was lower than the r^2 observed in the Landrace and Large White populations. This was observed because the proportion of variance explained by chromosome 8 in the Pietrain is clearly lower than in the Landrace and Large White populations. Having a closer look at the data of chromosome 8, we observed that the number of SNPs on this chromosome in the Pietrain population was on average 15% lower than in the Landrace and the Large White populations (n=1,632 in Pietrain, n=1,871 in Landrace, n=1,985 in Large White). Besides chromosome 8, the number of SNPs per chromosome was similar in all three populations. This difference in the number of SNPs on chromosome 8 is observed because in the Pietrain population, compared to the Landrace and the Large White populations, more SNPs presented low minor allele frequency or were completely fixed and, therefore, were excluded from the estimation of the variance components. This large number of SNPs with low minor allele frequency (or completely fixed) could be due to an ascertainment bias due to the selection of SNPs for the SNP chip used in this study. However, this could be also an indication that genes that influence traits included in the selection index of Pietrain are located on this chromosome. The breeding objectives in Pietrain (sire line) are distinct from those in the Landrace and Large White (dam lines) which are more similar. Given this difference, some alleles could have moved to fixation in Pietrain but not in Landrace and Large White.

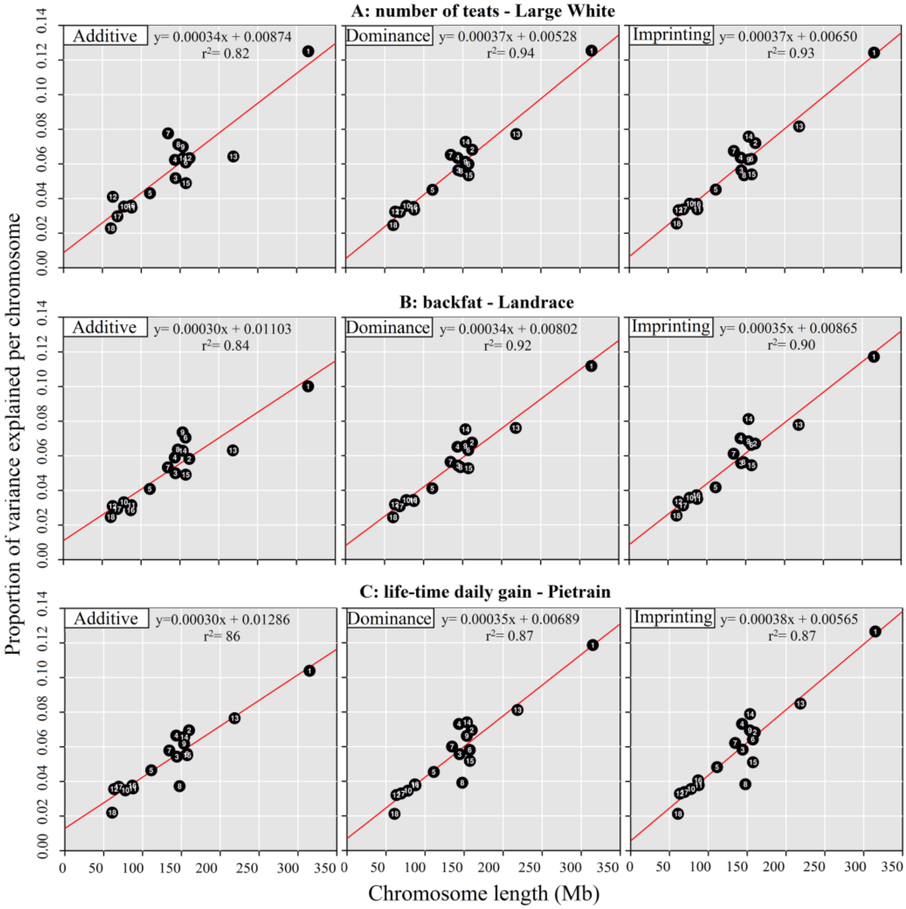

Figure 2.1 Proportion of additive, dominance and imprinting variance explained per individual chromosome against the physical length of the chromosome.

The proportion of additive variance explained by chromosome 7 for number of teats in the Large White population is relatively high (Figure 2.1A). This chromosome explained 21% more additive variance than chromosome 13, which is 62% longer than chromosome 7. This large proportion of explained variance is in agreement with the presence of a QTL for number of teats on this chromosome. In a previous study on a subset of the current Large White population, it was shown that on chromosome 7 a QTL is located in the region of the *VRTN* gene, explaining 2.5% of genetic variance (Duijvesteijn *et al.* 2014). In the current study, we showed that chromosome 7 accounted for 7.75% of the additive variance (5.64% of the total genetic variance using MADI).

2.5 Conclusions

Dominance effects account for a large proportion of the total genetic variance (up to 44%) for number of teats, backfat and lifetime daily gain in the pig populations evaluated. Although the contribution of imprinting effects to the total phenotypic variance of the evaluated traits was relatively small (1-3%), the SNP regression method allowed estimation of the additive, dominance and imprinting effects and resulting variances. Our results indicate a strong relationship between additive variance explained per chromosome and chromosome length, which has been previously described for other traits in other species. In addition, we also show that a similar linear relationship exists for dominance and imprinting variance. These novel results improve our understanding of the genetic architecture of the evaluated traits. The model can now be applied to other traits and species. Our results also open new perspectives for the inclusion of dominance and imprinting effects in the prediction of phenotypes, especially regarding mate allocation techniques in animal and plant breeding, and for assessment of the risk of disease in humans.

2.6 Acknowledgment

Financial support from Breed4Food Consortium is gratefully acknowledged. We also acknowledge input from anonymous reviewers and the associate editor Gustavo de los Campos whose comments improved the manuscript.

2.7 References

Abdollahi-Arpanahi, R., Pakdel, A., Nejati-Javaremi, A., Moradi Shahrbabak, M., Morota, G., Valente, B., Kranis, A., Rosa, G., Gianola, D. (2014) Dissection of additive genetic variability for quantitative traits in chickens using SNP markers. *J. Anim. Breed. Genet.,* **131,** 183-193.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Contr.,* **19,** 716-723.

Bruce, A.B. (1910) The Mendelian theory of heredity and the augmentation of vigor. *Science,* **32,** 627-628.

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

Cheng, Y., Rachagani, S., Cánovas, A., Mayes, M.S., Tait, R.G., Dekkers, J.C., Reecy, J.M. (2013) Body composition and gene expression QTL mapping in mice reveals imprinting and interaction effects. *BMC Genet.,* **14,** 103.

Da, Y., Wang, C., Wang, S., Hu, G. (2014) Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. *PLoS One,* **9,** e87666.

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

De Koning, D.-J., Bovenhuis, H., van Arendonk, J.A. (2002) On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. *Genetics,* **161,** 931-938.

de Los Campos, G., Gianola, D., Allison, D.B. (2010) Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.,* **11,** 880-886.

De Vries, A., Kerr, R., Tier, B., Long, T. (1994) Gametic imprinting effects on rate and composition of pig growth. *Theor. Appl. Genet.,* **88,** 1037-1042.

Duijvesteijn, N., Veltmaat, J.M., Knol, E.F., Harlizius, B. (2014) High-resolution association mapping of number of teats in pigs reveals regions controlling vertebral development. *BMC Genomics,* **15,** 542.

Falconer, D.S., Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics,* 4th edn. Longmans Green, Harlow.

Gilmour, A.R., Gogel, B., Cullis, B., Thompson, R. (2009) ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

Groenen, M.A., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature,* **491,** 393-398.

Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N., van der Werf, J.H. (2011) A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol.,* **43,** 12.

Hill, W.G. (2010) Understanding and using quantitative genetic variation. *Philosophical Transactions of the Royal Society B: Biological Sciences,* **365,** 73-85.

Hill, W.G., Goddard, M.E., Visscher, P.M. (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics,* **4,** e1000008.

Hirooka, H., De Koning, D., Harlizius, B., Van Arendonk, J., Rattink, A., Groenen, M., Brascamp, E., Bovenhuis, H. (2001) A whole-genome scan for quantitative trait loci affecting teat number in pigs. *J. Anim. Sci.,* **79,** 2320-2326.

Jeon, J.-T., Carlborg, Ö., Törnsten, A., Giuffra, E., Amarger, V., Chardon, P., Andersson-Eklund, L., Andersson, K., Hansson, I., Lundström, K. (1999) A paternally expressed QTL affecting skeletal and cardiac muscle mass in pigs maps to the IGF2 locus. *Nat. Genet.,* **21,** 157-158.

Krag, K., Janss, L., Shariati, M., Berg, P., Buitenhuis, A.J. (2013) SNP-based heritability estimation using a Bayesian approach. *Animal,* **7,** 531-539.

Lawson, H.A., Cheverud, J.M., Wolf, J.B. (2013) Genomic imprinting and parent-of-origin effects on complex traits. *Nat. Rev. Genet.,* **14,** 609-617.

Le Rouzic, A., Álvarez-Castro, J.M., Carlborg, Ö. (2008) Dissection of the genetic architecture of body weight in chicken reveals the impact of epistasis on domestication traits. *Genetics,* **179,** 1591-1599.

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

Maher, B. (2008) The case of the missing heritability. *Nature,* **456,** 18-21.

Melchinger, A.E., Piepho, H.-P., Utz, H.F., Muminović, J., Wegenast, T., Törjék, O., Altmann, T., Kusterer, B. (2007) Genetic basis of heterosis for growth-related traits in

Arabidopsis investigated by testcross progenies of near-isogenic lines reveals a significant role of epistasis. *Genetics,* **177,** 1827-1837.

Muñoz, P.R., Resende, M.F., Gezan, S.A., Resende, M.D.V., de los Campos, G., Kirst, M., Huber, D., Peter, G.F. (2014) Unraveling Additive from Non-Additive Effects Using Genomic Relationship Matrices. *Genetics***,** genetics. 114.171322.

Nezer, C., Moreau, L., Brouwers, B., Coppieters, W., Detilleux, J., Hanset, R., Karim, L., Kvasz, A., Leroy, P., Georges, M. (1999) An imprinted QTL with major effect on muscle mass and fat deposition maps to the IGF2 locus in pigs. *Nat. Genet.,* **21,** 155-156.

Nishio, M., Satoh, M. (2014) Including dominance effects in the genomic BLUP method for genomic evaluation. *PLoS One,* **9,** e85792.

Pante, M.J.R., Gjerde, B., McMillan, I., Misztal, I. (2002) Estimation of additive and dominance genetic variances for body weight at harvest in rainbow trout, *Oncorhynchus mykiss*. *Aquaculture,* **204,** 383-392.

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

Shen, G., Zhan, W., Chen, H., Xing, Y. (2014) Dominance and epistasis are the main contributors to heterosis for plant height in rice. *Plant Sci.,* **215,** 11-18.

Shull, G.H. (1908) The composition of a field of maize. *J. Hered.***,** 296-301.

Sorensen, D., Fernando, R., Gianola, D. (2001) Inferring the trajectory of genetic variance in the course of artificial selection. *Genet. Res.,* **77,** 83-94.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **64,** 583-639.

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

Sun, C., VanRaden, P.M., Cole, J.B., O'Connell, J.R. (2014) Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. *PLoS One,* **9,** e103934.

Tier, B., Meyer, K. (2004) Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. *J. Anim. Breed. Genet.,* **121,** 77-89.

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

Vinkhuyzen, A.A., Wray, N.R., Yang, J., Goddard, M.E., Visscher, P.M. (2013) Estimation and partitioning of heritability in human populations using whole genome analysis methods. *Annu. Rev. Genet.,* **47,** 75.

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

Visscher, P.M., Goddard, M.E. (2014) A general unified framework to assess the sampling variance of heritability estimates Uuing pedigree or marker-based relationships. *Genetics***,** genetics. 114.

Vitezica, Z.G., Varona, L., Legarra, A. (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics,* **195,** 1223-1230.

Wolf, J., Cheverud, J.M. (2012) Detecting maternal-effect loci by statistical cross-fostering. *Genetics,* **191,** 261-277.

Wray, N.R., Goddard, M.E., Visscher, P.M. (2007) Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.,* **17,** 1520-1528.

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

Yang, J., Manolio, T.A., Pasquale, L.R., Boerwinkle, E., Caporaso, N., Cunningham, J.M., de Andrade, M., Feenstra, B., Feingold, E., Hayes, M.G. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.,* **43,** 519-525.

Zeng, J., Toosi, A., Fernando, R.L., Dekkers, J.C., Garrick, D.J. (2013) Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genet. Sel. Evol.,* **45**.

2.8 Supporting information

Table 2.S1 Sampling correlation between variance estimates of the simulated trait

Model		$\sigma_{\mathbf{e}}^2$	$\sigma_{\mathbf{Aa}}^2$	$\sigma_{\mathbf{Dd}}^2$
MA	$\sigma_{\mathbf{Aa}}^2$	-0.504		
MAD	$\sigma_{\mathbf{Aa}}^2$	-0.186		
	$\sigma_{\mathbf{Dd}}^2$	-0.586	-0.208	
MADI	$\sigma_{\mathbf{Aa}}^2$	-0.204		
	$\sigma_{\mathbf{Dd}}^2$	-0.505	-0.215	
	$\sigma_{\mathbf{Ii}}^2$	-0.422	-0.006	0.018

File 2.S1 http://www.g3journal.org/content/suppl/2015/10/04/g3.115.019513.DC1/FileS1.docx

File 2.S2 http://www.g3journal.org/content/suppl/2015/10/04/g3.115.019513.DC1/FileS2.xlsx

3

Genomic prediction of growth in pigs based on a model including additive and dominance effects

Marcos S Lopes[1,2], John WM Bastiaansen[2], Luc Janss[3], Egbert F Knol[1], Henk Bovenhuis[2]

[1] Topigs Norsvin Research Center, 6640 AA, Beuningen, the Netherlands; [2] Wageningen University, Animal Breeding and Genomics Centre, 6700 AH, Wageningen, the Netherlands; [3] Aarhus University, Centre for Quantitative Genetics and Genomics, DK-8830, Tjele, Denmark

Journal of Animal Breeding and Genetics (2015) *Early online*

Abstract

Independent of whether prediction is based on pedigree or genomic information, the focus of animal breeders has been on additive genetic effects or “breeding values”. However, when predicting phenotypes rather than breeding values of an animal, models that account for both additive and dominance effects might be more accurate. Our aim with this study was to compare the accuracy of predicting phenotypes using a model that accounts for only additive effects (MA) and a model that accounts for both additive and dominance effects simultaneously (MAD). Lifetime daily gain (DG) was evaluated in three pig populations (1,424 Pietrain, 2,023 Landrace, and 2,157 Large White). Animals were genotyped using the Illumina SNP60K Beadchip and assigned to either a training dataset to estimate the genetic parameters and SNP effects or to a validation dataset to assess the prediction accuracy. Models MA and MAD applied random regression on SNP genotypes and were implemented in the program Bayz. The additive heritability of DG across the three populations and the two models was very similar at about 0.26. The proportion of phenotypic variance explained by dominance effects ranged from 0.04 (Large White) to 0.11 (Pietrain), indicating that importance of dominance might be breed-specific. Prediction accuracies were higher when predicting phenotypes using total genetic values (sum of breeding values and dominance deviations) from the MAD model compared to using breeding values from both MA and MAD models. The highest increase in accuracy (from 0.195 to 0.222) was observed in the Pietrain, and the lowest in Large White (from 0.354 to 0.359). Predicting phenotypes using total genetic values instead of breeding values in purebred data improved prediction accuracy and reduced the bias of genomic predictions. Additional benefit of the method is expected when applied to predict crossbred phenotypes, where dominance levels are expected to be higher.

Key words: SNP, variance component, phenotype prediction, pigs

3.1 Introduction

In the last decades, pig breeding programs have achieved a remarkable genetic improvement of production as well as reproduction traits (Merks *et al.* 2012). Most of this success can be attributed to the application of best linear unbiased prediction (BLUP), which uses family relationships (pedigree information) to predict breeding values via mixed model equations (Henderson 1975). With the incorporation of genomic data in genetic evaluations, e.g. genomic selection (Meuwissen *et al.* 2001), genetic progress is expected to further increase. While prediction methods have progressed from being based on pedigree information to genomic information, or a combination of both in a single-step approach (Misztal *et al.* 2009; Christensen & Lund 2010), the focus has remained on predicting additive effects (breeding values). However, if the aim is to predict the phenotype rather than the breeding value of an animal, models that account for dominance effects, in addition to additive effects (e.g. total genetic effects), might be more effective (Su *et al.* 2012).

Dominance effects are not directly transmitted to offspring as they are the result of interaction between alleles at the same locus. Until recently, estimation of dominance effects in livestock species has been limited because it is difficult to obtain accurate dominance estimates based on pedigree relationships, unless the appropriate family structure (i.e. large full-sib and half-sib families) is available (Vitezica *et al.* 2013). With the availability of genomic information, possibilities to estimate dominance variance have increased. However, although markers with dominance effects have been identified (Boysen *et al.* 2013; Lopes *et al.* 2014) and the potential of incorporating dominance effects in genetic evaluation has been described (Su *et al.* 2012; Ertl *et al.* 2014; Nishio & Satoh 2014), the use of dominance models in genomic prediction is still limited. In addition, the studies that to date have evaluated dominance effects in pig data have performed single-breed analyses (Su *et al.* 2012; Nishio & Satoh 2014). Therefore, the evaluation of dominance effects on the same phenotype in different breeds is still lacking.

The aim of this study was to compare the accuracy of predicting lifetime daily gain (DG) from three purebred pig populations using a model that accounts for only additive effects and a model that accounts for both additive and dominance effects simultaneously.

3.2 Material and methods

3.2.1 Animals and phenotypes

The phenotype evaluated was DG, which was defined as the average daily weight increase from birth to ~120 kg. This phenotype was measured on animals from three purebreed-based pig populations: Pietrain, Landrace and Large White. The data of each population was divided into three groups: a) ALL: consisted of all genotyped animals and their contemporaries (e.g. animals from the same breed and herd-year-week; 41,208 Pietrain, 49,074 Landrace, and 52,295 Large White animals). This group was used to calculate pre-adjusted DG values (corrected for systematic environmental factors) that were used as the response variables in subsequent analyses. The pre-adjusted phenotype was corrected for all non-genetic effects, except for the effect of litter. The litter effect was not corrected for because this effect could be confounded with dominance effects (Su *et al.* 2012). The non-genetic effects used in the correction of the phenotypes were estimated using a pedigree-based linear model in ASReml v3.0 (Gilmour *et al.* 2009). This model included sex and herd-year-week as fixed class effects, birth weight as a covariable, and pen, additive genetic, litter and residual as random effects. b) TRAINING: consisted of the oldest 80% animals with both genotype and phenotype records (1,138 Pietrain, 1,617 Landrace, and 1,725 Large White animals). This group was used for the estimation of genetic parameters and SNP effects. c) VALIDATION: consisted of the youngest 20% animals with both genotype and phenotype records (286 Pietrain, 406 Landrace, and 432 Large White animals). For this group, breeding values and total genetic values (sum of breeding values and dominance deviations) were estimated based on the SNP effects estimated in TRAINING. Descriptive statistics of the data are in Table 3.1.

3.2.2 Genotypes

Animals in TRAINING and VALIDATION of all populations were genotyped using the Illumina Porcine SNP60 Beadchip. SNPs with call rate <0.95, minor allele frequency <0.05, strong deviation from Hardy-Weinberg equilibrium (χ^2>600), GenCall<0.15, unmapped SNPs and SNPs located on sex chromosomes, according to the Sscrofa10.2 assembly of the reference genome (Groenen *et al.* 2012), were excluded from the data set. Out of the initial 64,232 SNPs, 38,116 SNPs for Pietrain, 39,131 SNPs for Landrace, and 37,574 SNPs for Large White were kept for further analyses. All genotyped animals had a frequency of missing genotypes <0.05. After cleaning procedures, the remaining missing genotypes were imputed using Beagle v3.3.2 (Browning & Browning 2007).

Table 3.1 Number of animals (N), mean and standard deviation (SD).

Line	Dataset	N	Mean	SD
Pietrain	ALL	41,208	603	75
	TRAINING	1,138	627	66
	VALIDATION	286	652	69
Landrace	ALL	49,074	602	71
	TRAINING	1,617	624	64
	VALIDATION	406	639	62
Large White	ALL	52,295	635	72
	TRAINING	1,725	645	69
	VALIDATION	432	662	64

ALL: the whole population used in the pre-adjustment of the phenotypes, which includes the animals from TRAINING, VALIDATION and their contemporaries; TRAINING: genotyped animals used for estimation of SNP effects and genetic parameters. VALIDATION: subset used for prediction accuracy.

3.2.3 Variance components and SNP effects

Variance components and SNP effects were estimated using models with random regression on SNP genotypes. Within each breed two different models were evaluated:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Lb} + \mathbf{Aa} + \mathbf{e} \qquad \text{(MA model)}$$

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Lb} + \mathbf{Aa} + \mathbf{Dd} + \mathbf{e} \qquad \text{(MAD model)}$$

where **y** is a vector of pre-adjusted phenotypic observations; μ is the mean of the population and **1** a vector of ones; **L** is the design matrix for the common litter effects; **b** is an unknown vector of common litter effects; **A** and **D** are design matrices for additive and dominance effects, respectively; **a** and **d** are unknown vectors of allele substitution effects and dominance deviations, respectively; and **e** is a vector of residuals. The entries of the design matrices **A** and **D** are SNP genotypes recoded following the classical parameterization described by Vitezica *et al.* (2013). Considering SNP *j* with alleles G and C, the recoding of genotypes for the i^{th} animal was carried out as follows:

$$\mathbf{A}_{i,j} = \begin{cases} 0 - 2p_j \\ 1 - 2p_j \\ 2 - 2p_j \end{cases} \text{for genotypes} \begin{cases} \text{GG} \\ \text{GC} \\ \text{CC} \end{cases}$$

$$\mathbf{D}_{i,j} = \begin{cases} -2p_j^2 \\ 2p_jq_j \\ -2q_j^2 \end{cases} \text{for genotypes} \begin{cases} \text{GG} \\ \text{GC} \\ \text{CC} \end{cases}$$

where p_j is the within-breed frequency of the C allele, and q_j is the within-breed frequency of the G allele.

We assumed distributions of **a** ~ *N*(0,**I**σ_a^2), **d** ~ *N*(0,**I**σ_d^2), **b** ~ *N*(0,**I**σ_{L}^2) and **e** ~ *N*(0,**I**σ_e^2), with σ_a^2 and σ_d^2 being respectively the additive and dominance genetic variances of a single SNP, σ_L^2 the common litter variance, and σ_e^2 the residual variance. The models MA and MAD were fitted using a Bayesian approach in the Bayz software package (http://www.bayz.biz/). The prior distributions for unknown variances were set as unbounded uniform. The generated Monte Carlo chain started with all regression parameters and other location parameters at zero, and all variance parameters at 1, and blocked Gibbs samplers were employed to facilitate mixing. The models evaluated in this study were SNP-based models and therefore they do not readily provide estimates of total explained variance. The total explained variance (within the population being analyzed) from a model term like **Aa** can be obtained by writing *var*(**Aa**)=**AA**'σ_a^2, and computing or evaluating the expected average diagonal of **AA**' to provide the constant to scale the explained variances of single SNPs to the total explained variance. In this way, the total phenotypic variance (σ_{P}^2) in the models can be expressed as σ_{P}^2= ($\sigma_{\mathbf{Aa}}^2$ + σ_L^2 + σ_e^2) for MA and σ_{P}^2= ($\sigma_{\mathbf{Aa}}^2$ + $\sigma_{\mathbf{Dd}}^2$ + σ_L^2 + σ_e^2) for MAD, with $\sigma_{\mathbf{Aa}}^2$=**AA'**σ_a^2, (total additive variance), and $\sigma_{\mathbf{Dd}}^2$=**DD'**σ_d^2 (total dominance variance). Alternatively, the variance contributed by a random effect could be estimated by evaluating the sample variance of the entries of the vectors **Aa** and **Dd** at each iteration of the Gibbs sampler (Sorensen *et al.* 2001). The additive heritability (narrow-sense heritability) was defined as the total additive variance divided by the total phenotypic variance ($\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$). The proportion of phenotypic variance explained by dominance effects was defined as the total dominance variance divided by the total phenotypic variance ($\sigma_{\mathbf{Dd}}^2/\sigma_{\mathrm{P}}^2$). Each model was run as a single chain with a length of 350,000. The first 50,000 iterations of each run were regarded as burn-in. After the burn-in, one sample was saved per 100 iterations.

3.2.4 Model comparison

Model comparison was performed using the Deviance Information Criterion (DIC, Spiegelhalter *et al.* 2002), which is analogous to the Akaike Information Criterion (Akaike 1974). DIC combines measurements of model fit (the expected deviance) and model complexity (the effective number of parameters) over all iterations, excluding burn-in. The model that provides the best fit to the data is the one with the lowest DIC (Spiegelhalter *et al.* 2002).

3.2.5 Validation

Accuracy of predicting phenotypes was assessed in the VALIDATION data as the correlation between the pre-adjusted phenotype and estimated breeding values ($\hat{\boldsymbol{u}}$= $\mathbf{A}\hat{\boldsymbol{a}}$**)** from models MA and MAD, and the total genetic values ($\hat{\boldsymbol{g}}$= $\mathbf{A}\hat{\boldsymbol{a}}$ + $\mathbf{D}\hat{\boldsymbol{d}}$) from model MAD. The allele substitution effects and dominance deviations ($\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{d}}$, respectively) were estimated using the data from the TRAINING group. Further, potential bias of predicted phenotypes was assessed by regressing the pre-adjusted phenotypes on the genomic predictions ($\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{g}}$). Bias of prediction was characterized when the regression coefficient of the pre-adjusted phenotypes on genomic predictions deviated from 1. Finally, we applied the Hotelling-Williams t-test (Steiger 1980) to verify if the accuracies of the different genomic predictions were significantly (P<0.10) different.

3.3 Results

In the Pietrain population, a substantial fraction of the phenotypic variation in DG was attributed to dominance variance: the additive heritability was 0.26 using both MA and MAD model, and the proportion of phenotypic variance explained by dominance effects was 0.11 (Table 3.2). In the Landrace population, the additive heritability was slightly higher than in the Pietrain population, being 0.28 using the MA model and 0.27 using the MAD model, but the proportion of phenotypic variance explained by dominance effects was quite lower at 0.06. In the Large White population, the additive heritability for in DG was the same as in the Pietrain population (0.26 using both MA and MAD model) and the proportion of phenotypic variance explained by dominance effects was the lowest of the three evaluated populations at 0.04. Although the dominance heritabilities for DG varied across the different populations, the MAD model presented lower DIC than the MA model in all populations (Table 3.2). In the Pietrain population, the accuracies of predicting phenotypes using the breeding values from the MA model were slightly higher than using the breeding values from the MAD model with values of 0.195 and 0.190, respectively (Table 3.3). The accuracy of predicting phenotypes using the total genetic values from the MAD model was 0.222 in this population. In the Landrace population, the accuracies of predicting phenotypes using the breeding values were 0.277 for both MA and MAD model, and the accuracy of predicting phenotypes using the total genetic values was 0.284. In the Large White population, the accuracies of predicting phenotypes using the breeding values were 0.354 for both MA and MAD model, and the accuracy of predicting phenotypes using the total genetic values was 0.359.

Breeding values from the MA model were more biased (deviated more from 1) than the total genetic values from the MAD model in the Pietrain population (0.77 and 0.84, respectively, Table 3.3), and in the Landrace population (0.87 and 0.90, respectively). In the Large White population, the bias of the breeding values from the MA model and of the total genetic values from the MAD model was the same (1.04).

3.4 Discussion

In this study, we found that the additive heritability of DG was very similar (~0.26) across three purebred pig populations. The proportion of phenotypic variance explained by dominance effects, however, varied considerably across populations (Table 3.2). The proportion of the total genetic variance explained by dominance effects ranged from 13% (Large White) to 30% (Pietrain). A large contribution of dominance effects to the total genetic variance in DG has been reported in a previous study, performed on a smaller number of records from the same populations (Lopes *et al.* 2015). In our previous study, however, a genotypic model instead of a breeding model (as used in the current study) was applied, and the influence of dominance effects on the prediction accuracy of phenotypes was not evaluated.

In the current study, we show that the use of total genetic values (breeding values and dominance deviations) from the MAD model instead of breeding values from MA (or MAD) results in a higher accuracy of predicting phenotypes in all populations (Table 3.3). The highest increase in accuracy (from 0.195 to 0.222) was observed in the Pietrain population, where the largest amount of dominance was detected. In the Large White population, where the lowest amount of dominance variance was detected, this increase in accuracy was only marginal (from 0.354 to 0.359).

Table 3.2 Variance estimates (± standard deviation), proportion of phenotypic variance (σ_{P}^2) explained by additive and dominance effects, and goodness of fit (DIC) for MA and MAD models.

Population	**Model**	**Variance estimates**				**Variance explained**		**DIC**
		$\sigma_{\mathbf{Aa}}^2$	$\sigma_{\mathbf{Dd}}^2$	σ_{L}^2	σ_e^2	$\sigma_{\mathbf{Aa}}^2/\sigma_{\mathrm{P}}^2$	$\sigma_{\mathbf{Dd}}^2/\sigma_{\mathrm{P}}^2$	
Pietrain	MA	758 ± 121		273 ± 131	1,868 ± 152	0.26 ± 0.04		10,016
	MAD	766 ± 121	325 ± 141	225 ± 121	1,602 ± 177	0.26 ± 0.04	0.11 ± 0.05	**9,949**
Landrace	MA	668 ± 88		155 ± 73	1,517 ± 92	0.28 ± 0.03		14,699
	MAD	639 ± 88	153 ± 94	127 ± 70	1,421 ± 103	0.27 ± 0.03	0.06 ± 0.04	**14,662**
Large White	MA	626 ± 81		229 ± 85	1,524 ± 99	0.26 ± 0.03		14,829
	MAD	628 ± 81	87 ± 55	204 ± 88	1,467 ± 107	0.26 ± 0.03	0.04 ± 0.02	**14,809**

$\sigma_{\mathbf{Aa}}^2$: total additive variance; $\sigma_{\mathbf{Dd}}^2$: total dominance variance; σ_e^2: residual variance; σ_{L}^2: common litter variance; DIC: Deviance Information Content. The lowest DIC per population is given in bold, indicating the model with the best fit.

Table 3.3 Accuracies and bias of predicted breeding values ($\hat{u}$) and total genetic values ($\hat{g}$) for MA and MAD models.

Population	Accuracy [*]			Bias [**]		
	MA ($\hat{u}$)	**MAD ($\hat{u}$)**	**MAD ($\hat{g}$)**	**MA ($\hat{u}$)**	**MAD ($\hat{u}$)**	**MAD ($\hat{g}$)**
Pietrain	0.195 ^{a}	0.190 ^{a}	0.222 ^{b}	0.77	0.76	0.84
Landrace	0.277 ^{a}	0.277 ^{a}	0.284 ^{a}	0.87	0.90	0.90
Large White	0.354 ^{a}	0.354 ^{a}	0.359 ^{a}	1.04	1.03	1.04

[*] Accuracy: was defined as the correlation between the pre-adjusted phenotypes and the genomic predictions ($\hat{u}$ and $\hat{g}$); [**] Bias: regression coefficient obtained by regressing the pre-adjusted phenotypes on the genomic predictions ($\hat{u}$ and $\hat{g}$). Within-row accuracies that do not share the same superscript ($^{a\text{-}b}$) differ significantly (P<0.10) according to the Hotelling-Williams t-test.

The influence of dominance effects on the later growth phase of pigs (from 30 kg to 100 kg) has been evaluated in a Duroc population (Su *et al.* 2012). The total genetic values (MAD model) resulted in higher accuracies of predicting phenotypes than using the breeding values from an additive model (0.330 and 0.319, respectively). In addition, Su *et al.* (2012) showed that accounting for dominance effects in the model reduced the bias of genomic predictions and improved the goodness of fit. In our study, predictions were biased using both MA and MAD models. One explanation for such bias might be the fact that the evaluated populations consist of breeding animals that are selected for DG. According to Vitezica *et al.* (2011), genomic prediction is expected to be biased when the genotyped population is highly selected for the evaluated trait. However, we also observed that in the Pietrain and Landrace populations, the total genetic values from the MAD model were less biased than the breeding values from the MA model (Table 3.3). In the Large White population, i.e. the population with the smallest proportion of dominance, the bias was about the same using both models (1.04). Based on the DIC (Table 3.2), the MAD model presented the best fit to the phenotypic observations in all populations, regardless of whether a large or small amount of dominance variance was found.

Large differences in the proportion of dominance were observed for the same trait across different populations. This has also been observed for milk performance traits in cattle. Ertl *et al.* (2014) reported that dominance deviations accounted for 28-41% of the total genetic variance of milk, fat, and protein yield in a Fleckvieh population, while Wittenburg *et al.* (2015) found no dominance variation for the same traits in a Holstein population. Therefore, our results and the results of the studies in cattle suggest that dominance effects differ considerably between populations (it might be breed-specific).

Breed-specific dominance effects can result from differences in allele frequencies in each population. The variance explained by a completely dominant locus would be all captured as additive variance if the recessive allele is segregating at high frequency (Falconer & Mackay 1996). Thus, in order to be able to discover dominance variance, the locus with this mode of gene action needs to be segregating at intermediate gene frequency (Hill *et al.* 2008). If differences in allele frequencies are responsible for the differences in dominance variation across the populations, then we would expect the Pietrain population to have the highest levels of heterozygosity, and the Large White population the lowest. However, the $\sum(2pq)^2$ (which in fact contributes to the dominance variation) was similar in the Pietrain and the Large White populations (5,864 and 5,813, respectively, Table 3.S1). The highest value of $\sum(2pq)^2$ was observed in the Landrace population (6,106), which presented intermediate values of total dominance variance. Because the patterns of heterozygosity do not seem to explain the differences in dominance variance across populations, we could argue that these differences could be due to differences in linkage disequilibrium (LD) between the markers and additive and dominance QTLs across populations.

The magnitude of LD in the population has recently been related to the amount of dominance variance (Hill & Mäki-Tanila 2015). Increased LD in a population can result in a small increase of dominance and epistasis variance in outbred populations. Analysis of LD patterns in the studied populations shows that the genomes of the Pietrain, Landrace, and Large White populations were composed of 2,640, 2,705, and 2,941 haplotype blocks, respectively (Veroneze *et al.* 2013). The average size of these blocks was 447, 387, and 378 kb in the Pietrain, Landrace, and Large White populations, respectively. The magnitude of LD does appear to be associated with the magnitude of dominance variance for DG in these populations. However, in a previous study (Lopes *et al.* 2015) we estimated additive, dominance, and imprinting variance of the trait backfat and observed that the largest dominance variation was found in the Large White population and the smallest in the Pietrain population. This is the opposite of the results for DG in this study. Therefore, if the magnitude of LD indeed influences the amount of dominance variation, this seems to be trait-related and not a general rule.

Dominance deviations accounted for up to 30% of the total genetic variance in DG in the evaluated populations. While this is a considerable proportion of the genetic variance, still, the majority of the genetic variation for this trait in these purebred populations is explained by additive effects. Hill *et al.* (2008) suggested that breeding companies should focus on additive effects because they often account for 100% (or at least over 50%) of the total genetic variation. However, breeding companies might also be interested in predicting the phenotype of a particular mating, especially in crossbreeding (Toro & Varona 2010; Ertl *et al.* 2014). In those situations, dominance might be of interest. Dominance effects could potentially be used for mate allocation without compromising additive genetic progress (Ertl *et al*. 2014). In a simulation study, Zeng *et al.* (2013) showed that especially in the presence of overdominance, models that account for dominance effects resulted in greater cumulative response to selection of purebred animals for crossbred performance compared to additive models. Although the inclusion of dominance effects in genetic evaluations may not be useful for all traits in all populations, in this study, the accuracy of predicting phenotypes was higher using the total genetic values from the MAD model than using the breeding values from both models. Therefore, models that account for dominance effects can contribute to a better prediction of phenotypes and should be considered when predicting individual phenotypes.

Large standard errors of dominance variance estimates were observed in all populations (Table 3.2), which might be related to the power to estimate these effects. Even though loci with large non-additive effects for complex traits may exist, the power to detect them in outbred populations is low, unless these loci have large effects and are segregating at an intermediate frequency (Hill *et al.* 2008). Others have also indicated that the size of current datasets (often < 2,000) may be too small to accurately estimate dominance effects (Su *et al.* 2012 and Ertl *et al.* 2014). Therefore, we expect that using larger datasets, dominance effects would be more accurately estimated and the standard errors of dominance variance estimates would decrease.

Additive heritability was very similar in all three populations, but the accuracy of prediction of the breeding values from the MA model was very different across the three populations. For instance, the prediction accuracy of the breeding values (MA model) in the Large White population was 0.354, while in the Pietrain population it was 0.195. This difference in accuracy is in accordance with the difference in the degree of relationship between the TRAINING and VALIDATION groups. In simulated data (Habier *et al.* 2007) as well as real data (Wu *et al.* 2015), it has been shown that higher relationship between training and validation sets results in higher prediction accuracies. In the Pietrain population, 26% of the animals in VALIDATION had both parents present in the TRAINING, while in the Landrace and the Large White populations this was 56% and 55%, respectively. Thus, this low percentage of animals in VALIDATION with close relationships to TRAINING could explain the lower accuracy of the breeding values of the Pietrain population compared to the other two populations. Another factor that influences the accuracies of genomic breeding values is the size of the training population (Daetwyler *et al.* 2010). The smallest size of TRAINING was observed in the Pietrain population (n= 1,138) and the largest in the Large White population (n= 1,725).

3.5 Conclusions

In this study, the accuracy of predicting phenotypes of purebred animals was improved by including dominance deviations in the prediction model. Predicting crossbred phenotypes with the total genetic values from the MAD model is expected to be even more beneficial as total dominance variance is expected to be higher in crossbreds. However, in both cases (purebred and crossbred populations) larger datasets than analyzed to date might be needed to more accurately estimate the importance of dominance effects in complex traits.

3.6 Acknowledgment

Financial support from Breed4Food Consortium and valuable comments from Esinam Nancy Amuzu-Aweh are gratefully acknowledged.

3.7 References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Contr.,* **19,** 716-723.

Boysen, T.-J., Heuer, C., Tetens, J., Reinhardt, F., Thaller, G. (2013) Novel use of derived genotype probabilities to discover significant dominance effects for milk production traits in dairy cattle. *Genetics,* **193,** 431-442.

Browning, S.R., Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.,* **81,** 1084-1097.

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

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

Ertl, J., Legarra, A., Vitezica, Z.G., Varona, L., Edel, C., Emmerling, R., Götz, K.-U. (2014) Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. *Genet. Sel. Evol.,* **46,** 40.

Falconer, D.S., Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics,* 4th edn. Longmans Green, Harlow.

Gilmour, A.R., Gogel, B., Cullis, B., Thompson, R. (2009) ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

Groenen, M.A., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature,* **491,** 393-398.

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

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

Hill, W., Mäki-Tanila, A. (2015) Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. *J. Anim. Breed. Genet.,* **132,** 176-186.

Hill, W.G., Goddard, M.E., Visscher, P.M. (2008) Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics,* **4,** e1000008.

Lopes, M.S., Bastiaansen, J.W., Harlizius, B., Knol, E.F., Bovenhuis, H. (2014) A genome-wide association study reveals dominance effects on number of teats in pigs. *PLoS One,* **9,** e105867.

Lopes, M.S., Bastiaansen, J.W.M., Janss, L., Knol, E.F., Bovenhuis, H. (2015) Estimation of additive, dominance, and imprinting genetic variance using genomic data. *G3: Genes|Genomes|Genetics*.

Merks, J., Mathur, P., Knol, E. (2012) New phenotypes for new breeding goals in pigs. *Animal,* **6,** 535-543.

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

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

Nishio, M., Satoh, M. (2014) Including dominance effects in the genomic BLUP method for genomic evaluation. *PLoS One,* **9,** e85792.

Sorensen, D., Fernando, R., Gianola, D. (2001) Inferring the trajectory of genetic variance in the course of artificial selection. *Genet. Res.,* **77,** 83-94.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A. (2002) Bayesian measures of model complexity and fit. *J. R. Statist. Soc. B,* **64,** 583-639.

Steiger, J.H. (1980) Tests for comparing elements of a correlation matrix. *Psychol. Bull.,* **87,** 245.

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

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

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

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

Vitezica, Z.G., Varona, L., Legarra, A. (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics,* **195,** 1223-1230.

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

Wu, X., Lund, M., Sun, D., Zhang, Q., Su, G. (2015) Impact of relationships between test and training animals and among training animals on reliability of genomic prediction. *J. Anim. Breed. Genet.,* **132,** 366–375.

Zeng, J., Toosi, A., Fernando, R.L., Dekkers, J.C., Garrick, D.J. (2013) Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genet. Sel. Evol.,* **45**.

3.8 Supporting information

Table 3.S1 Heterozygosity patterns of the evaluated populations.

Population	N	$(\sum 2pq)$/N	$\sum 2pq$	$\sum(2pq)^2$
All markers				
Pietrain	38,116	0.37	14,255	5,864
Landrace	39,131	0.38	14,759	6,106
Large White	37,574	0.38	14,111	5,813
Common set of markers				
Pietrain	30,335	0.38	11,646	4,864
Landrace	30,335	0.39	11,910	5,038
Large White	30,335	0.39	11,720	4,908

N= number of SNPs

4

A genome-wide association study reveals dominance effects on number of teats in pigs

Marcos S Lopes[1,2], John WM Bastiaansen[2], Barbara Harlizius[1], Egbert F Knol[1], Henk Bovenhuis[2]

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

PLoS ONE (2014) 9(8):e105867

Abstract

Dominance has been suggested as one of the genetic mechanisms explaining heterosis. However, using traditional quantitative genetic methods it is difficult to obtain accurate estimates of dominance effects. With the availability of dense SNP (Single Nucleotide Polymorphism) panels, we now have new opportunities for the detection and use of dominance at individual loci. Thus, the aim of this study was to detect additive and dominance effects on number of teats (NT), specifically to investigate the importance of dominance in a Landrace-based population of pigs. In total, 1,550 animals, genotyped for 32,911 SNPs, were used in single SNP analysis. SNPs with a significant genetic effect were tested for their mode of gene action being additive, dominant or a combination. In total, 21 SNPs were associated with NT, located in three regions with additive (SSC6, 7 and 12) and one region with dominant effects (SSC4). Estimates of additive effects ranged from 0.24 to 0.29 teats. The dominance effect of the QTL located on SSC4 was negative (-0.26 teats). The additive variance of the four QTLs together explained 7.37% of the total phenotypic variance. The dominance variance of the four QTLs together explained 1.82% of the total phenotypic variance, which corresponds to one-fourth of the variance explained by additive effects. The results suggest that dominance effects play a relevant role in the genetic architecture of NT. The QTL region on SSC7 contains the most promising candidate gene: *VRTN*. This gene has been suggested to be related to the number of vertebrae, a trait correlated with NT.

Key words: GWAS, heterosis, *VRTN*

4.1 Introduction

Dominance effects are non-additive effects due to the interaction between alleles at the same locus. In livestock and plant breeding, the main benefits of dominance effects are expected in crossbreeding, since dominance has been suggested as one of the genetic mechanisms explaining heterosis (Davenport 1908; Shull 1908; Bruce 1910; Xiao *et al.* 1995; Visscher *et al.* 2000). However, estimates of dominance effects have not been widely used in livestock breeding because it is difficult to estimate these effects accurately based on pedigree (Vitezica *et al.* 2013).

The development of dense SNP (Single Nucleotide Polymorphism) panels offered new opportunities for detection and use of dominance at individual loci. However, genomic selection or genome-wide association studies (GWAS) mainly focused on additive genetic effects and ignored dominance. Recently, a number of studies investigated the importance of non-additive effects in genomic prediction (Toro & Varona 2010; Su *et al.* 2012; Wellmann & Bennewitz 2012; Vitezica *et al.* 2013) and GWAS (Coster *et al.* 2012; Boysen *et al.* 2013), showing that accounting for these effects increased the accuracy and reduced the bias of genomically-predicted breeding values in comparison to an additive model (Toro & Varona 2010; Su *et al.* 2012; Wellmann & Bennewitz 2012; Vitezica *et al.* 2013). Su *et al.* (2012) showed that in a purebred Duroc population the dominance variance accounted for 6% of the total phenotypic variance in daily gain, emphasizing the relevance of dominance.

Significant dominance effects on number of piglets born alive and litter size were identified in a GWAS (Coster *et al.* 2012). In cattle, significant dominance effects were reported for milk production traits (Boysen *et al.* 2013). In both studies, additive and dominance effects were tested for each SNP using multiple regression, i.e. this approach simultaneously tested for the significance of the SNP and investigated its mode of gene action. An alternative way of testing for additive and dominance effects of an SNP consists of two steps: 1) SNP genotypes are fitted in the model as a class variable and the significance of a genetic association is tested, irrespective of the mode of gene action and subsequently, 2) only the SNPs with a significant genetic effect are tested for their mode of gene action. This two-step approach is favored over the multiple regression model because a single class variable is used to capture the total genetic variation that is explained by the SNP, while the multiple regression method applied by Coster *et al.* (2012) and Boysen *et al.* (2013) will divide the variation over two parameters which are then separately tested for significance. In addition, the multiple regression model requires approximately twice the number of tests that are performed by the two-step approach. Therefore, for certain modes of gene action, this multiple regression model leads to a reduction of power. Fitting an SNP as a class variable has been successfully applied in previous GWAS (Bouwman *et al.* 2011; Bouwman *et al.* 2012; Wijga *et al.* 2012). However, in these studies, the mode of gene action of the significant SNPs was not evaluated.

In QTL mapping studies in pigs, number of teats (NT) has been one of the most extensively studied traits. NT is an important trait for breeding programs because the number of piglets in a litter is often larger than the number of functional teats of the sow due to the remarkable improvement in sow prolificacy over the last decades (Rodriguez *et al.* 2005). A lower NT than the number of piglets induces suckling competition, which can lower pre-weaning growth and survival. Previous linkage studies on NT (Wada *et al.* 2000; Cassady *et al.* 2001; Hirooka *et al.* 2001; Geldermann *et al.* 2003; Holl *et al.* 2004; Sato *et al.* 2006; Zhang *et al.* 2007; Bidanel *et al.* 2008; Guo *et al.* 2008; Ding *et al.* 2009; Tortereau *et al.* 2010) have shown evidence of both additive and dominance effects on this trait. These studies applied low-density microsatellite panels to relatively small experimental crosses, resulting in the identification of QTL with wide confidence intervals. The use of dense SNP panels using a GWAS gives the opportunities to narrow down the QTL regions in purebred populations.

The aim of this study was to detect additive and dominance effects on number of teats, specifically to investigate the importance of dominance using a high-density SNP panel in a Landrace-based population of pigs.

4.2 Material and methods

4.2.1 Genotypes

DNA from 1,795 animals was extracted from blood, hair follicles or ear tissue. Genotyping was performed using the Illumina 60K+SNP Porcine Beadchip (Ramos *et al.* 2009). Positions of the SNPs were based on the Pig genome build10.2 (Groenen *et al.* 2012). The first step of the quality check consisted of excluding SNPs with GenCall score <0.15, with unknown position on the build10.2 (Groenen *et al.* 2012) and SNPs located on both sex chromosomes. Based on these criteria 8,990 SNPs were excluded from the data. Further, 13,315 SNPs were excluded because they failed at least one of the following criteria: call rate <0.95, minor allele frequency <0.01 and/or strong deviation from Hardy-Weinberg Equilibrium

(χ^2 values >600). Finally, 9,016 SNPs were excluded because a genotype class had a frequency <0.02. This last step was necessary because this study focused on both additive and dominance effects and, therefore, observations were necessary in all three genotype classes. After these quality checks, 32,911 out of 64,232 SNPs were used for the GWAS.

In total, 71 individuals with missing genotype frequency >0.05 (based on 32,911 SNPs that passed the quality check) were excluded. In addition, animals that had at least one of their parents genotyped were checked for pedigree inconsistencies. The parental check consisted of comparing the genotypes of the offspring and their parents (one or both parents) at all loci. If a Mendelian inconsistency was detected (e.g. offspring genotype=*BB* and parent genotype=*AA*), the genotype of the offspring at that specific locus was set to missing. Further, if the proportion of Mendelian inconsistencies was >0.01, either a pedigree mistake or a mistake during the genotyping process was assumed and the offspring was excluded from the data set. If the proportion of Mendelian inconsistencies was >0.01 for all offspring of a given parent, the parent was excluded as well, however, this was not observed in the current data set. A total of 17 animals (offspring) were excluded based on the described procedure. A further 68 animals were excluded because their NT was not recorded. Finally, 89 animals were excluded because they were the unique observation from their herd-year-season class, leaving 1,550 genotyped and phenotyped animals for this study.

4.2.2 Animals and phenotypes

The evaluated population consisted of 630 males and 920 females from a Landrace-based line. These animals were born between 2005 and 2012 on 30 different farms. A total of 952 genotyped animals had at least one of their parents genotyped as well. The group of genotyped parents consisted of 138 sires and 145 dams. The NT of each individual was counted at birth as part of standard data recording in a commercial breeding program. Only the total NT was counted. The number of left and right teats, and teat malformations was not recorded. The average NT in the dataset was 15.61±1.05, ranging from 12 to 20 teats.

4.2.3 Association analyses

A single-SNP GWAS for additive and dominance effects on NT was performed using an animal model. To capture both additive and dominance contributions to the variance explained by an SNP in a single model parameter, the genotypes were fitted as a class variable with three levels. The following model was used:

$$y_{ijkl} = \mu + sex_i + hys_j + SNP_k + animal_l + e_{ijkl} \quad (1)$$

where y_{ijkl} was the phenotype of animal l; μ is the overall mean; sex_i was the fixed effect of sex i; hys_j was the fixed effect of the herd (h) year (y) season (s) j of birth (j= 1 to 291); SNP_k was the SNP genotype k (AA, AB or BB) fitted as a fixed effect; $animal_l$ was the random additive genetic effect which was assumed to be distributed as ~ $N(0,\boldsymbol{G}\sigma_a^2)$, which accounted for the (co)variances between animals due to genomic relationships by formation of a ***G*** matrix (genomic relationship matrix); and e_{ijkl} was the random residual effect which was assumed to be distributed as ~ $N(0,\mathbf{I}\sigma_e^2)$. Variance components were re-estimated in each SNP association analysis. The analyses were performed using ASReml v3.0 (Gilmour *et al.* 2009). The ***G*** matrix was used to account for genomic relationships and to reduce the risk of false-positive associations due to population stratification and was computed as described by (VanRaden 2008):

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ'}}{2\sum_{i=1}^{n} p_i\,(1-p_i)}$$

where ***Z*** is a matrix that contains all SNP genotypes of all animals corrected for the allele frequency per SNP; n is the total number of SNPs present in ***Z*** and p_i is the frequency of the allele B of SNP i. The SNP genotypes were coded as 0, 1 and 2, being 0 =*AA*, 1=*AB* and 2=*BB*. Allele frequencies of the current sample were used in the calculations to obtain ***Z*** and p_i.

Residuals were visually inspected for normality based on a QQ-plot of the residuals from the model (1) without an SNP effect, using the *qqnorm*() function in R (R Development Core Team 2011). The inflation factor (lambda) for the distribution of *P*-values from the GWAS was estimated using the *estlambda*() function of the R package GenABEL (Aulchenko *et al.* 2007). A genome-wide False Discovery Rate (FDR) was applied using the R package qvalue (Dabney *et al.* 2010) to avoid false positives due to multiple testing. An FDR ≤0.10 was used to indicate a significant association.

All significant SNPs located within 5 Mb from another significant SNP were considered to belong to the same QTL region. When more than one QTL region was detected on the same chromosome, linkage disequilibrium (LD) was used to assess the dependence of these effects. If the LD (r^2) of all SNP-pairs between the two different regions was <0.70, these regions were considered independent. LD estimates were obtained using Haploview v4.2 (Barrett *et al.* 2005).

The total variance explained by each QTL (σ^2_{QTL}) was estimated as the sum of its additive ($\sigma^2_{QTL\_a}$) and dominance ($\sigma^2_{QTL\_d}$) variances, which were estimated as follows:

$$\hat{a} = (BB - AA)/2$$

$$\hat{d} = AB - (BB + AA)/2$$

$$\alpha = \hat{a} + (q - p)\hat{d}$$

$$\delta = 2pq\hat{d}$$

$$\sigma^2_{QTL\_a} = 2pq(\alpha)^2$$

$$\sigma^2_{QTL\_d} = (\delta)^2$$

$$\sigma^2_{QTL} = \sigma^2_{QTL\_a} + \sigma^2_{QTL\_d}$$

where *p* and *q* are the allele frequencies, $\hat{a}$ the additive and $\hat{d}$ the dominance effects estimated from the genotype effects (*AA*, *BB* and *AB*) of the most significant SNP in a QTL region, α is the allele substitution effect and δ is the dominance deviation. The QTL variance was expressed as a fraction of the total phenotypic variance (σ^2_P, being the summation of the additive and environmental variances) which was estimated based on model (1) without a SNP effect.

4.2.4 Testing for additive and dominance effects

To determine if the SNP had a significant additive effect, dominance effect or both, contrasts for additive and dominance effects were tested for the most significant SNP in each QTL region. Testing was performed using the option !CONTRAST in ASReml v3.0 (Gilmour *et al.* 2009) in model (1). Additive effects were declared when the contrast between the effects of the two homozygous genotypes was significantly different from zero ($P<0.01$). Dominance effects were declared when the contrast between the average effect of the two homozygous genotypes (*AA* and *BB*) and the effect of the heterozygous genotype was significantly different from zero ($P<0.01$).

Results from the current study were compared with previously identified QTL using the alignment of genetic and physical maps in PigQTLdb (Hu *et al.* 2013). Genes located in QTL regions, including flanking regions of 0.2 Mb upstream or downstream of QTL regions, were considered as candidates. Gene searches were carried out with NCBI map viewer (http://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?).

4.3 Results

The additive genetic variance for NT estimated using model (1) without a SNP effect was 0.43 and the corresponding heritability was 0.37±0.05. The estimated effects for sex showed that males presented 0.35±0.09 more teats than females. Although NT is a count variable, the residuals follow a normal distribution (Figure 4.S1). An inflation factor of 1.13 was estimated, indicating that any major effects of population stratification were accounted for in the analyses. In total, 21 SNPs were associated with NT (Figure 4.1). These SNPs were located in four different QTL regions on SSC4, 6, 7 and 12 (Table 4.1).

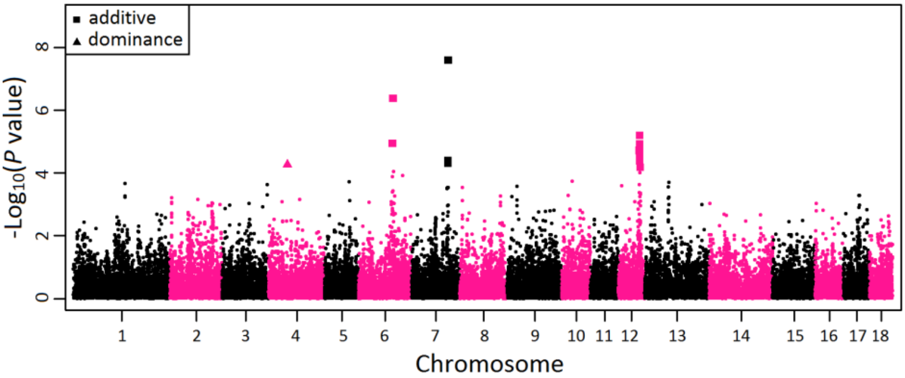

Figure 4.1 Genome-wide association study for additive and dominance effects on number of teats in pigs. On the y-axis is the -log10(*P*-values) of single-SNP association with number of teats in pigs. On the x-axis is the physical position of the SNPs across the 18 autosomes. SNPs associated (false discovery rate ≤0.10) with number of teats having additive and dominance effects are represented by squares and triangles, respectively.

One QTL region was characterized as dominant and three as additive. Estimated effects for QTLs that were characterised as showing additive gene action ranged from 0.24 to 0.29 teats (in absolute values). The QTL that was characterised as showing dominant gene action showed a negative dominance effect (-0.26 teats). The summation of $\sigma^2_{QTL\_a}$ of all four QTLs corresponds to 7.37% of σ^2_P and 23.25% of the additive genetic variance. The summation of $\sigma^2_{QTL\_d}$ of all four QTLs corresponded to 1.82% of σ^2_P, which is one-fourth of the variance explained by additive effects.

Table 4.1 Characterization of the QTL regions.

SSC[1]	Position (Mb)			Most sign. SNP	MAF	MGF	-log$_{10}$(*P*value)	SNP effects[2]		SNP variance (% of σ_P^2)	
	Start		End					$\hat{a}$	$\hat{d}$	$\sigma_{QTL\_a}^2$	$\sigma_{QTL\_d}^2$
4	44.53	-	44.53	ASGA0019540	0.31	0.09	4.27	0.04	-0.26*	0.13	1.14
6	101.77	-	104.41	ALGA0036369	0.36	0.14	6.37	0.27*	-0.18	1.97	0.60
7	103.03	-	103.59	ASGA0035500	0.32	0.12	7.59	0.29*	0.04	3.02	0.02
12	52.71	-	54.68	ALGA0120076	0.43	0.20	5.19	0.24*	0.05	2.25	0.06
Total										**7.37**	**1.82**

Minor allele frequency (MAF), minor genotypic frequency (MGF), -log$_{10}$(*P*-values) of the association analysis, additive ($\hat{a}$) and dominance ($\hat{d}$) effect estimates, proportion of the total phenotypic variance (σ_P^2) explained by the additive and dominance variances ($\sigma_{QTL\_a}^2$ and $\sigma_{QTL\_d}^2$) based on the most significant SNP of each QTL region. [1]SSC: *Sus scrofa* chromosome; *Contrast to evaluate the mode of gene action was significant (*P* <0.01); [2]Additive effects expressed in absolute values.

4.3.1 Additive QTL

The QTL region on SSC6 contained two SNPs with significant associations. This QTL region was located between 101.77 and 104.42 Mb and ALGA0036369 was the most significant SNP with -$\log_{10}$(P-value) of 6.37. This SNP showed an additive effect of 0.27 teats and a dominance effect of -0.18 teats. However, only the contrast for additive effects was significant for this SNP.

On SSC7, between 103.03 and 103.59 Mb, the highest GWAS peak was found for SNP ASGA0035500 with a -$\log_{10}$(P-value) of 7.59. This SNP showed an additive effect of 0.29 teats, a dominance effect of 0.04 teats and explained 3% of the phenotypic variance.

On SSC12, between 52.71 and 54.68 Mb, was located the third most significant QTL region which was also the region characterized by the largest number of significant SNPs in this study (15 SNPs). The most significant SNP in this region (ALGA0120076) showed an additive effect of 0.24 teats and a dominance effect of 0.05 teats.

4.3.2 Dominant QTL

The SNP ASGA0019540 located at 44.53 Mb on SSC4 was the only significant marker in this QTL region. This SNP showed a dominance effect of -0.26 teats and its $\sigma^2_{QTL\_d}$ corresponded to 1% of the total σ^2_P (Table 4.1). The additive effect for this QTL was not significant (P >0.01) and thus, the mode of gene action of this QTL seems purely dominant. This SNP presented a minor allele frequency of 0.31 and a minor genotypic frequency of 0.09 (Table 4.1), indicating that each genotype class consisted of a considerable number of observations.

4.4 Discussion

4.4.1 QTL and candidate genes

The majority of studies that use genomic information in livestock species have been directed towards discovery or use of additive genetic effects. Such studies are generally performed by applying a linear regression to obtain SNP allele substitution effects. In the current study, SNP genotypes were fitted in the model as a class variable. Using this approach, and basically the same data structure typically used in association studies, it was possible to distinguish additive and dominance genetic effects.

In the present study, four QTL regions related to NT were identified. Among these QTLs, three presented significant additive effects, while one only showed significant dominance effect. The proportion of the total phenotypic variance explained by the additive effects was also higher compared to the proportion explained by dominance effects, being respectively, 7.37 and 1.82% of σ_P^2 (Table 4.1). Although these percentages were likely overestimated due to the Beavis effect (Beavis 1998), which especially has an impact when the effects of a SNP are small, these results present convincing evidence that dominance plays a role in the genetic architecture of NT. These results also suggest that additive effects contribute more to the genetic variance of NT than dominance effects. In pigs, other authors have also demonstrated that additive effects contribute more to the genetic variance of traits than dominance effects. Su *et al.* (2012) showed that additive genetic variance of daily gain was 3.73 fold higher than the dominance genetic variance. Recently, Nishio and Satoh (2014) demonstrated for a number of traits in pigs that the contribution of additive effects to the genetic variance was 18-31% higher than the contribution of dominance effects.

All QTL regions identified in this study overlap with QTL regions that have been detected previously in one or more studies (Cassady *et al.* 2001; Holl *et al.* 2004; Sato *et al.* 2006; Guo *et al.* 2008; Ding *et al.* 2009; Tortereau *et al.* 2010). However, this study is the first to describe a dominant QTL effect on SSC4. On this chromosome, previous studies (Guo *et al.* 2008; Ding *et al.* 2009) have shown QTLs with additive effects. In addition, the length of the QTL regions in this study has been considerably reduced. For example, the most significant QTL in this study (SSC7) showed significant associations in the region between 103.03 and 104.35 Mb (length of the region is 1.32 Mb). Guo *et al.* (2008) reported a QTL related to NT on SSC7 with a confidence interval of 112 cM (~112 Mb).

The QTL region on SSC4 contained only a single significant SNP while the region on SSC12 contained 15 significant SNPs. The QTL region on SSC12 covered 1.97 Mb and the average LD (r^2) between the 15 SNPs was 0.78 (Figure 4.2) and the smallest pairwise r^2 between SNPs in this region was 0.56, except for the most distal SNP. The average LD between the significant SNP and the neighbouring SNPs (within 0.2 Mb) in the region on SSC4 (9 SNPs) was very low (0.15). The low LD between SNPs in this region, and with the single significant SNP in particular, explains why the significant associations could not be confirmed by significant associations of neighbouring SNPs with NT. An alternative explanation for observing only one single significant SNP on SSC4 could be that this SNP was misplaced in the Pig genome build10.2 (Groenen *et al.* 2012). However, Pearson correlations (r) between this SNP (genotypes coded as 0, 1 and 2) and all other SNPs used in the GWAS (across the whole-genome) showed that the highest correlations were found with SNPs, who according to the Pig genome build10.2, should be considered its neighbouring SNPs (data not shown). Therefore, there is no evidence suggesting that the location of this SNP is wrong. Thus, it was concluded that although the QTL on SSC4 is only picked up by a single SNP, this QTL is probably not an artefact. However, the effect of this QTL region needs to be confirmed based on independent studies.

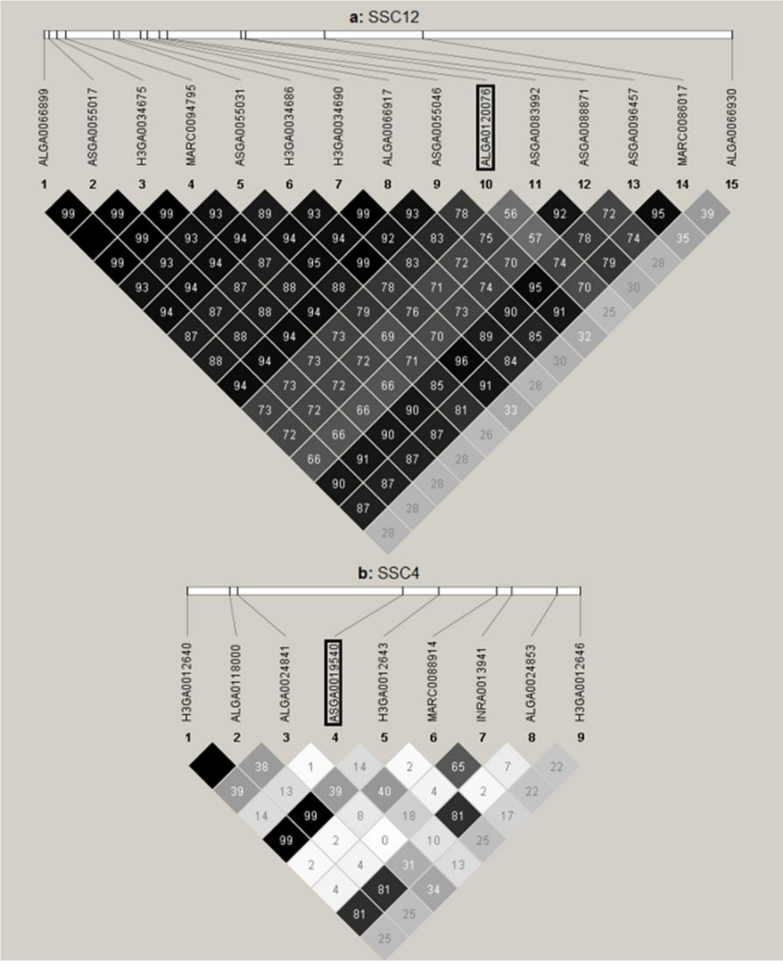

Figure 4.2 Difference in linkage disequilibrium (LD) between two distinct QTL regions. (a) LD (*r2*) between the significant SNPs of the QTL region on *Sus Scrofa* chromosome (SSC) 12; the most significant SNP in this region is surrounded by a square. (b) LD between the SNPs located 0.2 Mb downstream and upstream of the only significant SNP (surrounded by a square) of the QTL region on SSC4. The numbers inside the diamonds are the LD measurements (r^2) on a scale of 0 to 100%.

To distinguish between additive and dominance effects, observations are necessary for all three genotype classes. Therefore, a total of 9,016 SNPs with minor genotypic frequency <0.02 were excluded. The lowest minor genotypic frequency of a significant SNP observed in the current study was 0.09 (143 out of 1,550 individuals) for the dominant QTL in the region SSC4. As the proportion of SNPs excluded based on their minor genotypic frequency was relatively high, an additional analysis was performed to investigate whether any of these 9,016 excluded SNPs was associated with NT, even though these SNPs do not allow the investigation of the mode of gene action. For this analysis, the least frequent genotype class of these SNPs was set to missing and SNPs with two genotype classes <0.02 were not evaluated (n=115). None of these SNPs showed a significant association (FDR<0.10).

The region detected on SSC7 has been identified as a QTL for NT in other populations (Wada *et al.* 2000; Sato *et al.* 2006; Zhang *et al.* 2007; Guo *et al.* 2008; Ding *et al.* 2009), as well as a QTL for carcass length (Ma *et al.* 2009; Steibel *et al.* 2011; Wei *et al.* 2011) and number of vertebrae and ribs (Sato *et al.* 2003; Mikawa *et al.* 2005; Edwards *et al.* 2008; Choi *et al.* 2011; Ren *et al.* 2012). A phenotypic correlation of 0.24 between NT and number of thoracic vertebrae has been estimated (Ren *et al.* 2012), and a larger number of vertebrae is associated with an increase in carcass length and number of ribs (King & Roberts 1960; Borchers *et al.* 2004). Thus, the region is of great interest for pig breeders with favourable pleiotropic effects on economically important traits, including mothering ability of the sows due to the increase in NT, and increased pork production per animal due to longer carcasses.

The Vertnin *(VRTN)* gene appeared as the most promising candidate in this region. *VRTN* encodes a potential DNA binding factor and has been described as an essential factor for the development of the embryo in different species (Mikawa *et al.* 2011). Due to its biological function, this gene has been indicated as a candidate gene for number of vertebrae (Mikawa *et al.* 2011; Ren *et al.* 2012; Fan *et al.* 2013). Recently, Fan *et al.* (2013) performed a fine mapping study aiming to identify the causal mutation of a QTL for number of vertebrae in the same region. By applying an identity-by-descendent sharing method, the QTL region was narrowed down to a 128 Kb region that harboured the *VRTN* gene. The region was defined by two SNPs: ASGA0035500 and INRA0027623, which were, respectively, the first and the third most significant SNPs for NT in the current study. Later, Fan *et al.* (2013) identified a possible causal mutation in the *VRTN* gene. Due to the positive relation between NT and number of vertebrae and the similarities between the results on

SSC7 of the present study and the results of Fan *et al.* (2013), it can be assumed that the *VRTN* gene may also have an effect on NT.

In the other QTL regions, no obvious genes that could effectively affect NT were identified. The relationship between *VRTN* and NT needs to be further investigated in order to validate the effect of this gene on the genetic architecture of NT.

4.4.2 Implications

The term heterosis was coined by Shull (1914) to describe an improved performance of crossed individuals compared to the average performance of their parental inbred lines. However, the performance of crossbreds depends partly on the degree and sign of the dominance effects of the loci affecting the trait (Sellier 1976; Falconer & Mackay 1996), which can also lead to negative heterosis (crossbreds performing worse than the average of their parents). Thus, the definition of heterosis being the deviation of crossbred performance compared to the average performance of the two parental breeds (Falconer & Mackay 1996) is more appropriate.

In pigs, negative heterosis (also called outbreeding depression or hybrid inferiority) has not often been reported in the literature. Bereskin *et al.* (1971) showed that crossbred pigs on average had higher levels of backfat and lower levels of ham and loin percentage than their purebred parents. However, more examples of negative heterosis have been published in other species. In Drosophila, negative heterosis has been identified for the degree of deficient venation (Stern 1948) and in an F1 chicken population, negative heterosis was reported for leukocyte ratio at 8 weeks of age (Campo & Davila 2002). Minozzi *et al.* (2008) reported negative direct heterosis for general immune response traits in White Leghorn chickens. Barbato (1992) observed negative heterosis for abdominal fat in chickens. Denic *et al.* (2005) described that negative heterosis in humans is related to higher rates of breast and ovarian cancer.

In the current study, the dominant QTL identified on SSC4 showed a negative estimate for dominance effect (-0.26 teats). Based on this locus, negative heterosis would be expected for NT, assuming that dominance effects are the main cause of heterosis. However, it is important to keep in mind that the main cause of heterosis is still under debate. While it has been shown in few studies that dominance is an important factor contributing to heterosis (Davenport 1908; Shull 1908; Bruce 1910; Xiao *et al.* 1995; Visscher *et al.* 2000), in other studies, the main cause of heterosis has been attributed to epistasis (Melchinger *et al.* 2007; Moyle &

Nakazato 2009; Alvarez-Castro *et al.* 2012). Recently, Amuzu-Aweh *et al.* (2013) evaluating egg production traits in chickens, showed that although dominance cannot fully explain heterosis, a dominance model can achieve a considerable accuracy of prediction of heterosis. In pigs, the genetic background of heterosis has not been elucidated. Therefore, epistatic interactions also might play a role; however, in segregating populations, the power to identify epistatic interactions between QTLs is low (Melchinger *et al.* 2007).

As a further step, the role of dominance effects on the genetic architecture of NT effects should be evaluated in a crossbred population, since non-additive effects are expected to be of importance in crossbreeding (Dekkers *et al.* 2011; Goddard 2012). Nonetheless, the results of this study showed that dominance effects explain an important fraction of the phenotypic variance even in a purebred population.

The genotype effects of the QTL region on SSC7 (Figure 4.3) showed that this QTL has a clear additive effect. For such QTL, the traditional selection that is based on allele substitution effects would be sufficient, as selection for higher NT would lead to the fixation of the favourable allele B (ignoring the potential impact of drift).

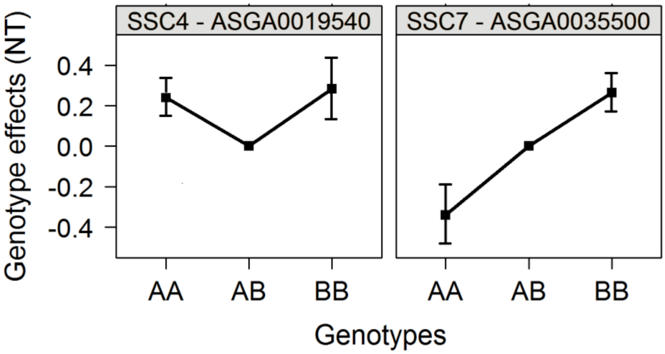

Figure 4.3 Genotype effects and their standard errors of the most significant SNPs on *Sus scrofa* chromosomes (SSC) 4 and 7 on number of teats (NT). The genotypic effects are relative to the effect of the heterozygous genotype, which was set to zero.

A more challenging situation is encountered with the QTL region on SSC4 which may require the adoption of different strategies, such as mate allocation. Applying an additive model for estimating breeding values would not be efficient because the additive effect of this QTL is close to zero (Figure 4.3). In cases of overdominance, selection tends to keep heterozygotes in the population instead of fixing one of the alleles (Wang *et al.* 2004). However, in order to improve the population mean, the goal for these two dominant QTL must be the fixation of one of the alleles in order to avoid heterozygous animals with their negative dominance effects. More specifically, for the QTL on SSC4, the selection should be towards the fixation of the A allele because this is the most frequent allele (f_A=0.69). If selection is aimed at fixation of the B allele, it would take longer before this allele becomes fixed, and in the meantime, an increase in the frequency of AB animals would be observed, negatively affecting the mean NT of the population. Finally, when this QTL presents the same effect on different lines, all lines within a breeding program should be fixed for the same allele in order to maximize the performance of crossbred animals.

According to Toro & Varona (2010), it is easier to include dominance effects in genomic evaluations compared to including them in the traditional selection using pedigree information. These authors concluded that the use of dominance effects in a scenario of genomic selection increases the accuracy of estimated breeding values and still offers the opportunity of applying mate-allocation. Wang *et al.* (2004) described that the genetic progress of traits controlled only by additive genetic effects will generally achieve the target genotype faster than traits with considerable overdominance. Although the genetic progress is slower in the presence of dominance compared to the situation when only additive effects play a role, if dominance effects exist and are not properly taken into account, the genetic progress may be even slower.

4.5 Conclusions

In this study, four QTLs, three additive and one dominant, were identified by applying a two-step approach; first testing for significant genetic effect and then testing for additive and/or dominant gene action only of the SNPs with significant genetic effects. In total, the $\sigma^2_{QTL\_d}$ corresponded to approximately one-fourth of the variance that was explained by $\sigma^2_{QTL\_a}$, demonstrating that dominance effects play a role in the genetic architecture of NT. The QTLs with significant additive effects overlap with earlier identified QTLs, however, the QTL regions were considerably reduced in size. Selection based on these QTLs would benefit mothering ability of the sows due to the increase in NT, as well as increasing pork production of finishing pigs due to pleiotropic effects on number of vertebrae and carcass length.

4.6 Acknowledgement

Financial support from Topigs Norsvin and Breed4Food Consortium is gratefully acknowledged.

4.7 References

Alvarez-Castro, J.M., Le Rouzic, A., Andersson, L., Siegel, P.B., Carlborg, Ö. (2012) Modelling of genetic interactions improves prediction of hybrid patterns–a case study in domestic fowl. *Gent. Res.,* **94,** 255-266.

Amuzu-Aweh, E., Bijma, P., Kinghorn, B., Vereijken, A., Visscher, J., van Arendonk, J.A., Bovenhuis, H. (2013) Prediction of heterosis using genome-wide SNP-marker data: application to egg production traits in white Leghorn crosses. *Heredity (Edinb.),* **111,** 530-538.

Aulchenko, Y.S., Ripke, S., Isaacs, A., van Duijn, C.M. (2007) GenABEL: an R package for genome-wide association analysis. *Bioinformatics,* **23,** 1294-1296.

Barbato, G. (1992) Genetic architecture of carcass composition in chickens. *Poult. Sci.,* **71,** 789-798.

Barrett, J.C., Fry, B., Maller, J., Daly, M. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* **21,** 263-265.

Beavis, W.D. (1998) QTL analyses: power, precision, and accuracy. *Molecular dissection of complex traits* (ed. A.H. Peterson), pp. 145-162. CRC Press, New York.

Bereskin, B., Shelby, C., Hazel, L. (1971) Carcass traits of purebred Durocs and Yorkshires and their crosses. *J. Anim. Sci.,* **32,** 413-419.

Bidanel, J.P., Rosendo, A., Iannuccelli, N., Riquet, J., Gilbert, H., Caritez, J.C., Billon, Y., Amigues, Y., Prunier, A., Milan, D. (2008) Detection of quantitative trait loci for teat number and female reproductive traits in Meishan× Large White F2 pigs. *Animal,* **2,** 813-820.

Borchers, N., Reinsch, N., Kalm, E. (2004) The number of ribs and vertebrae in a Pietrain cross: variation, heritability and effects on performance traits. *J. Anim. Breed. Genet.,* **121,** 392-403.

Bouwman, A.C., Bovenhuis, H., Visker, M.H., van Arendonk, J.A. (2011) Genome-wide association of milk fatty acids in Dutch dairy cattle. *BMC Genet.,* **12,** 43.

Bouwman, A.C., Visker, M.H., van Arendonk, J.A., Bovenhuis, H. (2012) Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. *BMC Genet.,* **13,** 93.

Boysen, T.-J., Heuer, C., Tetens, J., Reinhardt, F., Thaller, G. (2013) Novel use of derived genotype probabilities to discover significant dominance effects for milk production traits in dairy cattle. *Genetics,* **193,** 431-442.

Bruce, A.B. (1910) The Mendelian theory of heredity and the augmentation of vigor. *Science,* **32,** 627-628.

Campo, J., Davila, S. (2002) Estimation of heritability for heterophil: lymphocyte ratio in chickens by restricted maximum likelihood. Effects of age, sex, and crossing. *Poult. Sci.,* **81,** 1448-1453.

Cassady, J.P., Johnson, R., Pomp, D., Rohrer, G., Van Vleck, L.D., Spiegel, E., Gilson, K. (2001) Identification of quantitative trait loci affecting reproduction in pigs. *J. Anim. Sci.,* **79,** 623-633.

Choi, I., Steibel, J.P., Bates, R.O., Raney, N.E., Rumph, J.M., Ernst, C.W. (2011) Identification of carcass and meat quality QTL in an F2 Duroc × Pietrain pig resource population using different least-squares analysis models. *Front. Genet.,* **2,** 18.

Coster, A., Madsen, O., Heuven, H.C., Dibbits, B., Groenen, M.A., van Arendonk, J.A., Bovenhuis, H. (2012) The imprinted gene DIO3 is a candidate gene for litter size in pigs. *PLoS One,* **7,** e31825.

Dabney, A., Storey, J., Warnes, G. (2010) qvalue: Q-value estimation for false discovery rate control. R package version 1.26. 0.

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

Dekkers, J.C., Mathur, P.K., Knol, E.F. (2011) Genetic improvement of the pig. *The Genetics of the Pig* (eds M.F. Rothschild & A. Ruvinsky), pp. 390-425. CABI.

Denic, S., Khatib, F., Awad, M., Karbani, G., Milenkovic, J. (2005) Cancer by negative heterosis: breast and ovarian cancer excess in hybrids of inbred ethnic groups. *Med. Hypotheses,* **64,** 1002-1006.

Ding, N., Guo, Y., Knorr, C., Ma, J., Mao, H., Lan, L., Xiao, S., Ai, H., Haley, C., Brenig, B. (2009) Genome-wide QTL mapping for three traits related to teat number in a White Duroc × Erhualian pig resource population. *BMC Genet.,* **10,** 6.

Edwards, D.B., Ernst, C.W., Raney, N.E., Doumit, M.E., Hoge, M.D., Bates, R.O. (2008) Quantitative trait locus mapping in an F2 Duroc × Pietrain resource population: II. Carcass and meat quality traits. *J. Anim. Sci.,* **86,** 254-266.

Falconer, D.S., Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics,* 4th edn. Longmans Green, Harlow.

Fan, Y., Xing, Y., Zhang, Z., Ai, H., Ouyang, Z., Ouyang, J., Yang, M., Li, P., Chen, Y., Gao, J. (2013) A Further Look at Porcine Chromosome 7 Reveals VRTN Variants Associated with Vertebral Number in Chinese and Western Pigs. *PLoS One,* **8,** e62534.

Geldermann, H., Müller, E., Moser, G., Reiner, G., Bartenschlager, H., Cepica, S., Stratil, A., Kuryl, J., Moran, C., Davoli, R. (2003) Genome-wide linkage and QTL mapping in porcine F2 families generated from Pietrain, Meishan and Wild Boar crosses. *J. Anim. Breed. Genet.,* **120,** 363-393.

Gilmour, A.R., Gogel, B., Cullis, B., Thompson, R. (2009) ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

Goddard, M. (2012) Uses of genomics in livestock agriculture. *Anim. Prod. Sci.,* **52,** 73-77.

Groenen, M.A., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature,* **491,** 393-398.

Guo, Y.M., Lee, G., Archibald, A., Haley, C. (2008) Quantitative trait loci for production traits in pigs: a combined analysis of two Meishan × Large White populations. *Anim. Genet.,* **39,** 486-495.

Hirooka, H., De Koning, D., Harlizius, B., Van Arendonk, J., Rattink, A., Groenen, M., Brascamp, E., Bovenhuis, H. (2001) A whole-genome scan for quantitative trait loci affecting teat number in pigs. *J. Anim. Sci.,* **79,** 2320-2326.

Holl, J., Cassady, J., Pomp, D., Johnson, R. (2004) A genome scan for quantitative trait loci and imprinted regions affecting reproduction in pigs. *J. Anim. Sci.,* **82,** 3421-3429.

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

King, J., Roberts, R. (1960) Carcass length in the bacon pig: its association with vertebrae numbers and prediction from radiographs of the young pig. *Anim. Prod.,* **2,** 59-65.

Ma, J., Ren, J., Guo, Y., Duan, Y., Ding, N., Zhou, L., Li, L., Yan, X., Yang, K., Huang, L. (2009) Genome-wide identification of quantitative trait loci for carcass composition and meat quality in a large-scale White Duroc× Chinese Erhualian resource population. *Anim. Genet.,* **40,** 637-647.

Melchinger, A.E., Piepho, H.-P., Utz, H.F., Muminović, J., Wegenast, T., Törjék, O., Altmann, T., Kusterer, B. (2007) Genetic basis of heterosis for growth-related traits in Arabidopsis investigated by testcross progenies of near-isogenic lines reveals a significant role of epistasis. *Genetics,* **177,** 1827-1837.

Mikawa, S., Hayashi, T., Nii, M., Shimanuki, S., Morozumi, T., Awata, T. (2005) Two quantitative trait loci on Sus scrofa chromosomes 1 and 7 affecting the number of vertebrae. *J. Anim. Sci.,* **83,** 2247-2254.

Mikawa, S., Sato, S., Nii, M., Morozumi, T., Yoshioka, G., Imaeda, N., Yamaguchi, T., Hayashi, T., Awata, T. (2011) Identification of a second gene associated with variation in vertebral number in domestic pigs. *BMC Genet.,* **12,** 5.

Minozzi, G., Bidanel, J., Minvielle, F., Bed'Hom, B., Gourichon, D., Baumard, Y., Pinard-van der Laan, M. (2008) Crossbreeding parameters of general immune response traits in White Leghorn chickens. *Livest. Sci.,* **119,** 221-228.

Moyle, L.C., Nakazato, T. (2009) Complex epistasis for Dobzhansky–Muller hybrid incompatibility in Solanum. *Genetics,* **181,** 347-351.

Nishio, M., Satoh, M. (2014) Including dominance effects in the genomic BLUP method for genomic evaluation. *PLoS One,* **9,** e85792.

R Development Core Team (2011) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

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

Ren, D., Ren, J., Ruan, G., Guo, Y., Wu, L., Yang, G., Zhou, L., Li, L., Zhang, Z., Huang, L. (2012) Mapping and fine mapping of quantitative trait loci for the number of vertebrae in a White Duroc × Chinese Erhualian intercross resource population. *Anim. Genet.,* **43,** 545-551.

Rodriguez, C., Tomas, A., Alves, E., Ramirez, O., Arque, M., Munoz, G., Barragan, C., Varona, L., Silio, L., Amills, M. (2005) QTL mapping for teat number in an Iberian-by-Meishan pig intercross. *Anim. Genet.,* **36,** 490-496.

Sato, S., Atsuji, K., Saito, N., Okitsu, M., Komatsuda, A., Mitsuhashi, T., Nirasawa, K., Hayashi, T., Sugimoto, Y., Kobayashi, E. (2006) Identification of quantitative trait loci affecting corpora lutea and number of teats in a Meishan × Duroc F2 resource population. *J. Anim. Sci.,* **84,** 2895-2901.

Sato, S., Oyamada, Y., Atsuji, K., Nade, T., Sato, S.-i., Kobayashi, E., Mitsuhashi, T., Nirasawa, K., Komatsuda, A., Saito, Y. (2003) Quantitative trait loci analysis for growth and carcass traits in a Meishan × Duroc F2 resource population. *J. Anim. Sci.,* **81,** 2938-2949.

Sellier, P. (1976) The basis of crossbreeding in pigs; a review. *Livest. Prod. Sci.,* **3,** 203-226.

Shull, G.H. (1908) The composition of a field of maize. *J. Hered.***,** 296-301.

Shull, G.H. (1914) Duplicate genes for capsule-form inBursa bursa-pastoris. *Zeitschrift für Induktive Abstammungs-und Vererbungslehre,* **12,** 97-149.

Steibel, J.P., Bates, R.O., Rosa, G.J., Tempelman, R.J., Rilington, V.D., Ragavendran, A., Raney, N.E., Ramos, A.M., Cardoso, F.F., Edwards, D.B. (2011) Genome-wide linkage analysis of global gene expression in loin muscle tissue identifies candidate genes in pigs. *PLoS One,* **6,** e16766.

Stern, C. (1948) Negative heterosis and decreased effectiveness of alleles in heterozygotes. *Genetics,* **33,** 215.

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

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

Tortereau, F., Gilbert, H., Heuven, H., Bidanel, J.-P., Groenen, M., Riquet, J. (2010) Combining two Meishan F2 crosses improves the detection of QTL on pig chromosomes 2, 4 and 6. *Genet. Sel. Evol.,* **42,** 42.

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

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

Vitezica, Z.G., Varona, L., Legarra, A. (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics,* **195,** 1223-1230.

Wada, Y., Akita, T., Awata, T., Furukawa, T., Sugai, N., Ishii, K., Ito, Y., Kobayashi, E., Mikawa, S., Yasue, H. (2000) Quantitative trait loci (QTL) analysis in a Meishan × Göttingen cross population. *Anim. Genet.,* **31,** 376-384.

Wang, J., van Ginkel, M., Trethowan, R., Ye, G., DeLacy, I., Podlich, D., Cooper, M. (2004) Simulating the effects of dominance and epistasis on selection response in the CIMMYT Wheat Breeding Program using QuCim. *Crop. Sci.,* **44,** 2006-2018.

Wei, W., Duan, Y., Haley, C., Ren, J., de Koning, D., Huang, L. (2011) High throughput analyses of epistasis for swine body dimensions and organ weights. *Anim. Genet.,* **42,** 15-21.

Wellmann, R., Bennewitz, J. (2012) Bayesian models with dominance effects for genomic evaluation of quantitative traits. *Gent. Res.,* **94,** 21.

Wijga, S., Bastiaansen, J., Wall, E., Strandberg, E., de Haas, Y., Giblin, L., Bovenhuis, H. (2012) Genomic associations with somatic cell score in first-lactation Holstein cows. *J. Dairy Sci.,* **95,** 899-908.

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

Zhang, J., Xiong, Y., Zuo, B., Lei, M., Jiang, S., Li, F.e., Zheng, R., Li, J., Xu, D. (2007) Detection of quantitative trait loci associated with several internal organ traits and teat number trait in a pig population. *J. Gent. Genomics,* **34,** 307-314.

4.8 Supporting information

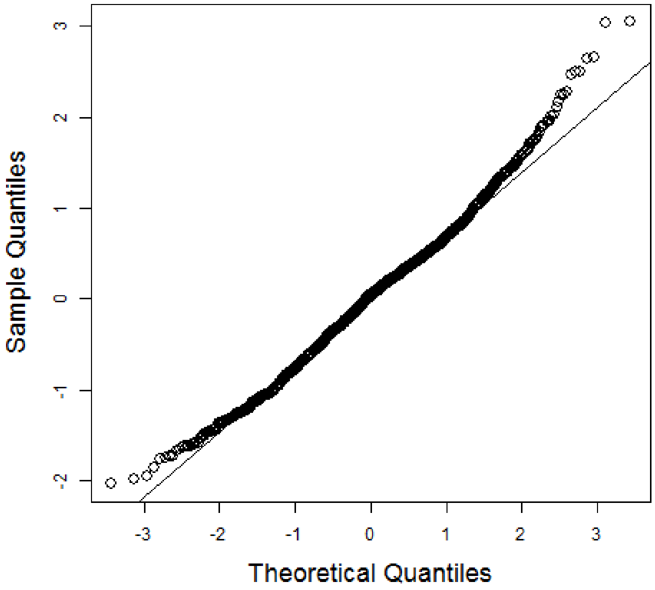

Figure 4.S1 QQ-plot of the residuals from the linear model without a SNP effect.

5

Using markers with large effect in genetic predictions

Marcos S Lopes[1,2], Henk Bovenhuis[2], Maren van Son[3], Øyvind Nordbø[3],
Eli H Grindflek[3], Egbert F Knol[1], John WM Bastiaansen[2]

[1] Topigs Norsvin Research Center, 6640 AA, Beuningen, the Netherlands;
[2] Wageningen University, Animal Breeding and Genomics Centre, 6700 AH, Wageningen, the Netherlands, [3] Norsvin, 2317, Hamar, Norway

Under review

Abstract

The first attempts of applying marker-assisted selection (MAS) in animal breeding were not very successful because the identification of reliable markers linked to Quantitative Trait Loci (QTL) using low-density microsatellite panels was difficult. More recently, the use of high-density Single Nucleotide Polymorphisms (SNPs) panels in Genome-Wide Association Studies (GWAS) have increased the power and precision of identifying markers linked to QTL. However, when GWAS started to be performed, the focus of many breeders had already shifted to the simultaneous use of all markers by applying genomic selection. In this study, we aimed to revive the traditional MAS approach by including GWAS findings in the prediction models. This approach consisted of including the most significant SNP from GWAS as a fixed effect in the prediction models: marker-assisted BLUP (MA-BLUP) and marker-assisted GBLUP (MA-GBLUP). To compare the prediction accuracies of BLUP, MA-BLUP, GBLUP, and MA-GBLUP we applied this approach to the trait "number of teats" in four distinct pig populations. In all four evaluated populations, the most significant SNP was located at ~103.50 Mb on chromosome 7. Accounting for the most significant SNP in the genetic predictions resulted in improved prediction accuracy for number of teats in all evaluated populations. By replacing BLUP by MA-BLUP, the increase in prediction accuracy ranged from 0.021 to 0.124, whereas by replacing GBLUP by MA-GBLUP, the increase in prediction accuracy ranged from 0.003 to 0.043. BLUP resulted in the lowest prediction accuracy and MA-GBLUP in the highest for all populations. In the same dataset, MA-BLUP can yield similar or superior accuracies compared to GBLUP. The superiority of MA-GBLUP over traditional GBLUP is more pronounced when training populations are smaller and when relationships between training and validation populations are smaller.

Key words: GWAS, MAS, genomic selection

5.1 Introduction

Marker-assisted selection (MAS) is a strategy where one or a few markers linked to Quantitative Trait Loci (QTL) are used as a selection tool. In the early 1990s, it was extensively discussed that the application of MAS would revolutionize the development of agricultural practices (Lande & Thompson 1990). In fact, with the use of low-density panels of DNA markers, such as microsatellites, many QTL were identified (Jonas & de Koning 2015) and some of these QTL was also included in MAS schemes (Boichard *et al.* 2002; Silva *et al.* 2014). However, the gain in genetic progress from MAS was much lower than expected, in both animal and plant breeding. One of the main reasons was that identification of reliable markers linked to QTL was difficult (Jonas & de Koning 2015), especially due to the low availability of markers (Heffner *et al.* 2009).

In the last decade, dense panels of Single Nucleotide Polymorphisms (SNPs) became available (Van Tassell *et al.* 2008; Ramos *et al.* 2009). Since then, SNPs have been used in many Genome-Wide Associations Studies (GWAS) for a large variety of traits. These GWAS enabled not only the identification of novel QTL, but also the reduction of confidence intervals of previously identified QTL (Lopes *et al.* 2014). However, when GWAS started to be performed, the focus of many breeders had already shifted from the use of MAS to the use of all markers (without pre-selection) by applying genomic selection (Meuwissen *et al.* 2001).

The most common genomic selection strategy is to replace the pedigree-based relationship matrix that is used for best linear unbiased prediction (BLUP) by a genomic-based relationship matrix (GBLUP). With GBLUP, it is assumed that quantitative traits are controlled by a large number of genes and each gene explains a small amount of the variance of the trait (infinitesimal model) (Goddard 2009). However, this assumption of GBLUP leads to a suboptimal prediction accuracy (Meuwissen & Goddard 2010) because many quantitative traits are expected to be controlled by a limited number of genes with moderate to large effects rather than by a large number of genes with small effects (Hayes & Goddard 2001).

An approach to overcome this limitation of GBLUP is to include GWAS findings in the GBLUP-based prediction models. Doing so, we can benefit from the linkage disequilibrium between SNPs and the QTL identified in GWAS as well as from the realized family relationship from GBLUP. In dairy cattle, Brøndum *et al.* (2015) showed that a GBLUP model that accounted for significant SNPs from GWAS presented higher prediction accuracy compared to traditional GBLUP. These authors first identified the most significant SNPs associated with the QTL using whole-genome sequence data. Afterwards, these SNPs were genotyped with a custom low-density SNP array and used in combination with the commonly-used 54K SNP array to predict breeding values. In pigs, the availability of whole-genome sequence data is still limited. However, GWAS with 60K SNP array data has successfully identified many QTL regions, such as the region on chromosome 7 for number of teats (Duijvesteijn *et al.* 2014; Lopes *et al.* 2014). Including the most significant SNP associated in this QTL region in the prediction model was expected to lead to an increased prediction accuracy, similar to results by Brøndum *et al.* (2015), even though SNPs were obtained using the 60K SNP array data instead of whole-genome sequence data.

In this study, we aimed to revive the MAS approach by including GWAS findings in the prediction models. This approach consisted of including the most significant SNP from GWAS as a fixed effect in the prediction models: marker-assisted BLUP (MA-BLUP) and marker-assisted GBLUP (MA-GBLUP). Therefore, the effect of the SNP (QTL) and the polygenic effect are estimated simultaneously, as already described in the late 1980s (Fernando & Grossman 1989). In order to validate the potential of incorporating GWAS findings in MA-BLUP and MA-GBLUP, we applied this approach to “number of teats”, a trait for which an important QTL is known to segregate (Duijvesteijn *et al.* 2014; Lopes *et al.* 2014) in two out of four pig populations here evaluated. The advantage of MA-BLUP and MA-GBLUP was assessed as increase in prediction accuracy over results from BLUP and GBLUP.

5.2 Material and methods

5.2.1 Data

Number of teats was recorded at birth in four pig populations: Large White, Dutch Landrace, Norwegian Landrace and Duroc (See Table 5.1 for descriptive statistics). The Large White and Dutch Landrace populations were located in Dutch nucleus farms. The Norwegian Landrace and Duroc populations were located in Norwegian nucleus farms and a boar testing station. Three datasets from each population were used in this study: ALL, TRAINING, and VALIDATION.

Table 5.1 Descriptive statistics.

Population	Dataset	N	Mean	SD
Large White	ALL	322,887	15.05	1.05
	TRAINING	2,620	15.37	0.96
	VALIDATION	665	15.65	0.98
Dutch Landrace	ALL	439,809	15.27	1.07
	TRAINING	2,491	15.61	1.02
	VALIDATION	622	15.78	1.04
Norwegian Landrace	ALL	210,289	15.70	0.99
	TRAINING	6,090	15.92	0.95
	VALIDATION	1,522	16.06	0.97
Duroc	ALL	8,118	13.02	1.05
	TRAINING	3,798	12.98	1.04
	VALIDATION	950	13.00	1.00

Number of phenotyped animals (N), mean and standard deviation (SD) of number of teats in each dataset of each population. ALL: the whole population used in the pre-adjustment of the phenotypes, which includes the animals from TRAINING, VALIDATION and their contemporaries; TRAINING: genotyped and phenotyped animals used for the GWAS and also as reference population in the genetic prediction analysis. VALIDATION: dataset used for prediction accuracy.

The dataset ALL consisted of all genotyped animals and their contemporaries that had phenotypes (322,887 Large White, 439,809 Dutch Landrace, 210,289 Norwegian Landrace and 8,118 Duroc). Using ALL, the phenotypes (number of teats) were pre-corrected for all non-genetic effects. The pre-corrected phenotype was used as the response variable in further analysis. The non-genetic effects were estimated by a pedigree-based linear model in ASReml v3.0 (Gilmour *et al.* 2009):

$$y_{ijkl} = \mu + \text{sex}_i + \text{hy}_j + \text{animal}_k + \text{litter}_l + e_{ijkl} \qquad (1)$$

where y_{ijkl} was the number of teats of animal *k*; μ is the overall mean; sex_i was the fixed effect of sex *i*; hy_j was the fixed effect of the herd-year *j* of birth; animal_k was the random additive genetic effect of animal k, which was assumed to be distributed as $\sim N(\mathbf{0},\mathbf{A}\sigma_a^2)$, which accounted for the (co)variances between animals due to relationships by formation of an **A** matrix (pedigree-based average numerator relationship matrix), σ_a^2 being the additive genetic variance; litter_l was the random effect of litter *l*, which was assumed to be distributed as $\sim N(\mathbf{0},\mathbf{I}\sigma_l^2)$, with **I** being an identity matrix and σ_l^2 the litter variance; and e_{ijkl} was the random residual effect which was assumed to be distributed as ~$N(\mathbf{0},\mathbf{I}\sigma_e^2)$, σ_e^2 being the residual variance.

The dataset TRAINING was a subset of ALL consisting of the oldest 80% of the animals that had both phenotypes and genotypes (2,674 Large White, 2,566 Dutch Landrace, 6,072 Norwegian Landrace and 3,795 Duroc animals). This dataset was used to perform the GWAS and also as the reference population for prediction of breeding values.

The dataset VALIDATION consisted of the remaining 20% youngest animals that had both phenotypes and genotypes (668 Large White, 641 Dutch Landrace, 1,518 Norwegian Landrace, and 949 Duroc animals). This dataset was used to assess the prediction accuracy of the evaluated models as described below in the "Prediction of breeding values" section.

5.2.2 Genotypes

Genotyping was performed at CIGENE (University of Life Sciences, Ås, Norway) and at GeneSeek (Lincoln, NE, USA), mainly using the Illumina Porcine SNP60 v1 Beadchip (Illumina, San Diego, CA, USA). Part of the animals from the Large White and Dutch Landrace population were genotyped using the Illumina Porcine SNP60 v2 Beadchip (Illumina, San Diego, CA, USA). All animals were imputed to the set of SNPs on the SNP60 Beadchip v1 that passed the quality control. Imputation was performed using Beagle v3.3.2 (Browning & Browning 2007). Quality control consisted of excluding SNPs with GenCall<0.15, call rate <0.95, minor allele frequency <0.02, strong deviation from Hardy-Weinberg equilibrium (χ^2>600), SNPs located on sex chromosomes and unmapped SNPs. The position of the SNPs was based on the Sscrofa10.2 assembly of the reference genome (Groenen *et al.* 2012). All genotyped animals had a frequency of missing SNP genotypes <0.05. With this, 43,439 SNPs for Large White, 41,077 SNPs for Dutch Landrace, 38,085 SNPs for Norwegian Landrace and 36,131 SNPs for Duroc were available for further analyses.

5.2.3 Association analyses

A single-SNP GWAS was performed within population using the following animal model:

$$y^*_k = \mu + \beta X + \text{animal}_k + e_k \qquad (2)$$

where y^*_k was the pre-corrected phenotype of animal *k*; μ and animal_k were as previously defined for model (1); X was the genotype (0, 1, 2) of animal *k* for the evaluated SNP; β was the unknown allele substitution effect of the evaluated SNP; and e$_k$ was the random residual effect which was assumed to be distributed as ~$N(\mathbf{0},\mathbf{I}\sigma_e^2)$. The association analyses were performed with the TRAINING dataset within each population using ASReml v3.0 (Gilmour *et al.* 2009).

The variance explained by an SNP (σ_{snp}^2) was estimated based on the allele frequency and the estimated allele substitution effect. The proportion of phenotypic variance explained by the SNP was defined as $\sigma_{snp}^2/\sigma_P^2$, where σ_P^2 is total phenotypic variance (sum of the additive and residual variances) which was estimated based on model (2) without a SNP effect.

5.2.4 Prediction of breeding values

From the GWAS, we selected the most significant (smallest *P* value) SNP in each population to be included in the marker-assisted models for within-line prediction. Four models were evaluated: BLUP, GBLUP, MA-BLUP and MA-GBLUP. The models BLUP and GBLUP were similar to model (2), but without the fixed effect βX. In GBLUP, a **G** matrix instead of an **A** matrix was used to account for the (co)variances between animals. The models MA-BLUP and MA-GBLUP were equal to model (2), except that in MA-GBLUP, a **G** matrix instead of an **A** matrix was used to account for the (co)variances between animals. The **G** matrix was built according to VanRaden (2008), using $\mathbf{G} = \mathbf{ZZ}'/2\sum pq$, where **Z** is a matrix of centered genotypes, and *p* and *q* are the allele frequencies of the SNPs.

In MA-GBLUP, the SNP that was fitted as fixed effect was also included in calculating the **G** matrix. To test whether using this SNP in both parts of the model has an effect on the accuracy of the MA-GBLUP, the SNP used as a fixed effect and all other SNPs in high linkage disequilibrium (LD, r^2>0.50) with it were excluded from the set of SNPs used to build the **G** matrix. The pairwise LD between the SNP used as a fixed effect in the model and all other SNPs on the chromosome was estimated on the TRAINING dataset using the software PLINK v1.07 (Purcell *et al.* 2007).

The prediction accuracy of the models was defined as the correlation between the estimated breeding values and the corrected phenotypes of animals in the VALIDATION dataset. For the models BLUP and GBLUP, breeding values were obtained directly from the analysis, e.g. the polygenic breeding value of animal *k* ($\hat{u}_k$) was defined as the term animal$_k$ from model (2). For MA-BLUP and MA-GBLUP, the breeding value was defined as the sum of the marker breeding value ($\hat{u}_{\text{snp}}$= βX) and the polygenic breeding value ($\hat{u}_k$). Finally, prediction bias was assessed by regressing the corrected phenotypes on the breeding values.

5.3 Results

5.3.1 Association analyses

In all four evaluated populations, the most significant SNP was located at ~103.5 Mb on chromosome 7 (Figure 5.1 and Table 5.2). In the Large White population, the most significant SNP explained 3.48% of the phenotypic variance. The phenotypic variance in the Large White population was 0.89±0.03 and the corresponding heritability was 0.41±0.04. In the Dutch Landrace population, the most significant SNP explained 3.67% of the phenotypic variance. The phenotypic variance in the Dutch Landrace population was 0.98±0.03 and the corresponding heritability was 0.36±0.04. In the Norwegian Landrace population, the most significant SNP explained 3.30% of the phenotypic variance. The phenotypic variance in the Norwegian Landrace population was 0.76±0.02 and the corresponding heritability was 0.27±0.03. In the Duroc population, the most significant SNP was the same as in the Large White population. In the Duroc population, this SNP explained 6.13% of the phenotypic variance, which is almost twice the variance explained by this SNP in the Large White population. The phenotypic variance in the Duroc population was 1.00±0.03 and the corresponding heritability was 0.29±0.04.

5.3.2 Prediction of breeding values

In all populations, the lowest prediction accuracy was observed for BLUP and the highest for MA-GBLUP (Table 5.3). In the Dutch Landrace population, we observed the lowest accuracies compared to the other populations for all models, except for BLUP, where the lowest accuracy was observed for the Duroc population. In the Norwegian Landrace population, which had the largest training dataset, the highest prediction accuracies were seen compared to the other populations for all models.

Using BLUP, predictions were more biased, overestimating the genetic variance, compared to MA-BLUP in the Large White, Dutch Landrace, and Duroc population (Table 5.3). In the Norwegian Landrace population, the regression coefficients were 1.12 using BLUP and 0.87 using MA-GBLUP. Both GBLUP and MA-GBLUP resulted in a similar bias of prediction in the Large White and Norwegian Landrace population. In the Dutch Landrace and Duroc population, the bias of predictions was more using GBLUP compared to MA-GBLUP.

When we excluded the SNP used as a fixed effect and all other SNPs in high linkage disequilibrium (r^2>0.50) with it from the set of SNPs used to build the **G** matrix, we obtained prediction accuracies and bias that were very similar to those described above (Supporting information). The only exception was the MA-GBLUP analysis in the Duroc population. In this population, the prediction accuracy decreased from 0.362 to 0.345 and regression coefficient of the corrected phenotypes on the breeding values decreased from 0.88 to 0.85.

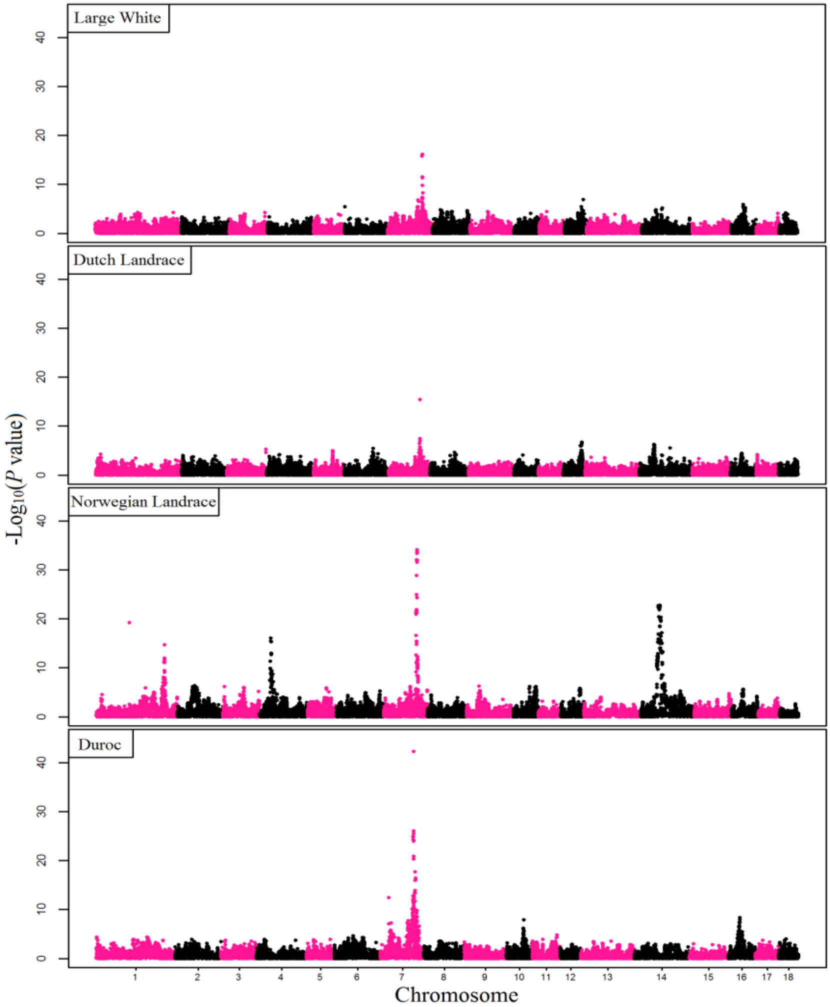

Figure 5.1 GWAS on number of teats in four pig populations. On the y-axis is the -Log$_{10}$(*P* values) of single SNP association with number of teats in pigs. On the x-axis is the physical position of the SNPs across the 18 autosomes.

Table 5.2 Description of the most significant SNP in each population.

Population	**Most. Sig. SNP**	**Chr**	**Pos**	**-Log$_{10}$(*P* value)**	**Allele Freq.**	**Effect**	**Var. explained (%)**
Large White	MARC0038565	7	103.50	16.12	0.30	0.27	3.48
Dutch Landrace	ASGA0035500	7	103.57	15.44	0.69	0.29	3.67
Norwegian Landrace	INRA0027623	7	103.37	34.09	0.71	0.25	3.30
Duroc	MARC0038565	7	103.50	42.26	0.38	0.36	6.13

Chromosome (Chr), position in Mb (Pos), frequency of the allele related to higher number of teats (Allele freq.), allele substitution effect (Effect), percentage of the total phenotypic variance explained by the most significant SNP (Var. explained).

Table 5.3 Accuracy of prediction and regression coefficient.

Population	BLUP	MA-BLUP	GBLUP	MA-GBLUP
		Accuracy [a]		
Large White	0.238	0.266	0.361	0.370
Dutch Landrace	0.199	0.259	0.239	0.271
Norwegian Landrace	0.315	0.336	0.474	0.477
Duroc	0.192	0.316	0.319	0.362
		Bias [b]		
Large White	0.84	0.87	0.97	0.96
Dutch Landrace	0.85	0.92	0.68	0.71
Norwegian Landrace	1.12	0.87	1.10	1.11
Duroc	0.82	1.01	0.80	0.88

[a]*Accuracy*: correlation between the corrected phenotypes and breeding values; [b]*Bias*: regression coefficient of the corrected phenotypes on the breeding values.

5.4 Discussion

In this study, we showed that reviving MAS by accounting for the most significant SNP (identified in GWAS) in the genetic predictions resulted in improved prediction accuracy for the trait "number of teats" in all evaluated populations (Table 5.3). Replacing BLUP by MA-BLUP, increased the prediction accuracy between 0.021 and 0.124, whereas replacing GBLUP by MA-GBLUP, resulted in increases between 0.003 and 0.043. Meuwissen & Goddard (1996) described that the advantage of MAS over non-MAS is related to the proportion of variance explained by the QTL linked to the markers used in the prediction. Changing either from BLUP to MA-BLUP or from GBLUP to MA-GBLUP, the highest increase in prediction accuracy was observed in the Duroc population and the lowest in the Norwegian Landrace. This result is concordant with the total phenotypic variance explained by the SNP used in the predictions (6.13% in the Duroc and 3.30% in the Norwegian Landrace, Table 5.2).

The smaller improvement observed when replacing GBLUP by MA-GBLUP compared to replacing BLUP by MA-BLUP can be explained by the fact that GBLUP already accounts for the Mendelian sampling, which is one of the greatest advantages of genomic selection compared to pedigree-based selection (VanRaden 2008; Lopes *et al.* 2013). Applying MA-GBLUP, however, we account for Mendelian sampling and also for some prior information on SNPs with a large effect which has additional benefits for the prediction accuracy.

Previous applications of MAS in livestock were mainly using QTL identified in linkage studies with experimental crosses between breeds or lines. Thus, the frequency of the identified QTL were often fixed in the pure lines and could not be used for selection within line (Dekkers 2004). In this study, we performed a GWAS within line and showed that the most significant SNP for number of teats in all evaluated populations is located in the same region on chromosome 7 (~103.50 Mb). For the Large White and Dutch Landrace populations, this was expected because this QTL region was reported in previous studies using about 28% (Duijvesteijn *et al.* 2014) and 50% (Lopes *et al.* 2014) of the current data from these populations, respectively. For the Norwegian Landrace and Duroc populations, however, this QTL region has not previously been reported. Identification of this QTL region in all evaluated populations provides additional independent replication of the previous studies (Duijvesteijn *et al.* 2014; Lopes *et al.* 2014) and reinforces its relevance for being used in selection for number of teats in pigs.

Although the QTL region on chromosome 7 was identified in all evaluated populations, the most significant SNP was not the same across populations. MARC0038565 was the most significant SNP in both Large White and Duroc population (Table 5.2). In the Norwegian Landrace, MARC0038565 was the second most significant SNP, being in high LD to INRA0027623, the most significant SNP in this population (r^2= 0.99, Figure 5.2). In the Dutch Landrace, however, the most significant SNP (ASGA0035500) showed no LD with MARC0038565 (r^2= 0). These differences in LD and the fact that we have different SNPs tagging the, presumably, same QTL across populations seems to indicate that the causal mutation is not present in the SNP panels used in this study. Finding the causal mutation would allow moving from marker-assisted (G)BLUP to gene-assisted (G)BLUP, which could lead to even higher accuracies of prediction because complete LD exists between the marker and the QTL (Villanueva *et al.* 2002).

The Norwegian Landrace presented the highest prediction accuracies for all models, whereas the Dutch Landrace population presented the lowest (except for BLUP, Table 5.3). In both simulated and real data (Habier *et al.* 2007; Wu *et al.* 2015), it has been shown that higher relationship between training and validation populations can lead to higher prediction accuracies. As can be observed in Figure 5.3, the highest average of pedigree-based relationships between the TRAINING and the VALIDATION datasets was observed for the Norwegian Landrace (average=0.06) and the lowest for the Dutch Landrace (average=0.03). For the Norwegian Landrace population, pairwise pedigree-based relationship coefficients between the animals from the TRAINING and VALIDATION datasets were all greater than zero. On the other hand, for the Dutch Landrace, a large proportion of the pairwise pedigree-based relationship coefficient was zero. Intermediate prediction accuracies and relationship between training and validation datasets were observed for both Large White and Duroc populations. These differences between populations indicate that the relationship between training and validation populations may indeed have affected the observed accuracies of prediction.

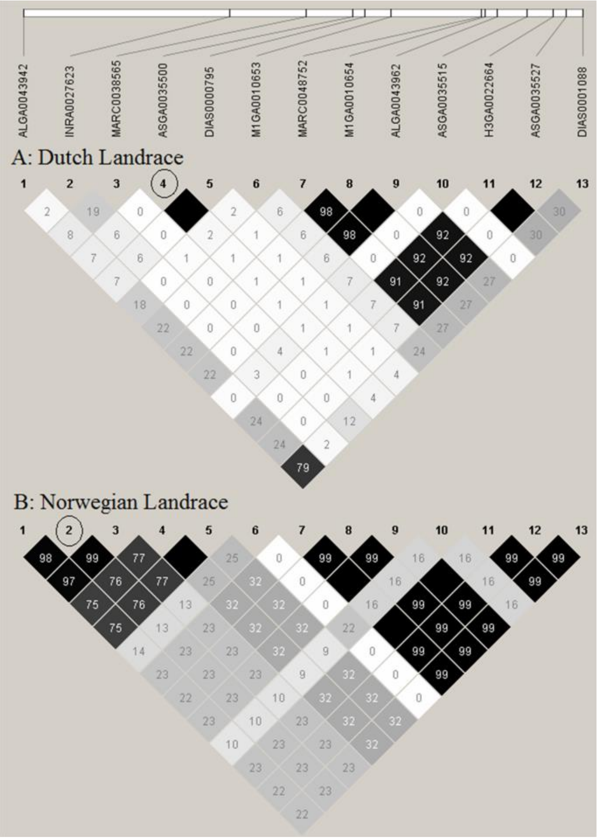

Figure 5.2 Linkage disequilibrium (LD) on chromosome 7. LD (r^2) between SNPs located between 103 Mb and 104 Mb in the Dutch Landrace population (A) and the Norwegian Landrace Population (B). The most significant SNP in each population is marked with a circle. The numbers inside the diamonds are the r^2 values on a scale of 0 to 100%.

Another factor that influences the accuracies of breeding values is the size of the training population (Daetwyler *et al.* 2010). The size of the TRAINING dataset in this study varied considerably across populations, ranging from 2,490 for the Dutch Landrace to 6,090 for the Norwegian Landrace. The population with the highest prediction accuracy (Norwegian Landrace) also had the largest training population. To evaluate the effect of the size of the TRAINING dataset on the value of adding individual QTL in the model, we performed the prediction analysis in a smaller dataset (N=3,000), within the Norwegian Landrace. The 3,000 oldest animals of this population were divided in training (N= 2,400) and validation (N=600) datasets according to their date of birth (validation animals were the 20% youngest animals from the dataset). In this scenario, the prediction accuracy for BLUP, MA-BLUP, GBLUP, and MA-GBLUP were respectively 0.287, 0.339, 0.423, and 0.446. Using the complete data (training on 6,090 animals), the prediction accuracy for BLUP, MA-BLUP, GBLUP, and MA-GBLUP were respectively 0.315, 0.336, 0.474, and 0.477. As expected, the prediction accuracies tended to decrease with the smaller training population. The decrease was bigger for the traditional models (BLUP and GBLUP), indicating that MAS has more added value with smaller training populations. Increases in accuracy were 0.021 (MA-BLUP) and 0.003 (MA-GBLUP) compared to BLUP and GBLUP, respectively. With the reduced dataset, these increases were 0.052 and 0.023, respectively.

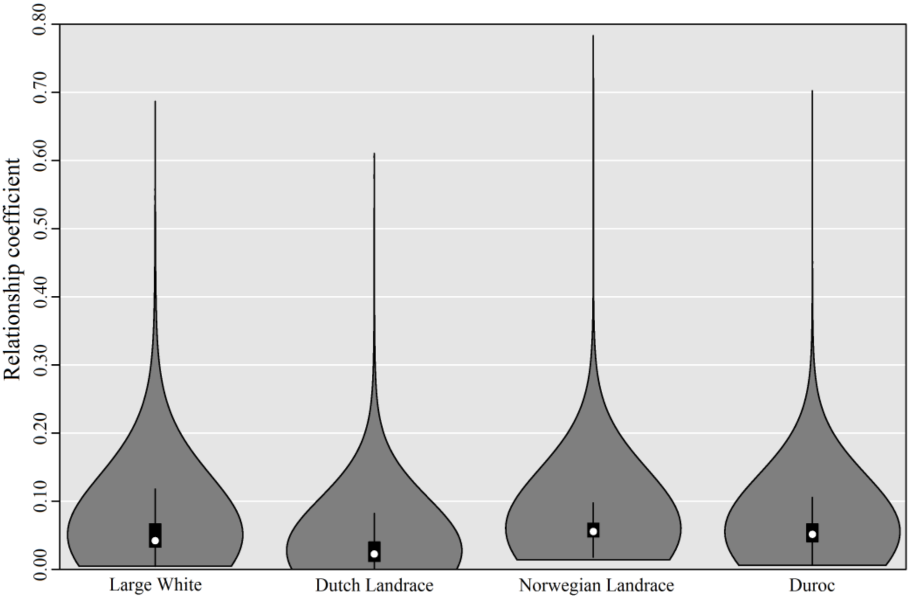

Figure 5.3 Relationship coefficient between animals from TRAINING and VALIDATION dataset. Violin plot (box plot and probability density) of the pedigree-based relationship coefficient between TRAINING and VALIDATION dataset of the four evaluated populations. The median relationship coefficient is indicated with a white dot inside the box plot.

As discussed above and in previous studies (Habier *et al.* 2007; Daetwyler *et al.* 2010; Wu *et al.* 2015), the accuracies of breeding values seem to be influenced by the relationships between validation and training populations, and the size of the reference population. However, the estimation of the SNP effects seems to be less affected by these two factors. The correlation between the marker breeding value and the corrected phenotype of the VALIDATION dataset was 0.132 in the Large White, 0.150 in the Dutch Landrace, 0.175 in the Norwegian Landrace, and 0.260 in the Duroc (data not shown). These values seem to correlate with the amount of phenotypic variance explained by the marker (~3.48% in the dam lines and 6.13% in the Duroc) and not with the relationships between training and validation or the size of the reference population.

As previously discussed, and as expected (Sato *et al.* 2006; Guo *et al.* 2008; Ding *et al.* 2009; Duijvesteijn *et al.* 2014; Lopes *et al.* 2014), the QTL region on chromosome 7 was the most significant region for number of teats in all populations. We choose this QTL region for our analyses because of its known effect. However, in the Norwegian Landrace, highly significant peaks [-log$_{10}$(*P* value) >10] were also observed on chromosomes 1, 4, and 14 (Figure 5.1). The most significant SNPs in each of these regions explains >1% of the phenotypic variance (Supporting information). Accounting only for the most significant SNP from chromosome 7, the prediction accuracies of MA-BLUP and MA-GBLUP were 0.336 and 0.477, respectively. Additionally accounting for the most significant SNP from chromosome 14, the prediction accuracies became 0.372 and 0.474, respectively (Table 5.4). Further additions of the most significant SNPs from the peaks on chromosomes 1 and 4 accuracies became 0.399 and 0.482, respectively. Thus, including a marker from each of these other three QTL regions in Norwegian Landrace increased the prediction accuracy of MA-BLUP and MA-GBLUP by 0.063 and 0.005, above the effect of the marker from chromosome 7. Including all markers above our threshold of explaining >1% of the phenotypic variance lead to increased prediction accuracy. However, this threshold was arbitrary and the marker selection strategy needs further evaluation.

When using only the marker breeding value ($\hat{u}_{\text{snp}} = \sum_{m=1}^{4} \beta_m X_m$) based on the four markers described above, the prediction accuracy for the Norwegian Landrace was 0.302 (using SNP effects estimated in MA-BLUP, Table 5.4). This accuracy is similar to the polygenic breeding value accuracy from BLUP (0.315), indicating that for number of teats in this population, four markers have almost the same prediction ability as pedigree. With such accuracies, marker breeding values can be an important tool for selection of animals that have no genotypes nor pedigree information, which is often the case in crossbred sows.

Table 5.4 Accuracy of prediction in the Norwegian Landrace population using multiple QTL regions.

QTL regions Included*	MA-BLUP			MA-GBLUP		
	$\hat{u}_{\mathrm{g}}$	$\hat{u}_{\mathrm{snp}}$	$\hat{u}$	$\hat{u}_{\mathrm{g}}$	$\hat{u}_{\mathrm{snp}}$	$\hat{u}$
**	0.315	-	0.315	0.474	-	0.474
7	0.296	0.175	0.336	0.453	0.173	0.477
7, 14	0.302	0.245	0.372	0.423	0.245	0.474
7, 14, 4	0.296	0.293	0.392	0.409	0.296	0.479
7, 14, 4, 1	0.291	0.302	0.399	0.400	0.306	0.482

*The chromosome number(s) from which the most significant SNP from the most pronounced peaks was selected to be included in the prediction analysis. **No SNPs were used, therefore, it correspond to traditional BLUP and GBLUP. $\hat{u}_{\mathrm{g}}$: polygenic breeding value; $\hat{u}_{\mathrm{snp}}$: marker breeding value ($\sum_{m=1}^{M} \beta_m \mathrm{X}_m$); and $\hat{u}$: total breeding value ($\hat{u}_{\mathrm{g}}$ + $\hat{u}_{\mathrm{snp}}$).

An alternative approach to using GWAS results in genetic predictions was described by Zhang *et al.* (2010). With this approach, the traditional **G** matrix in GBLUP is replaced by a trait-specific **G** matrix that gives different weights to each SNP. This approach favours (gives more weight to) SNPs that contribute more to the genetic variance of the evaluated trait. In a traditional genomic relationship matrix (**G** matrix), all SNPs are expected to contribute equally (they have the same weights). Zhang *et al.* (2010) showed that the breeding values from the model that applies the trait-specific **G** matrix were more accurate, but also more biased than the breeding values from both BLUP and GBLUP. The practical application of this approach is troublesome because the **G** matrix is trait-specific and would, therefore, require single-trait genetic evaluations. However, breeding programs, in general, apply multi-trait genetic evaluation to capitalize on the genetic correlations between traits.

Using the MA-BLUP or MA-GBLUP, GWAS results could be incorporated in the genetic prediction using both single-trait and multi-trait genetic evaluation, as the same **G** matrix could be used for all traits of interest. For the marker-assisted models evaluated in this study, modification on the **G** matrix would also be required if the double use of the SNP as both a fixed effect and a contributor to the **G** matrix would give problems. If double counting of effects would occur, the SNPs used in the prediction models should not be used to build the **G** matrix. However, as shown in the Supporting information, removing the SNPs in LD with the QTL from the **G** matrix resulted in similar prediction results, indicating no problems due to double counting.

Another alternative for the use of GWAS results in genetic predictions was described by Boichard *et al.* (2012). These authors showed that including a random effect of haplotypes in significant regions from GWAS was more accurate than traditional BLUP and GBLUP. The marker-assisted models presented in the current study are, however, more straightforward than the haplotype approach because phasing of haplotypes is not required.

The increasing amount of research aiming to develop models and methods for genomic selection is showing the potential of this relatively novel breeding tool (Jonas & de Koning 2015). With this study we showed that with improved technologies, such as dense SNP panels, we can also revive "old" models and methods, such as MAS, to improve the accuracy of prediction. We found prediction to be improved using the marker-assisted models compared to BLUP and GBLUP models. These marker-assisted models should be considered for prediction of breeding values in traits with well-defined QTL regions. In pigs, these traits would include, for example, androstenone level (Duijvesteijn *et al.* 2010; Hidalgo *et al.* 2014), and host response to porcine reproductive and respiratory syndrome virus challenge (Boddicker *et al.* 2012). In dairy cattle, traits affected by the *DGAT1* region (Jiang *et al.* 2010; Bouwman *et al.* 2012) would be good candidates. However, not all traits have QTL regions that explain a substantial proportion of the phenotypic variance. For those traits, the application of traditional GBLUP remains a good method to use genomic data to obtain higher prediction accuracies.

5.5 Conclusions

For number of teats, BLUP resulted in the lowest prediction accuracy and MA-GBLUP in the highest for four distinct populations. In the same dataset, MA-BLUP can yield similar or superior accuracies compared to GBLUP. The superiority of MA-GBLUP over traditional GBLUP is more pronounced when training populations are smaller and when relationships between training and validation populations are smaller.

5.6 Acknowledgement

This work was financed by the Research Council of Norway and the Breed4Food Consortium. The authors thank Cigene (NMBU, Norway) and Hanne Hamland in Norsvin for sample handling and genotyping.

5.7 References

Boddicker, N., Waide, E., Rowland, R., Lunney, J., Garrick, D., Reecy, J., Dekkers, J. (2012) Evidence for a major QTL associated with host response to porcine reproductive and respiratory syndrome virus challenge. *J. Anim. Sci.,* **90,** 1733-1746.

Boichard, D., Fritz, S., Rossignol, M.-N., Boscher, M.Y., Malafosse, A., Colleau, J.J. (2002) Implementation of marker-assisted selection in French dairy cattle. *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France, August, 2002. Session 22.*, pp. 1-4. Institut National de la Recherche Agronomique (INRA).

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M.-N., Boscher, M.Y., Druet, T., Genestout, L., Colleau, J.-J., Journaux, L. (2012) Genomic selection in French dairy cattle. *Anim. Prod. Sci.,* **52,** 115-120.

Bouwman, A.C., Visker, M.H., van Arendonk, J.A., Bovenhuis, H. (2012) Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. *BMC Genet.,* **13,** 93.

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

Browning, S.R., Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.,* **81,** 1084-1097.

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

Dekkers, J.C. (2004) Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. *J. Anim. Sci.,* **82,** E313-E328.

Ding, N., Guo, Y., Knorr, C., Ma, J., Mao, H., Lan, L., Xiao, S., Ai, H., Haley, C., Brenig, B. (2009) Genome-wide QTL mapping for three traits related to teat number in a White Duroc × Erhualian pig resource population. *BMC Genet.,* **10,** 6.

Duijvesteijn, N., Knol, E.F., Merks, J.W.M., Crooijmans, R.P.M.A., Groenen, M.A.M., Bovenhuis, H., Harlizius, B. (2010) A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. *BMC Genet.,* **11,** 42.

Duijvesteijn, N., Veltmaat, J.M., Knol, E.F., Harlizius, B. (2014) High-resolution association mapping of number of teats in pigs reveals regions controlling vertebral development. *BMC Genomics,* **15,** 542.

Fernando, R., Grossman, M. (1989) Marker assisted selection using best linear unbiased prediction. *Genet. Sel. Evol.,* **21,** 467.

Gilmour, A.R., Gogel, B., Cullis, B., Thompson, R. (2009) ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

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

Groenen, M.A., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature,* **491,** 393-398.

Guo, Y.M., Lee, G., Archibald, A., Haley, C. (2008) Quantitative trait loci for production traits in pigs: a combined analysis of two Meishan × Large White populations. *Anim. Genet.,* **39,** 486-495.

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

Hayes, B., Goddard, M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.,* **33,** 209-230.

Heffner, E.L., Sorrells, M.E., Jannink, J.-L. (2009) Genomic selection for crop improvement. *Crop. Sci.,* **49,** 1-12.

Hidalgo, A.M., Bastiaansen, J.W., Harlizius, B., Megens, H.-J., Madsen, O., Crooijmans, R.P., Groenen, M.A. (2014) On the relationship between an Asian haplotype on chromosome 6 that reduces androstenone levels in boars and the differential expression of SULT2A1 in the testis. *BMC Genet.,* **15,** 4.

Jiang, L., Liu, J., Sun, D., Ma, P., Ding, X., Yu, Y., Zhang, Q. (2010) Genome wide association studies for milk production traits in Chinese Holstein population. *PLoS One,* **5,** e13661.

Jonas, E., de Koning, D.-J. (2015) Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. *Front. Genet.,* **6**.

Lande, R., Thompson, R. (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics,* **124,** 743-756.

Lopes, M.S., Bastiaansen, J.W., Harlizius, B., Knol, E.F., Bovenhuis, H. (2014) A genome-wide association study reveals dominance effects on number of teats in pigs. *PLoS One,* **9,** e105867.

Lopes, M.S., Silva, F.F., Harlizius, B., Duijvesteijn, N., Lopes, P.S., Guimarães, S.E., Knol, E.F. (2013) Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genet.,* **14,** 92.

Meuwissen, T., Goddard, M. (1996) The use of marker haplotypes in animal breeding schemes. *Genet. Sel. Evol.,* **28,** 161-176.

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

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

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.,* **81,** 559-575.

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

Sato, S., Atsuji, K., Saito, N., Okitsu, M., Komatsuda, A., Mitsuhashi, T., Nirasawa, K., Hayashi, T., Sugimoto, Y., Kobayashi, E. (2006) Identification of quantitative trait loci affecting corpora lutea and number of teats in a Meishan × Duroc F2 resource population. *J. Anim. Sci.,* **84,** 2895-2901.

Silva, M.V., dos Santos, D.J., Boison, S.A., Utsunomiya, A.T., Carmo, A.S., Sonstegard, T.S., Cole, J.B., Van Tassell, C.P. (2014) The development of genomics applied to dairy breeding. *Livest. Sci.,* **166,** 66-75.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature methods,* **5,** 247-252.

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

Villanueva, B., Pong-Wong, R., Woolliams, J.A. (2002) Marker assisted selection with optimised contributions of the candidates to selection. *Genet. Sel. Evol.,* **34,** 679-704.

Wu, X., Lund, M., Sun, D., Zhang, Q., Su, G. (2015) Impact of relationships between test and training animals and among training animals on reliability of genomic prediction. *J. Anim. Breed. Genet.,* **132,** 366–375.

Zhang, Z., Liu, J., Ding, X., Bijma, P., de Koning, D.-J., Zhang, Q. (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One,* **5,** e12648.

5.8 Supporting information

Table S5.1 SNPs with linkage disequilibrium (r^2) >0.50 with the most significant SNP of each population. All SNPs are located on chromosome 7.

Population	Most significant SNP		Linked SNP		r^2
	SNP	Pos	SNP	Pos	
Large White	MARC0038565	103.50	H3GA0022644	102.90	0.56
Dutch Landrace	ASGA0035500	103.57	DIAS0000795	103.59	1.00
Norwegian Landrace	INRA0027623	103.37	H3GA0022589	100.26	0.60
			CADI0000203	100.79	0.67
			ALGA0043854	100.84	0.67
			ALGA0043856	100.87	0.67
			ALGA0043860	100.93	0.67
			DRGA0008010	100.96	0.67
			H3GA0022595	101.00	0.68
			MARC0070023	101.10	0.68
			INRA0027559	101.15	0.68
			MARC0050498	101.18	0.68
			DRGA0008014	101.27	0.70
			INRA0027569	101.33	0.70
			DIAS0002786	101.40	0.70
			MARC0106494	101.80	0.89
			ALGA0043906	101.97	0.64
			INRA0027600	102.22	0.95
			ALGA0043913	102.38	0.95
			DIAS0000908	102.43	0.95
			MARC0113727	102.46	0.96
			INRA0027605	102.74	0.95
			H3GA0022644	102.90	0.98
			ALGA0043941	103.00	0.98
			ALGA0043942	103.02	0.98
			MARC0038565	103.50	0.99
			ASGA0035500	103.57	0.77
			DIAS0000795	103.59	0.77
Duroc	MARC0038565	103.50	INRA0027622	103.10	0.58
			ASGA0035500	103.57	0.61
			DIAS0000795	103.59	0.61
			H3GA0022659	103.72	0.55
			ASGA0035515	103.87	0.59
			H3GA0022664	103.91	0.60
			ASGA0035527	103.93	0.59
			DIAS0001088	103.96	0.59

Table S5.2 Prediction accuracy using a **G** matrix built using all markers (*all-in*), and excluding the SNP used a fixed effect in the model and the linked (r^2> 0.50) SNPs (*LD-out*).

Population	MA-GBLUP	
	all-in	*LD-out*
Accuracy [a]		
Large White	0.370	0.371
Dutch Landrace	0.271	0.271
Norwegian Landrace	0.477	0.472
Duroc	0.362	0.345
Bias [b]		
Large White	0.96	0.99
Dutch Landrace	0.71	0.71
Norwegian Landrace	1.11	1.12
Duroc	0.88	0.85

[a] *Accuracy*: correlation between the corrected phenotypes and breeding values; [b] *Bias*: regression coefficient of the corrected phenotypes on the breeding values.

Table S5.3 Description of the most significant SNP in the most pronounced QTL peaks in the Norwegian Landrace population.

Most. Sig. SNP	Chr	Pos	-Log$_{10}$(*P* value)	Freq.	Effect	Var. expl. (%)
INRA0027623	7	103.37	34.09	0.71	0.25	3.30
ALGA0077463	14	51.01	22.78	0.52	0.18	2.15
INRA0013436	4	31.64	16.06	0.46	0.15	1.53
ASGA0006373	1	268.77	14.63	0.27	0.16	1.28

Chromosome (Chr), position in Mb (Pos), frequency of the allele related to higher number of teats (Freq.), allele substitution effect (Effect), percentage of the total phenotypic variance explained by the most significant SNP (Var. expl.).

6

Genomic selection for crossbred performance accounting for breed-specific effects

Marcos S Lopes[1,2], Henk Bovenhuis[2], André M Hidalgo[2], Johan AM van Arendonk[2], Egbert F Knol[1], John WM Bastiaansen[2]

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

In preparation

Abstract

Breed-specific effects are observed when the same allele of a given marker has a different effect depending on its breed origin, resulting in different allele substitution effects across breeds. When this is the case, single-breed breeding values may not be efficient to predict crossbred performance. Our aim was to estimate the contribution of each parental breed to the genetic variance of traits measured in their crossbred offspring, and to compare the prediction accuracies of breeding values from a traditional genomic selection model (GS), trained on purebred or crossbred data, with accuracies of breeding values from a model that accounts for breed-specific effects (BS), trained on crossbred data. Traits evaluated were litter size (LS) and gestation length (GL) of pigs. The genetic correlation between purebred and crossbred performance for both LS and GL was 0.90. For both traits, the additive genetic variance was higher for alleles inherited from the Large White (LW) compared to alleles inherited from the Landrace (LR) breed (0.74 and 0.56 for LS, and 0.42 and 0.40 for GL, respectively). Highest accuracies of predicting performance of crossbred sows were observed when training on crossbred data. For LS, prediction accuracies were the same for GS and BS breeding values (0.23), while for GL, prediction accuracy for BS was slightly greater than accuracy of GS breeding values (0.53 and 0.52, respectively). Evidence of breed-specific effects for LS and GL were demonstrated. In future studies, traits with lower genetic correlation between purebred and crossbred performance should be evaluated to confirm the potential use of BS models in genomic predictions.

Key words: SNP, epistasis, breeding value, prediction accuracy

6.1 Introduction

In pig breeding, selection takes place in purebred lines and genetic evaluations have been performed mainly using information collected at the purebred level, in high-health environments, even though the final product of the pig industry is a crossbred animal. This strategy may not be optimal when the objective is to improve crossbred performance. The genetic progress realized at the purebred level may not fully translate to improved crossbred performance (under field conditions) when the genetic correlation between purebred and crossbred performance is less than 1 (Dekkers 2007; Hidalgo *et al.* 2015). Low genetic correlations between purebred and crossbred performance have been observed for many production traits (Zumbach *et al.* 2007; Cecchinato *et al.* 2010) and can be caused by genotype-by-environment interaction, non-additive effects (such as dominance) or breed-specific effects. Therefore, if the goal is to improve crossbred performance by purebred selection, effects that influence the genetic correlation between purebred and crossbred performance should be evaluated. In addition, the use of crossbred data in genetic evaluations should be considered (Wei & Van der Steen 1991; Dekkers 2007; Ibañez-Escriche *et al.* 2009; Esfandyari *et al.* 2015a; Van Grevenhof & Van der Werf 2015).

Using either real or simulated data, several studies have investigated the relevance of genotype-by-environment interactions and dominance effects for pig breeding (Knap & Su 2008; Silva *et al.* 2014; Esfandyari *et al.* 2015b; Lopes *et al.* 2015). Breed-specific effects, however, have not yet been studied extensively. Breed-specific effects are observed when the same allele, say allele A, of a given marker has a different effect on the crossbred phenotype depending on its breed origin. Breed-specific marker effects may occur when the linkage disequilibrium (LD) between the markers and the quantitative trait loci (QTL) differ between breeds or when the allele frequencies of the QTL vary across breeds (Ibañez-Escriche *et al.* 2009). When breed-specific effects are present, allele substitution effects will vary across breeds, and therefore, the single-breed breeding values, estimated using purebred data, may not accurately predict crossbred performance.

With the recent availability of high-density marker genotypes on both purebred and crossbred animals, we can now include crossbred data in the genomic evaluations. Breed origin of alleles in crossbreds can also be determined and used to build breed-specific relationship matrices (Christensen *et al.* 2014). Replacing the genomic relationship of a traditional genomic selection model (GS model) by the breed-specific relationship matrices (BS model), allows us to quantify the contributions of each parental breed to the total genetic variance of the trait in crossbreds. In addition, we can also estimate breed-specific breeding values that can be backsolved to breed-specific marker effects (Wang *et al.* 2012). Breed-specific marker effects could then be used to predict breeding values of purebred animals for crossbred performance. Doing so, we would benefit from training on crossbred field data using only the alleles inherited from the purebred population where selection takes place.

In simulation studies, Ibañez-Escriche *et al.* (2009) concluded that the BS models may not be required to effectively select purebreds for crossbred performance, while Esfandyari *et al.* (2015a) concluded that if the size of the training population is sufficiently large and the parental breeds are not very closely related, accounting for the breed origin of alleles can improve accuracy of genomic prediction. However, studies using real data are still necessary to determine the relevance of breed-specific effects for genomic selection. In this study, we investigated the value of breed-specific effects for predicting crossbred performance using real data. First, the contribution of each parental breed to the genetic variance was quantified for traits measured in a two-way crossbred population. Second, prediction accuracies were estimated with the GS model with either purebred or crossbred training data and with the BS model with crossbred training data.

6.2 Material and methods

6.2.1 Data

Phenotypic and genotypic data were available on pigs from two purebred populations: Large White (LW) and Landrace (LR); and a two-way crossbred population (F1) consisting of animals produced by reciprocal crosses of the purebred populations (LW♂ x LR♀ and LR♂ x LW♀). Phenotypic data consisted of litter size (LS, sum of piglets born alive and stillborn in the same litter) and gestation length (GL, number of days between insemination and farrowing). Both traits were recorded from parities 2 to 7. Records from the first parity were excluded because LS and GL measured in the first or in later parities, have been described as different traits due to their low genetic correlation (Irgang *et al.* 1994; Hanenberg *et al.* 2001).

Phenotypic data on both traits were available for 22,597 LW, 27,035 LR, and 29,847 F1 animals (Table 6.1). The F1 population consisted of 14,964 animals from the LW♂ x LR♀ cross, and 14,883 animals from the LR♂ x LW♀ cross. On average, data from 3.8, 3.6, and 2.5 parities per animal were available in the LW, LR, and F1 populations, respectively. Data from the purebred populations were recorded on genotyped animals (3,723 LW and 3,291 LR) and their non-genotyped contemporaries (e.g. animals from the same breed and farm as the genotyped animals; 18,874 LW and 23,744 LR). The purebred animals were located on 18 (LW) and 20 (LR) farms and were born between 2004 and 2014. Data from the F1 population was recorded on 1,126 genotyped animals and their 1,120 non-genotyped contemporaries. These genotyped F1 animals and their contemporaries were located on 6 farms. Finally, data were also recorded on 27,601 non-genotyped F1 offspring of the genotyped purebred animals. This additional group of F1 animals was located across 111 farms and was only used to increase the size of the crossbred population to estimate the genetic correlation between purebred and crossbred performance. All F1 animals were born between 2010 and 2014 and were.

Phenotypes of genotyped animals were pre-adjusted for fixed effects using the larger data set, i.e. including contemporaries. Therefore, fixed effects were accounted more accurately. Fixed effects were estimated by fitting a single trait, pedigree-based linear model for each population, using ASReml v3.0 (Gilmour *et al.* 2009). The model used for LS included fixed effects: parity, interval between weaning and pregnancy (days), whether more than one insemination was performed (yes or no), litter type (purebred or crossbred), and herd-year-season; and random effects: service sire, permanent environmental effect, and the additive genetic effect. The model used for GL included fixed effects: parity, whether more than one insemination procedure was performed (yes or no), litter type (purebred or crossbred), herd-year-season, and the covariate LS. The random effects of the model for GL were the same as for LS. The fixed effect litter type was not included in the model for both traits when evaluating the performance of the F1 animals.

6.2.2 Genetic correlation between purebred and crossbred performance

Genetic correlations between purebred and crossbred performance for both traits were estimated using all available phenotypic data in a pedigree-based bivariate analyses in ASReml 3.0 (Gilmour *et al.* 2009). Purebred and crossbred performance were analysed as different traits. Models for estimating the genetic correlation were the same as for the pre-adjustment of the phenotypes, except that a fixed effect for breed composition (LW, LR or F1) was added.

Table 6.1 Number of animals and phenotypic records from each population.

Population [1]	Phenotypes [2]		Genotypes [3]	Genotypes and phenotypes [4]	
	Animals	Records	Animals	Animals	Records
LW	22,597	84,837	3,723	924	3,358
LR	27,035	96,431	3,291	924	3,319
F1	29,847	75,143	1,126	924	3,771

[1] Large White (LW), Landrace (LR), and two-way crossbred (F1). [2] Number of animals with phenotypic information and total number of phenotypic records of these animals. [3] Number of genotyped animals used for imputation and phasing procedures. [4] Number of genotyped animals and the number of phenotypic records of these animals used for estimating the variance components and SNP effects.

6.2.3 Genotyping

Genotypic data were available on 3,723 LW, 3,291 LR, and 1,126 F1 animals (Table 6.1). In the purebred populations, both males and females were genotyped. In the F1 population, only females were genotyped. Genotyping was performed mainly using the Illumina Porcine SNP60 Beadchip, but part of the animals from all populations was genotyped using the Illumina Porcine SNP60 v2 Beadchip. All animals were imputed to have genotypes for all the SNPs on the SNP60 Beadchip that passed the quality control. Quality control consisted of excluding SNPs with GenCall<0.15, call rate <0.95, minor allele frequency <0.01, and strong deviation from Hardy-Weinberg equilibrium (χ^2>600). The SNPs located on sex chromosomes and unmapped SNPs were also excluded. Positions of the SNPs were based on the Sscrofa10.2 assembly of the reference genome (Groenen *et al.* 2012). All genotyped animals had a frequency of missing genotypes above the threshold of 0.05 for excluding poorly-genotyped animals. After quality control and imputation, 39,788 SNPs for LW, 41,299 SNPs for LR, and 45,515 SNPs for F1 were available for further analyses.

6.2.4 Imputation and phasing of the data

Imputation and phasing of the data were performed using AlphaImpute (Hickey *et al.* 2011), combining genomic and pedigree information to determine the parental origin of alleles. Imputation of missing genotypes of the purebred populations was performed within population using all SNPs that passed the quality control. For the F1 population, the imputation of missing genotypes and phasing of the data was performed combining the F1 data with the imputed purebred data. The latter analysis was performed using only the 36,733 SNPs that segregated (minor allele frequency >0.01) in all populations.

To ensure the use of accurately phased haplotypes to determine the breed origin of alleles, a threshold was applied to F1 phased data. For each SNP genotype of each individual, AlphaImpute (Hickey *et al.* 2011) generates two probabilities: P_1 is the probability that a specific allele was received from the father, say the allele G of a G/C genotype, and P_2 is the probability that the same allele was received from the mother. Considering a heterozygous animal (CG) where the C allele was inherited with certainty from the father (and therefore a G allele from the mother), the probabilities would be P_1= 0 and P_2= 1. When the phasing cannot be performed with certainty, these probabilities will have values between 0 and 1. Values of P_1 or P_2 between 0.1 and 0.9 were considered indicative of poor phasing. SNPs that were considered poorly-phased in >95% of animals were excluded from the dataset. After this, animals that had >5% poorly-phased genotypes were excluded. After this quality control, 924 F1 animals with genotypes for 31,930 SNPs were available for estimation of variance components and SNP effects. The same set of SNPs was also used for estimation of variance components and SNP effects for the purebred populations.

After phasing of the data, the breed origin of alleles was easily determined because the breeds of the F1's parents were known. The final F1 population was formed by 414 animals from the LW♂ x LR♀ cross, and 510 animals from the LR♂ x LW♀ cross.

6.2.5 Estimation of variance and SNP effects

The number of animals with both phenotypes and genotypes was larger in the purebred than in the crossbred population. As the size of the training population influences the estimation of SNP effects and consequently the prediction accuracy (Daetwyler *et al.* 2010), we randomly selected 924 animals born between 2010 and 2014 from each purebred population to be the training population. With this, we attempted to have a fair comparison, as the size of the training population and range of birth years were the same for all populations (Table 6.1). In order to have independent datasets for validation analysis (discussed below), the purebred animals used as training populations had no offspring or sibs in the F1 population.

Variance components and SNP effects were estimated within population using a traditional genomic selection model (GS model) and a model that accounts for breed-specific effects (BS model). The GS model was applied to both purebred and crossbred data while the BS model was applied only to the crossbred data. These models were implemented in ASReml (Gilmour *et al.* 2009), as follows:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ss} + \mathbf{Pp} + \mathbf{Zu}_{GS} + \mathbf{e} \qquad \text{(GS model)}$$
$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Ss} + \mathbf{Pp} + \mathbf{Z}_{LW}\mathbf{u}_{BS|LW} + \mathbf{Z}_{LR}\mathbf{u}_{BS|LR} + \mathbf{e} \qquad \text{(BS model)}$$

where **y** is a vector of phenotypes pre-adjusted for fixed effects; μ is the mean of the populations and **1** a vector of ones; **S** is the design matrix for the service sire effects; **s** is an unknown vector of service sire effects; **P** is the design matrix for the permanent environmental effects; **p** is an unknown vector of permanent effects; **Z**, **Z**$_{LW}$ and **Z**$_{LR}$ are design matrices for the additive genetic effects; $\mathbf{u}_{GS}$ is an unknown vector of additive effect (i.e. breeding values); $\mathbf{u}_{BS|LW}$ and $\mathbf{u}_{BS|LR}$are unknown vectors of breed-specific additive genetic effects (i.e. breed-specific breeding values). Assumed distributions were **s** ~N(**0**,**I**σ_s^2), **p** ~N(**0**,**R**σ_r^2), $\mathbf{u}_{GS}$ ~N(**0**,**G**$\sigma^2_{a_{GS}}$), $\mathbf{u}_{BS|F1_{LW}}$ ~ N(**0**,**B**$_{\text{LW}}\sigma^2_{a_{BS|F1_{LW}}}$), and $\mathbf{u}_{BS|F1_{LR}}$ ~N(**0**,**B**$_{\text{LR}}\sigma^2_{a_{BS|F1_{LR}}}$), where **I** is an identity matrix, σ_{s}^2 is the service sire variance, **R** is a diagonal matrix with the number of observations per sow on the diagonal, σ_r^2 is the permanent environmental variance, **G** is the traditional genomic additive relationship matrix, $\sigma^2_{a_{GS}}$ is the additive variance, **B**$_{\text{LW}}$ and **B**$_{\text{LR}}$ are the breed-specific genomic relationship matrices, and $\sigma^2_{a_{BS|F1_{LW}}}$ and $\sigma^2_{a_{BS|F1_{LR}}}$ are the breed-specific additive variances. The heritability was defined as $\sigma^2_{a_{GS}}/\sigma^2_{\text{P}}$ for the GS model, and $(\sigma^2_{a_{BS|F1_{LW}}} + \sigma^2_{a_{BS|F1_{LR}}})/\sigma^2_{\text{P}}$ for the BS model, where σ^2_{P} is the total phenotypic variance (sum of all variance estimates from each model). The **G** matrix was built according to VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{MM}'}{2\sum_{i=1}^{n} p_i q_i}$$

where p and q are the allele frequencies, and **M** is a matrix of centered genotypes (0-2p, 1-2p, 2-2p). The **B**$_{\text{LR}}$ and **B**$_{\text{LW}}$ matrices were built according to the genomic gametic relationship matrices described by Nishio and Satoh (2015):

$$\mathbf{B} = \frac{\mathbf{LL}'}{\sum_{i=1}^{n} p_i^* q_i^*}$$

where p^* and q^* are the frequencies of the alleles from either LW (**B**$_{\text{LW}}$) or LR (**B**$_{\text{LR}}$) in the F1 population, and **L** is matrix of centered alleles (0-p^*, 1-p^*) from either breed LW (**B**$_{\text{LW}}$) or LR (**B**$_{\text{LR}}$). After obtaining the estimated breeding values from the GS model ($\hat{\mathbf{u}}_{GS}$) and BS model ($\hat{\mathbf{u}}_{BS|F1_{LW}}$ and $\hat{\mathbf{u}}_{BS|F1_{LR}}$), we backsolved these breeding values to obtain SNP effects which were used to estimate the breeding values of the validation animals. Backsolving of breeding values from the GS models ($\widehat{\boldsymbol{a}}_{GS}$) was performed as described by Wang *et al.* (2012):

$$\widehat{\boldsymbol{a}}_{GS} = \frac{\mathbf{M}'\mathbf{G}^{-1}\widehat{\boldsymbol{u}}_{GS}}{2\sum_{i=1}^{n} p_i q_i}$$

and the per-SNP variance from the GS model was estimated as $\hat{\sigma}^2_{a_{GS}} = 2pq\hat{a}^2_{GS}$. Backsolving of breeding values from the BS model ($\widehat{\boldsymbol{a}}_{BS|F1_{LW}}$ and $\widehat{\boldsymbol{a}}_{BS|F1_{LR}}$) was performed as an extension of the method described by Wang *et al.* (2012):

$$\widehat{\boldsymbol{a}}_{BS} = \frac{\mathbf{L}'\mathbf{B}^{-1}\widehat{\boldsymbol{u}}_{BS}}{\sum_{i=1}^{n} p_i^* q_i^*}$$

and the per-SNP variance from the BS model was estimated as $\hat{\sigma}^2_{a_{BS}} = p^*q^*\hat{a}^2_{BS}$.

6.2.6 Predicting crossbred performance

Individual performance of genotyped crossbred sows was predicted with the SNP effects estimated by the GS model and the BS model. A 40-fold cross-validation was performed to evaluate prediction accuracies. For each replicate 10% of the genotyped F1 animals (N= 92) were randomly assigned to the validation population and the other 90% (N= 832) were assigned to the F1 training population. The purebred training populations were the same for each replicate. Within each replicate, three different traditional genomic (GS) breeding values of validation animals were estimated as: $\widehat{\boldsymbol{u}}_{GS|j} = \mathbf{M}_{F1}\widehat{\boldsymbol{a}}_{GS|j}$, where $\mathbf{M}_{F1}$ is a matrix of centered genotypes of the F1 validation animals and $\widehat{\boldsymbol{a}}_{GS|j}$ is a vector of SNP effects estimated using the GS model on the training animals. The subscript *j* indicates the breed of the animals included in the three different training populations (LW, LR or F1). Across the replicates, values of $\widehat{\boldsymbol{a}}_{GS|F1}$ will vary while the values of $\widehat{\boldsymbol{a}}_{GS|LW}$ and $\widehat{\boldsymbol{a}}_{GS|LR}$ will be the same as the LW and LR training populations were always the same. Also within each replicate, two groups of breed-specific genomic (BS) breeding values of the validation animals were estimated. The first group of BS breeding values used SNP effects estimated within the parental purebred populations as: $\widehat{\boldsymbol{u}}_{BS|j} = \mathbf{L}_{F1_j}\widehat{\boldsymbol{a}}_{GS|j}$, where $\mathbf{L}_{F1_j}$ is a matrix of centered alleles that the F1 validation animals inherited from the *j*th parental purebred populations (LW or LR) and $\widehat{\boldsymbol{a}}_{GS|j}$ is as defined above. So, separate BS breeding values were estimated, one for each parental purebred population. In addition, total BS breeding values ($\widehat{\boldsymbol{u}}_{BS|LW} + \widehat{\boldsymbol{u}}_{BS|LR}$) were calculated for the validation animals.

The second group of BS breeding values does the same as the first, except that this group used SNP effects estimated within the crossbred population as: $\hat{\boldsymbol{u}}_{BS|F1_j} = \mathbf{L}_{F1_j}\hat{\boldsymbol{a}}_{BS|F1_j}$, where $\hat{\boldsymbol{a}}_{BS|F1j}$ is a vector of SNP effects estimated using the BS model on the alleles that the F1 training animals inherited from the j^{th} parental purebred populations, and $\mathbf{L}_{F1_j}$ is as defined above. Also here, total BS breeding values ($\hat{\boldsymbol{u}}_{BS|F1_{LW}} + \hat{\boldsymbol{u}}_{BS|F1_{LR}}$) were calculated for the validation animals. Prediction accuracy was defined as the correlation of the GS breeding values, the BS breeding values, or the total BS breeding values with the pre-adjusted phenotypes in the validation population. Prediction accuracies presented are the averages over 40-fold cross-validation replicates.

6.3 Results

The pedigree-based estimate of the genetic correlation between purebred and crossbred performance was the same for both traits (0.90). The pedigree-based heritability for LS was 0.16 based on the purebred performance and 0.15 based on crossbred performance. The pedigree-based heritability for GL was 0.39 based on the purebred performance and 0.37 based on crossbred performance.

Using the GS model, the heritability for LS was 0.15 for the LW population and 0.12 for the LR population (Table 6.2). For the F1 population, the heritability obtained using the GS and BS models was the same (0.12). Using the GS model, the heritability for GL was 0.34 for the LW population and 0.33 for the LR population (Table 6.2). For the F1 population, the heritability obtained using the GS model was slightly lower than using the BS model (0.39 and 0.40, respectively). For both traits, the breed-specific additive genetic variance was higher for the alleles inherited from the LW population compared to the alleles inherited from the LR population (0.74 and 0.56 for LS, and 0.42 and 0.40 for GL, respectively).

The highest accuracies for predicting performance of crossbred sows were observed when training on crossbred data (Table 6.3). For LS, when training on crossbred data, prediction accuracies were the same for the GS breeding values as for the total BS breeding values (0.23). For GL, when training on crossbred data, slightly higher prediction accuracy was obtained for the total BS breeding values compared to the GS breeding values (0.53 and 0.52, respectively). For both traits, the BS breeding values from the LW alleles resulted in higher prediction accuracies than the BS breeding value from the LR alleles (0.21 vs. 0.12 for LS; 0.43 vs. 0.34 for GL).

Table 6.2 Variance estimates (and standard errors) from a traditional genomic selection model (GS) and a model that accounts for breed-specific effects (BS) for the traits litter size and gestation length.

Population [1]	Model	σ_s^2	σ_r^2	σ_a^2	$\sigma_{a_{LW}}^2$	$\sigma_{a_{LR}}^2$	σ_e^2	h^2
				Litter size				
LW	GS	0.43 (0.13)	1.70 (0.30)	1.81 (0.36)	-	-	8.26 (0.25)	0.15 (0.03)
LR	GS	0.02 (0.09)	1.22 (0.28)	1.30 (0.31)	-	-	8.80 (0.27)	0.12 (0.03)
F1	GS	0.10 (0.10)	1.38 (0.28)	1.42 (0.34)	-	-	8.54 (0.24)	0.12 (0.03)
F1	BS	0.11 (0.10)	1.37 (0.29)	-	0.74 (0.23)	0.56 (0.245)	8.52 (0.24)	0.12 (0.03)
				Gestation length				
LW	GS	0.21 (0.03)	0.33 (0.05)	0.68 (0.09)	-	-	0.78 (0.02)	0.34 (0.04)
LR	GS	0.23 (0.03)	0.26 (0.05)	0.64 (0.09)	-	-	0.82 (0.03)	0.33 (0.04)
F1	GS	0.16 (0.02)	0.23 (0.06)	0.81 (0.11)	-	-	0.90 (0.03)	0.39 (0.04)
F1	BS	0.16 (0.02)	0.17 (0.06)	-	0.42 (0.08)	0.40 (0.08)	0.90 (0.03)	0.40 (0.04)

[1] Large White (LW), Landrace (LR), two-way crossbred (F1). Variance components: service sire (σ_s^2), permanent environment (σ_r^2), additive (σ_a^2), additive for the alleles of the F1 population inherited from the LW ($\sigma_{a_{LW}}^2$) and LR ($\sigma_{a_{LR}}^2$) populations, and error (σ_e^2). h^2: heritability.

Table 6.3 Accuracy of predicting performance of crossbred sows for gestation length and litter size (40-fold cross-validation).

Model [1]	Training [2]	Accuracy [3]	SD [4]
		Litter size	
GS	LW	0.07	0.10
	LR	0.07	0.12
	F1	**0.23**	0.08
BS	LW	0.06	0.11
	LR	0.07	0.13
	LW and LR *	0.09	0.13
	$F1_{LW}$	0.21	0.08
	$F1_{LR}$	0.12	0.09
	$F1_{LW}$ and $F1_{LR}$ *	**0.23**	0.08
		Gestation length	
GS	LW	0.43	0.08
	LR	0.30	0.10
	F1	**0.52**	0.08
BS	LW	0.40	0.08
	LR	0.23	0.10
	LW and LR *	0.46	0.08
	$F1_{LW}$	0.43	0.08
	$F1_{LR}$	0.34	0.09
	$F1_{LW}$ and $F1_{LR}$ *	**0.53**	0.08

[1] GS: traditional genomic selection model; BS: model that accounts for breed-specific effects. [2] LW: Large White; LR: Landrace; F1: two-way crossbred; $F1_{LW}$: alleles of the F1 population inherited from the LW population; $F1_{LR}$: alleles of the F1 population inherited from the LR population. * Predicted breeding value was the "total breeding value" (sum of the breed-specific breeding values). [3] Average of the 40 replicates. Accuracy was defined as the correlation between the breeding values of the validation population and their average pre-adjusted phenotypes. [4] Standard deviation over replicates. The highest accuracies for each model and trait are given in bold.

6.4 Discussion

6.4.1 Breed-specific effects

The parental breeds LW and LR contributed differently to the additive genetic variance of LS and GL in their F1 offspring (Table 6.2), indicating the presence of breed-specific effects. Because prediction of breeding values from the different models required backsolving to obtain the SNP effects we can inspect these breed-specific effects from the different models in standard genome-wide association study (GWAS) plots (Figures 6.1 and 6.2). Plotting the per-SNP variances in GWAS plots, we have no traditional significance levels thresholds, such as *P* values, but we can identify regions (peaks) that explain more or less variance. For the trait LS, for example, pronounced peaks were observed on chromosomes 17 and 18 when evaluating the alleles that F1 animals inherited from the LW breed (Figure 6.1 D). However, when analyzing the alleles inherited from LR breed, a pronounced peak was observed only on chromosome 8 (Figure 6.2 E). These differences in GWAS peaks may indicate that breed-specific effects exist for LS in the F1 population.

Peaks on chromosomes 8, 17 and 18 were also observed when the F1 population was evaluated using the GS model, which does not distinguish breed origin of alleles (Figure 6.1 C). Peaks appeared to be sharper when the BS model was applied. Further, when evaluating the GWAS plots based on the two purebred populations (Figure 6.1 A and B), the peaks on chromosomes 8, 17, and 18 were not observed. A possible explanation for the absence of the peaks in purebred data is the interaction of SNPs with the environment (e.g. genotype-by-environment interactions) due to differences between the environments where the purebred and crossbred animals were housed. Purebred animals were housed in high-health environment (nucleus farms), while crossbred animals were housed in field conditions. Interactions of SNPs with the environment have been previously described for LS in pigs (Silva *et al.* 2014) and milk yield in cattle (Lillehammer *et al.* 2009). Genes with large and different effects in different environments could provide interesting tools for selecting the best parents to produce offspring for specific environments. Also, crossbred populations supply a different genetic background for the SNP effects (Bastiaansen *et al.* 2014). Therefore, the difference in the GWAS peaks from crossbred and purebred data could also be due to gene-by-gene interactions (e.g. epistasis), which is expected to make an important contribution to the genetic architecture of complex traits (Mackay 2014).

Even though the pattern we highlight with the peaks on chromosomes 8, 17 and 18 is consistent, the large difference between GWAS results of purebred and crossbred data was surprising. With the high genetic correlations of 0.90 between purebred and crossbred performance for both traits, largely the same QTLs were expected to affect the purebred as well as the crossbred traits. Power to detect associations might have contributed to the differences in the GWAS plots.

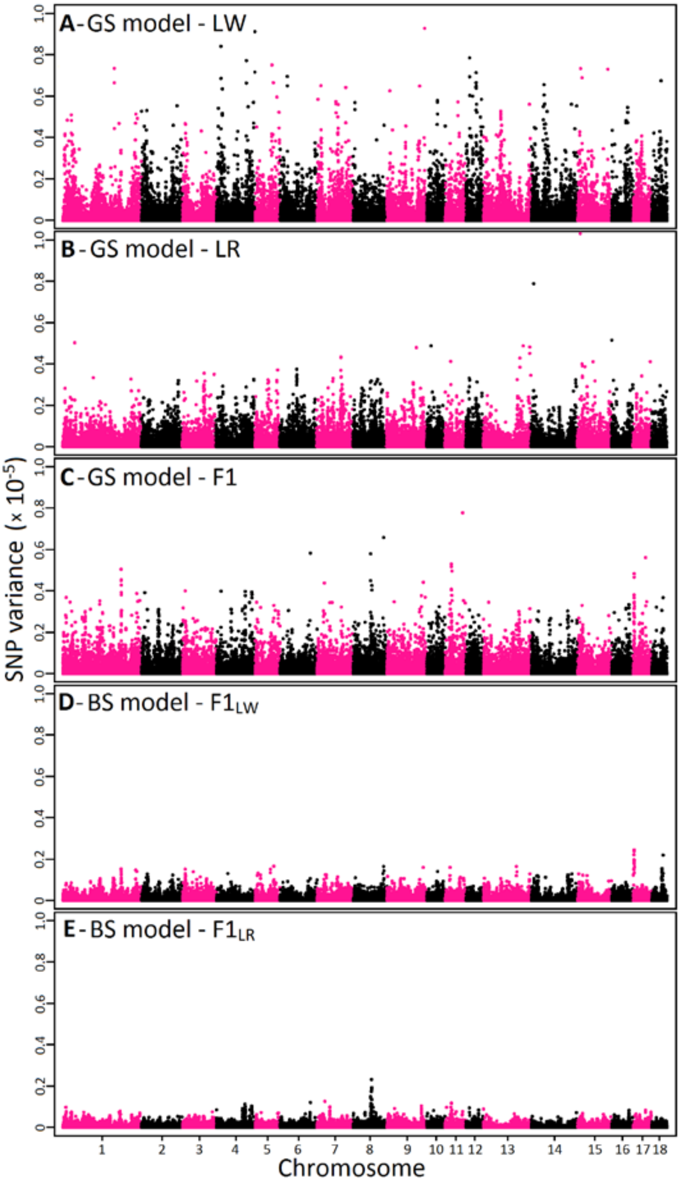

Figure 6.1 SNP variance for litter size estimated using a traditional genomic section model (GS) and a model that account for breed-specific effects (BS) in different training populations. LW: Large White; LR: Landrace; F1: two-way crossbred; $F1_{LW}$: alleles of the F1 population inherited from the LW population; $F1_{LR}$: alleles of the F1 population inherited from the LR population.

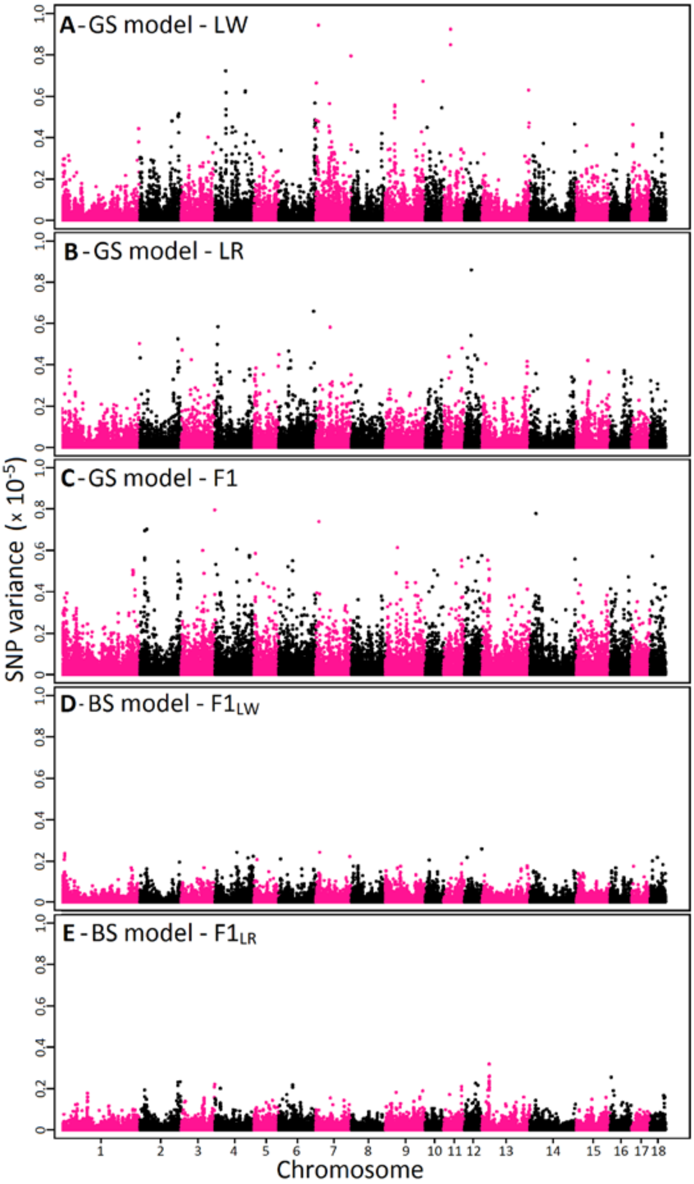

Figure 6.2 SNP variance for gestation length estimated using a traditional genomic section model (GS) and a model that account for breed-specific effects (BS) in different training populations. LW: Large White; LR: Landrace; F1: two-way crossbred; $F1_{LW}$: alleles of the F1 population inherited from the LW population; $F1_{LR}$: alleles of the F1 population inherited from the LR population.

In this study, we did show, for both traits, that the same SNP allele in the F1 population can contribute differently to the additive genetic variance depending on its breed origin and accounting for this can have an, albeit small, impact on prediction accuracy (Table 6.3). The standard errors of the breed-specific variance estimates were however rather high, especially for LS where it was up to 43% of the variance estimate (Table 6.2). These high standard errors are expected in a small dataset (N=924). While standard errors can increase due to inaccurate determination of the breed origin of alleles, this is expected to be limited because a stringent quality control was applied to the phased genotypes. When larger crossbred populations are genotyped, more accurate estimates of the contribution of the alleles inherited from different breeds to the genetic variation of the crossbred performance will be obtained.

6.4.2 Prediction accuracies

Predicting performance of genotyped crossbred sows was more accurate when training on crossbred data instead of purebred data (Table 6.3) even though training on crossbred data was at a small disadvantage because 10% of the dataset was set aside as validation population. These results are in line with previous studies that have shown the potential of training on crossbred data for predicting crossbred performance using both simulated and real data. In simulation studies, training on crossbred data has been reported to yield either the same (Toosi *et al.* 2010) or higher accuracies (Esfandyari *et al.* 2015a) compared to training on purebred data, while using real data, training on crossbred data has been described to yield the highest prediction accuracies (Hidalgo *et al.* 2015).

Predicting performance of genotyped crossbred sows based on the total BS breeding values resulted in similar or higher accuracies compared to using GS breeding values (Table 6.3). Additional benefit of using the BS model over the GS model is expected when crossbred populations are larger and more distant parental breeds are crossed (Esfandyari *et al.* 2015a). In the current study, we evaluated a small F1 population (N=924) that was obtained by crossing two dam lines. If a sire line was used instead of one dam line, the parental breeds may, depending on the lines chosen, become more distant (Veroneze *et al.* 2014) and breed-specific effects could eventually have a larger impact on the prediction accuracy. In pig breeding, the cross between sire and dam lines is typically done in the next generation, mating F1 sows to boars from a sire line. Applying the BS model to such a three-way crossbred population may show larger benefits when compared to the GS model because the parental lines are more distantly related. However, for evaluating three-way crossbred populations, even larger crossbred populations might be required because each crossbred will only carry two (grand) parental alleles.

Predictions of performance of genotyped crossbred sows using the BS breeding values from alleles of LW breed resulted in higher accuracies than using the BS breeding values from alleles of LR breed. This advantage of LW breed compared to LR is consistent with the larger amount of variance that the alleles of the LW breed explain (Table 6.2). Further, when training was performed on purebred data, performance of genotyped crossbred sows was more accurately estimated using the total BS breeding values than using the GS breeding values based on the SNP effects estimated in each purebred (Table 6.3). This suggests that determining the breed origin of the alleles on crossbred sows is beneficial even if the training is performed on purebred data.

In this study, the BS model was applied to a training population composed of crossbred animals only. As a further step, the benefits of accounting for breed-specific effects could be evaluated under a combined crossbred and purebred selection (CCPS), which has been described as an efficient way of obtaining genetic progress on both purebred and crossbred populations (Wei & Van der Steen 1991; Bijma & Van Arendonk 1998). Such an approach was proposed by Christensen *et al.* (2014) and consists of performing genomic evaluations using a combination of the genomic relationship of the purebred populations with the breed-specific relationship matrices of the crossbred population. One of the limitations of applying CCPS using pedigree-based models was that it also resulted in an increase of the rates of inbreeding (Bijma *et al.* 2001). However, with genomic-based models this increased inbreeding from CCPS is expected to be limited, or even absent, because the genomic information allows the estimation of Mendelian sampling and, therefore, reduces the emphasis on family information (Dekkers 2007).

Such a CCPS approach to estimate BS breeding values could be applied for both two-way and three-way crossbred populations. In the current study, we evaluated a two-way crossbred population and the determination of the breed origin of alleles was dependent upon pedigree information. In three-way crossbred populations pedigree information is not commonly recorded. Therefore, different strategies would be required for determining the breed origin of alleles. Recently, Bastiaansen *et al.* (2014) proposed a method to determine breed origin of alleles in crossbreds using long-range phasing that can be applied for crossbred populations where pedigree information is lacking. Also, close relationships between the crossbred and purebred genotyped animals would not be required because long-range phasing will work even with distant purebred relatives of the crossbreds.

Therefore, future studies on the practical application of BS models in CCPS should evaluate a combination of the methods proposed by Christensen *et al.* (2014) and Bastiaansen *et al.* (2014).

In this study, evaluation of the relevance of breed-specific effects when genomic selection is applied to real data was started. In further studies, evaluation of traits with lower genetic correlation between purebred and crossbred performance should be included because benefits of BS models are expected to be larger in those cases. In addition to less correlated traits, investigating breed-specific effects in crosses of more distant purebred populations may result in higher benefits of the BS model. Further, evaluation of larger datasets than the ones evaluated in the current study will also be required for more conclusive results and to quantify the benefits of accounting for breed-specific effects in prediction models.

6.5 Conclusions

In this study, we demonstrated evidence of breed-specific SNP effects for litter size and gestation length in a two-way crossbred population. Predicting performance of crossbred sows was shown to be more accurate when training was performed on crossbred data instead of purebred data. In addition, predictions based on the total breeding values from the BS model resulted in the same or higher accuracies compared to predictions based on breeding values from the GS model.

6.6 Acknowledgement

Financial support from the Breed4Food Consortium is gratefully acknowledged.

6.7 References

Bastiaansen, J., Bovenhuis, H., Lopes, M., Silva, F., Megens, H., Calus, M. (2014) SNP effects depend on genetic and environmental context. *Proceedings of the 10th World Congress of Genetics Applied to Livestock Production*. Vancouver.

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

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

Cecchinato, A., de los Campos, G., Gianola, D., Gallo, L., Carnier, P. (2010) The relevance of purebred information for predicting genetic merit of survival at birth of crossbred piglets. *J. Anim. Sci.,* **88,** 481-490.

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

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

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

Esfandyari, H., Sørensen, A.C., Bijma, P. (2015a) A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genet. Sel. Evol.,* **47,** 1-12.

Esfandyari, H., Sørensen, A.C., Bijma, P. (2015b) Maximizing crossbred performance through purebred genomic selection. *Genet. Sel. Evol.,* **47,** 16.

Gilmour, A.R., Gogel, B., Cullis, B., Thompson, R. (2009) ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

Groenen, M.A., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature,* **491,** 393-398.

Hanenberg, E., Knol, E., Merks, J. (2001) Estimates of genetic parameters for reproduction traits at different parities in Dutch Landrace pigs. *Livest. Prod. Sci.,* **69,** 179-186.

Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N., van der Werf, J.H. (2011) A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genet. Sel. Evol.,* **43,** 12.

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

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

Irgang, R., Fávero, J.A., Kennedy, B.W. (1994) Genetic parameters for litter size of different parities in Duroc, Landrace, and large white sows. *J. Anim. Sci.,* **72,** 2237-2246.

Knap, P., Su, G. (2008) Genotype by environment interaction for litter size in pigs as quantified by reaction norms analysis.

Lillehammer, M., Hayes, B., Meuwissen, T., Goddard, M. (2009) Gene by environment interactions for production traits in Australian dairy cattle. *J. Dairy Sci.,* **92,** 4008-4017.

Lopes, M.S., Bastiaansen, J.W.M., Janss, L., Knol, E.F., Bovenhuis, H. (2015) Estimation of additive, dominance, and imprinting genetic variance using genomic data. *G3: Genes|Genomes|Genetics***,** g3.115.019513.

Mackay, T.F. (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.,* **15,** 22-33.

Nishio, M., Satoh, M. (2015) Genomic best linear unbiased prediction method including imprinting effects for genomic evaluation. *Genet. Sel. Evol.,* **47,** 32.

Silva, F., Mulder, H., Knol, E., Lopes, M., Guimarães, S., Lopes, P., Mathur, P., Viana, J., Bastiaansen, J. (2014) Sire evaluation for total number born in pigs using a genomic reaction norms approach. *J. Anim. Sci.,* **92,** 3825-3834.

Toosi, A., Fernando, R., Dekkers, J., Quaas, R. (2010) Genomic selection in admixed and crossbred populations. *J. Anim. Sci.,* **88,** 32.

Uimari, P., Sironen, A. (2014) A combination of two variants in PRKAG3 is needed for a positive effect on meat quality in pigs. *BMC Genet.,* **15,** 29.

Van Grevenhof, I.E., Van der Werf, J.H. (2015) Design of reference populations for genomic selection in crossbreeding programs. *Genet. Sel. Evol.,* **47,** 1-9.

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

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

Wang, H., Misztal, I., Aguilar, I., Legarra, A., Muir, W. (2012) Genome-wide association mapping including phenotypes from relatives without genotypes. *Gent. Res.,* **94,** 73-83.

Wei, M., Van der Steen, H. (1991) Comparison of reciprocal recurrent selection with pure-line selection systems in animal breeding (A review). *Anim. Breed. Abstracts,* **59,** 281-298.

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

7

Artificial selection on introduced Asian haplotypes shaped the genetic architecture in European commercial pigs

Mirte Bosse[1¶], Marcos S Lopes[1,2¶], Ole Madsen[1], Hendrik-Jan Megens[1],
Richard PMA Crooijmans[1], Laurent A.F. Frantz[1], Barbara Harlizius[2],
John WM Bastiaansen[1], Martien AM Groenen[1]

[¶] These authors contributed equally to this work

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

Proceeding of the Royal Society B (2015) 282: 20152019

Abstract

Early pig farmers in Europe imported Asian pigs to cross with their local breeds in order to improve traits of commercial interest. Current genomics techniques enabled genome-wide identification of these Asian introgressed haplotypes in modern European pig breeds. We propose that the Asian variants are still present because they affect phenotypes that were important for ancient traditional, as well as recent commercial pig breeding. Genome-wide introgression levels were only weakly correlated with gene content and recombination frequency. However, regions with an excess or absence of Asian haplotypes contained genes that were previously identified as phenotypically important such as *FASN*, *ME1* and *KIT*. Therefore, the Asian alleles are thought to have an effect on phenotypes that were historically under selection. We aimed to estimate the effect of Asian haplotypes in introgressed regions in Large White pigs on the traits backfat and litter size. The majority of regions we tested that retained Asian DNA showed significantly increased backfat from the Asian alleles. Our results suggest that the introgression in Large White pigs has strongly been determined by the selective pressure acting upon the introgressed Asian haplotypes. We therefore conclude that human-driven hybridization and selection contributed to the genomic architecture of these commercial pigs.

Key words: Hybridization, domestication, *Sus scrofa*

7.1 Introduction

Introgression and subsequent selection on introgressed variants are thought to be a widespread phenomenon among many species (Currat *et al.* 2008). Introgression can occur naturally, due to mixture of populations in hybrid zones or occasional invasions. Then, selection for introgressed haplotypes can occur, a process known as adaptive introgression (Hedrick 2013). However, introgression can also be human-driven through hybridization, either accidental or on purpose (Crispo *et al.* 2011; Harrison & Larson 2014). Domestic animals are a clear example of species that have experienced population admixture due to human interference (Larson & Burger 2013). Introduction of and selection on novel alleles into a population has been observed in e.g. chicken (Eriksson *et al.* 2008), cagebirds (Rheindt & Edwards 2011) and cattle (Flori *et al.* 2014). Also in pigs, human-mediated hybridization has introduced haplotypes that cause desired effects on phenotypes (Bosse *et al.* 2014a).

Pig farming has undergone a true metamorphosis from first domestication until the intensified industry we know today. Pigs were domesticated independently leading to separate European and Asian domestic pigs some 10,000 years ago (Larson *et al.* 2005; Larson *et al.* 2007). Subsequent selection and breeding resulted in highly distinct breeds on these continents (Kijas & Andersson 2001; Megens *et al.* 2008). Especially in Europe, farmers were herding their swine in surrounding forests, and it was not before the Industrial Revolution during the eighteenth century that pigs were kept in sties and became an important farm animal (White 2011). Pigs were selected based on their phenotypes regarding traits of commercial interest.

In the last 100 years, with the improvements in performance recording and making use of genetic evaluation methods based on pedigree information, breeding programs have achieved a remarkable genetic progress in reducing backfat for carcass quality and improving growth rate for production efficiency. Since the 1990s, using the same traditional breeding strategies (pedigree-based), genetic progress has also been observed in reproduction traits, especially litter size (Merks *et al.* 2012). Part of the success of these breeding programs can also be attributed to the introduction of genes from Chinese breeds in commercial European breeds. About three centuries ago, with the intensification of global trade, farmers in Europe realized that Chinese pigs possessed particular characteristics that would be beneficial to introduce to their breeding stock. Therefore, pigs from Chinese breeds were imported to Europe and multiple crosses between European and Asian breeds were made with the purpose of combining beneficial traits, such as backfat (BF) and litter size (LS) from Asian pigs, and body length from European pigs (Jones *et al.* 1998; White 2011).

With the advent of genomic selection (Meuwissen *et al.* 2001) the genetic progress is expected to speed up even more. The design of a 60K SNPchip for pigs in 2009 (Ramos *et al.* 2009) and the publication of the pig reference genome in 2012 (Groenen *et al.* 2012) greatly contributed to the applicability of these techniques in pig breeding. This genomic information can be used also to pinpoint regions in the genome that have been under selective pressure. The resulting changes at the DNA level have been detected as selective sweeps in a multitude of breeds (Rubin *et al.* 2012; Wilkinson *et al.* 2013). Interestingly, some of these loci have been identified as not only being under selection but also introgressed. Asian alleles at the *EDNRB* (Wilkinson *et al.* 2013), *IGF2* (Ojeda *et al.* 2008) and *KITLG* (Okumura *et al.* 2008) locus have proven effects on phenotypes (e.g. meat content and coat color) of European commercial breeds.

With the current genomic techniques, it has become possible to trace back the haplotypes that were introduced during the Industrial Revolution. In (Bosse *et al.* 2014a), we examined the occurrence of Asian haplotypes in a population of European pigs that belong to the commercial Large White breed. Our results showed that Asian haplotypes are widely present in the genomes of these commercial pigs and highlighted the effect of the introgressed Asian variant of the *AHR* gene on litter size. Since the Asian haplotypes were introduced for a specific purpose, the effect on the phenotype should co-occur with the presence of Asian variants in regions of the genome that are associated with the traits known to have been under selection. However, how much influence the Asian introgression had on the selective history of commercial traits remains unknown.

We hypothesize that the introgression landscape in commercial breeds is shaped mostly by artificial selection and, therefore, most introgressed regions should have an effect on commercial traits (BF and LS). Also, an absence of Asian haplotypes in some parts of the Large White genomes could be the result of purifying selection. In this study, we examine the introgression signatures that we identified previously in (Bosse *et al.* 2014a) in more detail, showing that the majority of the regions we tested has a significant effect on BF in a commercial Large White population. These findings have important implications for the knowledge on natural and human-driven evolutionary forces shaping genomes after hybridization.

7.2 Material and methods

7.2.1 Introgression data

The analyses in this paper build upon the dataset and results that were obtained in Bosse *et al.* (2014a) and Bosse *et al.* (2014b). Briefly, the proportion of Asian introgressed haplotypes in a population of Large White pigs was assessed over all autosomes. Introgression mapping was performed on a group of 9 Large White pigs and the background of their haplotypes was assigned to be European or Asian (Bosse *et al.* 2014a). In bins of 10kb over the genome, the relative Asian introgression signal in the Large White population was obtained (Figure 7.1A). We assessed whether the Large White haplotypes were identical by descent (IBD) with Asian and/or European haplotypes, and calculated the relative frequency of IBD with Asian haplotypes in the population for each bin (rIBD). We used these genome-wide rIBD signals to understand the details of the Asian introgression (See Supplementary Text 7.1 for more details about the introgression mapping).

7.2.2 Genome characteristics

To assess the correlation between gene density, recombination frequency and introgression signal, we averaged the rIBD in 1MB bins and counted the number of genes within each bin. The recombination map from Tortereau *et al.* (2012) was used to obtain the recombination frequency per Mb. To test whether the probability of introgression decreases with an increase in number of genes in a region, we used the Pearson's product-moment correlation in R.

Selection of regions

We used the ~400 regions of introgression previously identified by Bosse *et al.* (2014a) with a Z-transformed rIBD (ZrIBD) >2. The regions were extended with one 10kb bin at the time to the left and right flank of each identified region of introgression, until the threshold of >2 ZrIBD was no longer met and/or the rIBD value for one particular 10kb bin was <0. We found 33 regions of introgression that were longer than 150kb and checked whether they physically overlapped with markers on the Illumina Porcine 60K iSelect Beadchip (60K chip). Three regions were discarded because they contained less than 3 segregating markers on the chip. Table 7.S1 contains the list of 30 regions that were included in the further analyses.

7.2.3 Genotyping and phasing

We genotyped a total of 9,970 pigs and wild boar for 488 markers on the 60K chip that fell within the 30 identified regions of introgression. Phasing of haplotypes was done independently for each region with Beagle V. 3.3.2 (Browning & Browning 2007), using the genotype information for all individuals. After the phasing step, we used haplotype data from three groups: Asian (N= 448), European wild (N= 920) and European commercial (N= 18,572). Because the introgression analysis was done on Large White pigs, we extracted only those individuals from the European commercial group that were known to be purebred Large Whites, leaving us with a total of 4,764 Large White haplotypes.

7.2.4 Determination of haplotype origin

The total number of observed reference haplotypes in the group of European wild boar was 920, and the total number of observed Asian reference haplotypes was 448. For each of the 4,764 haplotypes observed in Large White animals, we determined the Asian (AS) or European (EU) origin based on the frequencies of this haplotype in the European and/or Asian group of animals. For each region, we counted the number of unique haplotypes among the 448 haplotypes in the Asian group. Because we have unequal sampling, the number of unique haplotypes in a random sample of 448 haplotypes from the European group was also counted. Then, for each unique haplotype that was observed within the group of Large White haplotypes, we counted the number of times the haplotype was observed in the European and in the Asian group. To avoid a bias due to the (generally) higher diversity in Asia, we adjusted the counts for the amount of diversity in the Asian and European groups. This adjustment was done by calculating the ratio of unique haplotypes in both AS and EU groups. The number of times that particular haplotype was observed in the European group was multiplied by the proportion of unique EU compared to AS haplotypes, and the number of times the haplotype was observed in the Asian reference group was multiplied by the proportion of total observed Asian haplotypes. We then checked whether the corrected number of observed haplotypes in the European and Asian reference groups differed at least by a factor 4. If so, the haplotype was assigned to the group in which it was observed most. If not, it was assigned to the group for which both backgrounds were considered ("Both").

7.2.5 Cleaning of the data

Introgression regions in which the genotypes showed strong deviation from Hardy-Weinberg Equilibrium (P<0.00001) and the frequency of Asian haplotypes was <0.60 or >0.99 were excluded from further analysis. As we evaluated regions with a strong signal of Asian introgression, we expected that the Asian haplotypes should be in higher frequency than European haplotypes. We assumed that regions where the frequency of Asian haplotypes were <0.60 were observed due to possible errors during the phasing procedures and, therefore, they were excluded. We excluded regions where the frequency of Asian haplotypes were >0.99 because this means that these regions were fixed, which does not allow to test the association of these regions with the evaluated phenotypes. In order to test the independence of the evaluated regions, we also estimated the pairwise Pearson's correlations (r) for all regions using the recoded haplotypes. When two regions showed r >0.80, the shortest regions were excluded. After cleaning procedures, a total 1,384 animals with haplotype information on 11 regions were available for further steps (see Supplementary Text 7.2 for more details).

7.2.6 Breeding values and association analyses

In this study we evaluated the traits backfat (BF) and litter size (LS). Deregressed estimated breeding values (DEBV) were used as the response variable for each trait under study. The estimated breeding value (EBV) was deregressed for each trait separately using the methodology described by Garrick *et al.* (2009). The EBV of each animal was obtained from the routine genetic evaluation by Topigs Norsvin using an animal model (pedigree-based). The model for BF included genetic line, sex, herd-year-month and weight as fixed effects and an additive genetic effect (animal) and a common litter effect as random effects. For LS, the model included genetic line, parity number, interval weaning-pregnancy (days), whether more than one insemination procedures were performed (yes or no) and herd-year-season, while the random effects consisted of service sire, a permanent effect to account for the repeated observations of a single sow, and an additive genetic effect (animal). The reliabilities per animal for the purpose of deregression were extracted from the genetic evaluation based on the methodology of Tier and Meyer (2004). The heritabilities used for the deregression were also extracted from the routine genetic evaluation.

The association analyses were performed using the software ASReml (Gilmour *et al.* 2009) applying the following model:

$$\mathrm{DEBV}_{ij}\ \mathrm{w} = \mu + \mathrm{R}_i + \mathrm{a}_j + \mathrm{e}_{ij}$$

where $DEBV_{ij}$ is the observed DEBV for the animal j, μ is the overall DEBV mean of the population, R_i is the count of Asian haplotypes (AS) of the region i, a_j is the additive genetic effect estimated using a pedigree-based relationship matrix and e_{ij} the residual error. The weighting factor *w* was used in the association analyses to account for the differences in the amount of offspring information available for the estimation of the DEBV (Garrick *et al.* 2009). To ensure the quality of the DEBV, only animals with a *w* higher than zero and a reliability of the DEBV >0.20 were used. The reliability of the DEBV was obtained according to Garrick *et al.* (2009).The association analyses were performed per region. In addition, a combined analysis was done where R represented the count of AS summed over all regions.

7.3 Results and Discussion

7.3.1 Effect of introgression on commercial traits

We hypothesized that the pattern of introgression in the Large White population is mainly determined by artificial selection acting upon the Asian haplotypes. Following this rationale, the Asian introgression should persist mainly in those regions of the genome where the Asian variant has a favorable effect on a phenotype of interest. To test this, we extracted haplotypes in the introgressed regions and estimated the effect of their origin (European or Asian) on two commercially important traits: backfat (BF) and litter size (LS). A total of 2,382 individuals from the commercial Large White line were genotyped for markers on the 60K SNPchip (Ramos *et al.* 2009) that cover those regions. More specifically, we extracted 11 regions that had an introgression signal that persisted over more than 150 Kbp from the data presented by (Bosse *et al.* 2014a), and that passed our thresholds for data cleaning (see Material and methods and Supporting information for more details). The further analyses for these regions were based on this selection of 60K markers.

7.3.2 Effects per region

We evaluated whether these 11 regions were significantly associated with the traits BF and LS. None of these regions were found to have a significant effect on LS (Table 7.1). However, six of these 11 regions showed a significant association with BF (Table 7.1). For all these significant regions, we observed an increase in BF when a European haplotype was replaced by an Asian haplotype.

Most introgressed regions were identified on chromosome 2 (Figure 7.1B, Figure 7.S1A with the strongest effect in the gene-dense region 2_2 (0.22 mm of increase in BF). This region contains multiple genes coding for intercellular adhesion molecules (*ICAMs*) that have been shown to have an effect on obesity

(Dong *et al.* 1997). Whether the effect of BF is caused by these genes is however unclear, because the regions contain a total of 39 annotated genes in the current Ensembl release 76.

Table 7.1 Effect of Asian haplotypes on backfat and litter size in introgression regions.

Region	Backfat		Litter size	
	P-value	Effect	P-value	Effect
15_1	0.560	-0.07	0.110	0.21
18_1	**0.024**	0.15	0.617	-0.03
2_2	**0.002**	0.22	1.000	0.00
2_4	0.420	0.07	0.471	0.07
2_6	**0.027**	0.15	0.203	0.09
2_7	**0.011**	0.17	0.234	-0.09
3_1	0.527	-0.05	0.729	0.03
6_1	0.299	-0.42	0.348	-0.41
9_1	0.823	-0.05	1.000	0.00
9_3	**0.081**	0.14	0.590	0.05
9_5	**0.054**	0.13	1.000	0.00
All combined	**<0.001**	0.09	0.610	0.01

An animal model was used to estimate the effect of Asian haplotypes on the deregressed estimated breeding values for the two traits. P-values of significant regions (P<0.10) are given in bold. Effect for BF is in mm BF per Asian haplotype, and effect for LS is in number of piglets per Asian haplotype. 'All combined' indicates the regression analysis over all 11 regions.

Region 2_7 is the region with the strongest introgression signal on chromosome 2 and, therefore, it was used as an example of the applied method in Figure 7.S1. The substitution of a European haplotype for an Asian haplotypes in this region on average increased BF by 0.17 mm. As can be seen in the Ensembl annotation for these regions (Figure 7.S1B), one candidate gene, *COMMD10*, lies within the peak of region 2_7. COMMD proteins contain a conserved and unique 'COMM' domain involved in cellular homeostasis including copper and the NFkβ pathway, and at least 10 COMM family members exist that are conserved in all vertebrates (Maine & Burstein 2007). Murgiano *et al.* (2010) found another *COMMD* gene differentially expressed in *longissimus lumborum* muscle samples between Large White and Casertana pigs, and suggest that *COMMD* negatively regulates NFkβ signaling which in turn can result in triggering the adipogenetic cascades.

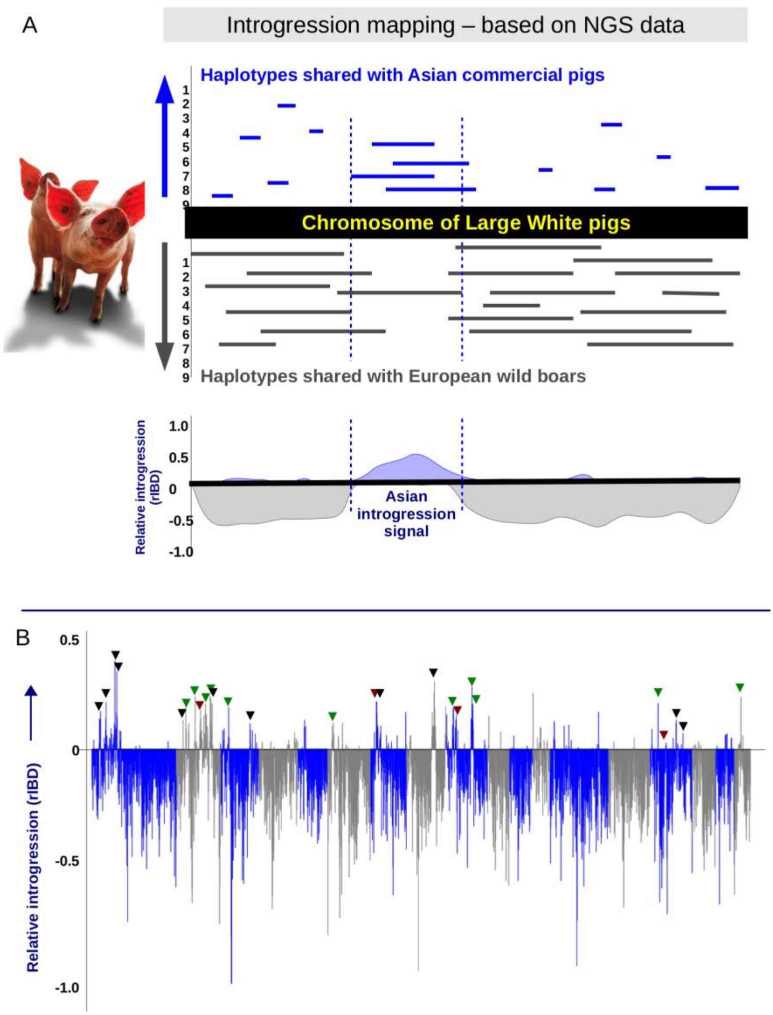

Figure 7.1 The principle of introgression mapping. The purpose of introgression mapping is to determine what the background is of haplotypes in a particular part of the genome. **1A**. Haplotypes that were shared with either Asian or European pigs were mapped to the genome for all 9 Large White pigs. The total numbers of overlapping haplotypes were counted and the relative introgression signal (rIBD) was obtained by taking the difference of haplotype frequencies as described in Bosse et al (2014a). **1B.** The y-axis displays the rIBD signal averaged in bins of 1Mb, and the x-axis contains the physical position for all 18 autosomes. Regions that were selected for the commercial trait analysis are indicated, with green triangles indicating included regions, black triangles indicating selected regions with strong HW disequilibrium and red triangles indicating regions discarded because of allele frequencies.

The other significant region (2_6) on this chromosome contains 7 candidate genes for increased BF (*CAST, ERAP1, ERAP2, LNPEP, LIX1 RIOK2* and *RGMB*). *LNPEP* is an insulin-regulated amino peptidase which acts as membrane protein associated with glucose transporter vesicles in cultured mouse adipocytes according to Larance *et al.* (2005). *CAST*, calpastatin, has no direct known function in fat synthesis, but interestingly it is a well-known locus involved in meat tenderness in multiple species and studies, see: Ropka-Molik *et al.* (2014) and meat quality traits of pigs in general (Ciobanu *et al.* 2004).

Regions 9_3 and 9_5 are both located on chromosome 9, in the vicinity of the *AHR* gene that was identified in (Bosse *et al.* 2014a), with an effect of ~0.15 mm of BF per Asian haplotype (Table 7.1). Region 9_3 overlaps with the AHR gene, suggesting that this gene is involved in multiple biological processes, as discussed by (Denison *et al.* 2011) and (Hernandez-Ochoa *et al.* 2009). Region 9_5, in addition to its proximity to *AHR*, contains two *TWIST* neighbor genes (*TWISTNB*) that are involved in transcription and the *TMEM196* gene coding for transmembrane protein 196. The last significant region 18_1 on chromosome 18 contains only one gene, protection of telomeres1 (*POT1),* that has previously been shown to have a higher expression level in multiple tissues from the fat-type Wujin pigs, compared to Large White pigs, including *longissimus dorsi* muscle (Yong *et al.* 2012).

In summary, for the 11 regions with a strong introgression signal, the Asian haplotypes displayed a significant positive effect on BF in the majority of regions. Zooming into the genes in the BF-associated regions, a multitude of candidate genes could be identified that possibly caused the effect on BF. We suggest further experiments that focus on these specific genes to confirm their role in the accumulation of BF in pigs. Selecting those regions with the most important signal of introgression might have introduced some bias making the results less representative. In Supplementary Text 7.4 we discuss why we believe the potential bias is minor in our analysis.

7.3.3 Additive effect over regions

The effect of the Asian haplotypes is an increase in BF in all significant regions, suggesting that the Asian haplotypes could have an additive effect on BF over all regions. We examined the association of Asian haplotypes with BF and LS combining all regions (summing the count of Asian haplotypes of all regions). We performed the regression of the counts of Asian haplotypes (ranging from 9 to 22, Figure 7.S2) on both BF and LS. For LS, the association test was not significant, but for BF the association was even stronger than when individual regions were analyzed (Table 7.1). For BF, an additive effect of 0.09 mm of BF was observed per

Asian haplotype that replaced a European haplotype (Table 7.1). The overall phenotypic standard deviation of BF in the Large White population was 1.61 mm. Analyzing all 11 regions together, an individual that presents only Asian haplotypes (n=22) will show 1.98 mm of BF more than an animal that presents only European haplotypes, which means 1.23 phenotypic standard deviations of BF. To demonstrate that the associations between the introgression regions and BF found in this study are significantly different from those expected by chance, we performed a theoretical exercise which is described in detail in the Supplementary Text 7.5.

Chinese breeds were thought to be superior for the traits fatness and litter size according to the early European pig farmers, and these traits were artificially selected after introgression (White 2011). Müller *et al.* (2000) showed that the amount of BF observed Meishan pigs was 2.38 fold the amount of BF observed in Pietrain pigs (32.83 mm and 13.78 mm, respectively). Haley *et al.* (1995) comparing litter size between Meishan and Large White sows at first parity, showed that the Meishan sows had in average 28% more piglets than Large White sows (13.03 piglets and 10.20 piglets, respectively). Our initial hypothesis was, therefore, that the regions with a strong introgression signal would have a significant effect on both BF and LS. However, our results showed that regions of introgression only have a significant effect on BF. Selection signatures for complex traits do not necessarily leave a sweeplike signature in the genome (Heidaritabar *et al.* 2014; Kemper *et al.* 2014). This could explain why none of the introgressed regions display a significant association with LS. Indeed, if we look at the previous finding for the *AHR* gene (Bosse *et al.* 2014a) it is one particular Asian allele rather than all Asian haplotypes at that locus that have the effect on LS. Another explanation why the introgressed Asian haplotypes have no effect on LS could be that the specific loci that are involved in a complex trait like litter size contain genes that are also involved in other (life history) traits. The pleiotropic nature of these genes may restrict the selection on Asian haplotypes, resulting in less obvious signatures in the genome than for BF related genes, although this is speculative.

Our results demonstrate that by screening a population for signals of introgression, regions can be pinpointed where introduced haplotypes have an effect on selected traits. Popular methods that are developed to detect selective sweeps in a population, like Fst (Weir & Cockerham 1984) or homozygosity tests, use increase or reduction of genetic variation as a signal. Ongoing selection for introduced haplotypes that are genetically more diverse or distant than haplotypes from the source population will not be picked up by these methods. We therefore suggest consideration of alternative methods (such as a test for the background of haplotypes as we have described in the manuscript) when the studied population has a known history of admixture, or when the goal is to screen specifically for adaptive introgression.

7.3.4 General patterns of introgression

Our hypothesis was that artificial selection is the main factor in shaping the introgression pattern in Large White pigs. According to Hedrick (2013), the probability of an introgressed haplotype to be maintained in a population is strongly increased when it has some selective advantage. The results of the association analysis support the hypothesis that introgression signals are enriched for associations with commercially-interesting traits. In line with this hypothesis, general genome characteristics like gene density and recombination frequency should contribute little to the introgression pattern. To assess whether the introgression signal is correlated with gene density or recombination frequency, the rIBD was averaged over 1MB bins over the genome. We found a very modest significant negative correlation of -0.05, as well as a significant correlation between rIBD and (log transformed) recombination frequency of 0.10 (Figure 7.2). In a recent study on Neanderthal introgression in modern humans, gene deserts were enriched for Neanderthal ancestry (Sankararaman *et al.* 2014). This finding suggests that purifying selection removed the majority of Neanderthal haplotypes from the population and that introgressed haplotypes mainly occur in regions with relaxed selection pressure. In pigs, these general patterns in the genome explain only a fraction of the variation in introgression signatures. The circumstances of introgression in these two species are however very different, since the admixture in commercial pigs has been deliberate, and selection for some of the introgressed haplotypes is expected to be positive because of the known differences between Asian and European pigs.

Although the purpose of the introgression of Asian haplotypes was to maintain beneficial variants in the population, probably not all introduced Asian haplotypes had the desired effect on the European stock. Selection pressure on Asian haplotypes could have been either positive or purifying, and should have resulted in either an excess or an absence of the Asian variants, depending on the location in the genome and the associated genes in those regions. Commercial pig breeds are known to be under strong artificial selection, and our results have shown that indeed a majority of the strongly introgressed regions we tested had a significant effect on BF. Therefore, we expect that regions with an excess or absence of Asian haplotypes contain genes of commercial interest. The 1% extreme tails of the introgression distribution (see Table 7.S2) were scanned for genes that are known to be related to commercially important traits. Introgressed regions should contain genes that have an effect on traits that are present in the Asian breeds and that had a selective advantage in the European breeds. By contrast, within those regions that do not contain Asian haplotypes, we expect to identify genes that have an effect on traits that are typical for European breeds. For the regions with lack of introgression signal, it would be interesting to create experimental crosses to introduce Asian haplotypes and test whether we can validate that indeed phenotypes of commercial interest are affected by locus-specific Asian introgression.

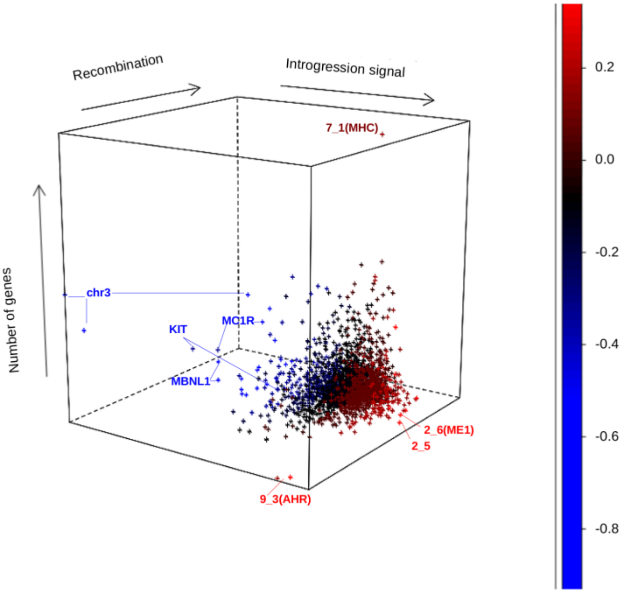

Figure 7.2 rIBD and genome characteristics. For each 1Mb bin in the genome, the rIBD signal was plotted against the number of genes and the recombination frequency in the bin. Coloration is based on the rIBD signal so that bright red indicates bins with the strongest Asian introgression signal, and blue bins indicate the strongest underrepresentation of Asian haplotypes. Number of genes per bin ranged from 0 to 128 and recombination frequency ranged from -16.6 to 3.3 (log2-transformed) cM/Mb. Bins containing interesting genes that are discussed in the main text are indicated.

7.3.5 Fat related genes

The two 1Mb bins containing the highest rIBD scores are both on chromosome 1. When we look for candidate genes in the first bin, the malic enzyme 1 (*ME1*) gene has previously been described as an important QTL region for

BF and meat quality (Vidal *et al.* 2006; Bartz *et al.* 2013; Ramírez *et al.* 2014). This bin overlaps with region 1_6 from the association analysis, but that region was discarded because of an excess of heterozygotes. The second bin contains *AMD1* (cell proliferation, polyamine synthesis pathway) and zeta catalytic subunit DNA polymerase (rev3L in human) as candidate genes. This bin overlaps with region 1_5 from the association analysis, but it was also discarded due to an excess of heterozygotes.

On the beginning of chromosome 12, we identified another bin with a high rIBD signal, hinting at a locus that contains Asian introgressed haplotypes. This region overlaps the FASN gene that was previously described as fatty acid synthase and an important gene involved in fat deposition Braglia *et al.* (2014). In addition to the genes described in this paragraph, our 11 regions that are used in the association analysis can be found in the 25 bins that span the top 1% of introgressed regions in the genome. These findings hint at a prominent role for selection on fatness in shaping the introgression landscape in Large White pigs.

The reason for the observation that more heterozygous individuals are observed than one would expect based on Hardy-Weinberg equilibrium remains unclear. Apart from the technical issues, these regions are potentially very interesting if the signal is genuine. Balancing selection or a heterozygote advantage can result in more heterozygous individuals than expected. Also, if selection for a previously low-frequency haplotype is ongoing, an excess of heterozygous individuals could be observed. As shown by Merks *et al.* (2012), BF was included in the selection index of breeding companies over 100 years ago, and nowadays leanness is preferred in commercial pigs. This switch in preference and direction of selection could result in more heterozygosity in some regions. Further experiments for the regions with an excess of heterozygotes should indicate what causes this peculiar pattern.

7.3.6 Pigmentation genes

Skin pigmentation is an important trait for modern pig breeders, and therefore we expect a distinct introgression signal at pigmentation loci (Figure 7.S3). In the top 1% of bins with low Asian introgression are two regions that have previously been identified as important candidate regions for coat coloration in pigs, containing the *KIT* gene and the *MC1R* gene. The *KIT* gene is a very well-known gene that is involved in coat color (Okumura *et al.* 2008). The European pigs contain a copy number variable dominant white allele, mostly in homozygous form (Rubin *et al.* 2012; Paudel *et al.* 2013). We clearly see European haplotypes surrounding this gene rather than Asian, suggesting selection against introgression at this locus. Also present in this peak region is *MAP9*, a gene involved in mitotic spindle formation (Figure 7.S3). *MC1R* is known to be involved in pigmentation in multiple species including pigs. No introgression is expected in Large Whites for this region based on previous results (Fang *et al.* 2009), and indeed we see a clear lack of introgression at this locus (Figure 7.S3). Interestingly, among those bins that contain the highest introgression signal in the Large Whites, were two other genes found to be involved in pigmentation (*TYR* and *RAB38*). *RAB38* and *TYR* are both involved in pigmentation patterns of skin, eyes and hair, according to a multitude of different studies (del Marmol & Beermann 1996; Loftus *et al.* 2002). Tyrosinase seems to have some temperature-dependent coloration patterns, but different forms exist. *RAB38* lies within region 9_1, and has been identified in (Wilkinson *et al.* 2013) as well as a region in Large White pigs that is introgressed and selected.

Morphology

The 1Mb bin covering *MBNL1* (muscleblind-like splicing regulator 1 gene) has the lowest introgression signal. In human and mouse, this gene has been shown to be associated with muscle dystrophy (Kanadia *et al.* 2003). Since European commercial pigs and Asian pigs are known to be very different in terms of muscle content, selection for muscle related traits in European pigs might select against Asian variants in this region. In the introgression peak at the very end of chromosome 8 we identified another interesting gene, *BMP3* (bone morphogenesis protein 3), that has previously been identified as being involved in growth restriction in human (Bonnet *et al.* 2010) and a mutation in this gene has an effect on skull shape in zebrafish and dogs (Schoenebeck *et al.* 2012). This region was also found to be introgressed and selected in LW pigs in Wilkinson *et al.* (2013).

7.4 Conclusions

With this work, we demonstrate that the introgression landscape in Large White pigs seems to be strongly determined by the selective pressure acting upon the introgressed Asian haplotypes. The majority of the regions we tested with a high frequency in Asian haplotypes turn out to have an effect on BF, and many of these regions overlap with previously identified fat-related genes, and therefore we conclude that artificial selection on fatness influenced the introgression landscape in Large White pigs. To investigate whether this is expected behavior of introgressed haplotypes under selection, we propose future simulation studies on the introgression landscape of populations under a neutral scenario and with selection pressure. We then hypothesize that introgressed haplotypes will be elevated to high frequency due to positive selection, and other introgressed haplotypes will quickly be removed because of purifying selection. The fact that the proportion of Asian material is relatively similar for most European commercial breeds (Groenen *et al.* 2012; Bosse *et al.* 2014b) suggest that the introgression occurred before the establishment of modern breeds. After many generations since the introgression, Asian haplotypes could have been purged if they had a selective disadvantage. On a genome-wide scale, however, we observed a general pattern of low frequency of Asian haplotypes, suggesting a more neutral scenario for the introgressed genetic material in this Large White pig population. Regions with an excess or absence of Asian haplotypes indeed contain genes where the Asian variants are thought to have an effect on phenotypes of interest, and therefore we illustrate that human-driven introgression and selection may have broadly shaped the genomic architecture of this commercial pig breed. Extending this study to more commercial traits and different breeds will provide more insight into the process of selective introgression. Our findings provide a unique insight about how the selection history of pig breeding influenced the genomic haplotype patterns of the commercial pigs that we know today. How general this introgression pattern is, should be pointed out by future studies on other organisms that likely experienced introgression and consecutive selection.

7.5 Acknowledgements

The DNA samples that we used for genotyping were provided by Dr Ning Li, China Agricultural University, China; Dr Alain Ducos, UMR INRA-ENVT, France; Sem Genini, Parco technologico Padano, Italy; Dr Gono Semiadi, Puslit Biologi, Indonesia; Dr Naohiko Okumura, Staff Institute 446-1 Ippaizuka, Japan; Dr Alan Archibald, Roslin Institute and the Royal (Dick) School of Veterinary Studies, University of Edinburgh , Scotland; Topigs Norsvin Research Center, The Netherlands; Dr Oliver Ryder, San Diego Zoo, USA; Cheryl L. Morris, Ph.D., Omaha's Henry Doorly Zoo, USA. We thank Bert Dibbits for lab work, B. van de Water for graphics production and Yogesh Paudel and Egbert Knol for valuable discussion. We also thank Oscar Gaggiotti, Pablo Orozco-ter Wengel, and an anonymous reviewer for all their valuable comments on our manuscript.

7.6 References

Bartz, M., Kociucka, B., Mankowska, M., Switonski, M., Szydlowski, M. (2013) Transcript level of the porcine ME1 gene is affected by SNP in its 3′ UTR, which is also associated with subcutaneous fat thickness. *J. Anim. Breed. Genet.,* **131,** 271–278.

Bonnet, C., Andrieux, J., Béri-Dexheimer, M., Leheup, B., Boute, O., Manouvrier, S., Delobel, B., Copin, H., Receveur, A., Mathieu, M.a. (2010) Microdeletion at chromosome 4q21 defines a new emerging syndrome with marked growth restriction, mental retardation and absent or severely delayed speech. *J. Med. Genet.,* **47,** 377-384.

Bosse, M., Megens, H.-J., Frantz, L.A., Madsen, O., Larson, G., Paudel, Y., Duijvesteijn, N., Harlizius, B., Hagemeijer, Y., Crooijmans, R.P. (2014a) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. *Nature communications,* **5**.

Bosse, M., Megens, H.J., Madsen, O., Frantz, L.A., Paudel, Y., Crooijmans, R.P., Groenen, M.A. (2014b) Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent Sus scrofa populations. *Mol. Ecol.,* **23,** 4089-4102.

Browning, S.R., Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics,* **81,** 1084-1097.

Ciobanu, D., Bastiaansen, J., Lonergan, S.M., Thomsen, H., Dekkers, J.C., Plastow, G.S., Rothschild, M.F. (2004) New alleles in calpastatin gene are associated with meat quality traits in pigs. *J. Anim. Sci.,* **82,** 2829-2839.

Crispo, E., Moore, J.S., Lee-Yaw, J.A., Gray, S.M., Haller, B.C. (2011) Broken barriers: Human-induced changes to gene flow and introgression in animals. *BioEssays,* **33,** 508-518.

Currat, M., Ruedi, M., Petit, R.J., Excoffier, L. (2008) The hidden side of invasions: massive introgression by local genes. *Evolution,* **62,** 1908-1920.

del Marmol, V., Beermann, F. (1996) Tyrosinase and related proteins in mammalian pigmentation. *FEBS letters,* **381,** 165-168.

Denison, M.S., Soshilov, A.A., He, G.C., DeGroot, D.E., Zhao, B. (2011) Exactly the Same but Different: Promiscuity and Diversity in the Molecular Mechanisms of Action of the Aryl Hydrocarbon (Dioxin) Receptor. *Toxicological Sciences,* **124,** 1-22.

Dong, Z.M., Gutierrez-Ramos, J.-C., Coxon, A., Mayadas, T.N., Wagner, D.D. (1997) A new class of obesity genes encodes leukocyte adhesion receptors. *Proceedings of the National Academy of Sciences,* **94,** 7526-7530.

Eriksson, J., Larson, G., Gunnarsson, U., Bed'hom, B., Tixier-Boichard, M., Strömstedt, L., Wright, D., Jungerius, A., Vereijken, A., Randi, E. (2008) Identification of the yellow skin gene reveals a hybrid origin of the domestic chicken. *PLoS genetics,* **4,** e1000010.

Fang, M., Larson, G., Ribeiro, H.S., Li, N., Andersson, L. (2009) Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS Genetics,* **5,** e1000341.

Flori, L., Thevenon, S., Dayo, G.K., Marcel, S., Sylla, S., Berthier, D., Moazami-Goudarzi, K., Gautier, M. (2014) Adaptive admixture in the West African bovine hybrid zone: insight from the Borgou population. *Mol. Ecol.*

Garrick, D.J., Taylor, J.F., Fernando, R.L. (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol,* **41,** 44.

Gilmour, A.R., Gogel, B., Cullis, B., Thompson, R. (2009) ASReml user guide release 3.0. *VSN International Ltd, Hemel Hempstead, UK*.

Groenen, M.A., Archibald, A.L., Uenishi, H., Tuggle, C.K., Takeuchi, Y., Rothschild, M.F., Rogel-Gaillard, C., Park, C., Milan, D., Megens, H.-J. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature,* **491,** 393-398.

Haley, C., Lee, G., Ritchie, M. (1995) Comparative reproductive performance in Meishan and Large White pigs and their crosses. *Anim Sci,* **60,** 259-267.

Harrison, R.G., Larson, E.L. (2014) Hybridization, Introgression, and the Nature of Species Boundaries. *J Hered,* **105,** 795-809.

Hedrick, P.W. (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol Ecol,* **22,** 4606-4618.

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

Hernandez-Ochoa, I., Karman, B.N., Flaws, J.A. (2009) The role of the aryl hydrocarbon receptor in the female reproductive system. *Biochemical Pharmacology,* **77,** 547-559.

Jones, G., Rothschild, M., Ruvinsky, A. (1998) Genetic aspects of domestication, common breeds and their origin. *The genetics of the pig.***,** 17-50.

Kanadia, R.N., Johnstone, K.A., Mankodi, A., Lungu, C., Thornton, C.A., Esson, D., Timmers, A.M., Hauswirth, W.W., Swanson, M.S. (2003) A muscleblind knockout model for myotonic dystrophy. *Science,* **302,** 1978-1980.

Kemper, K.E., Saxton, S.J., Bolormaa, S., Hayes, B.J., Goddard, M.E. (2014) Selection for complex traits leaves little or no classic signatures of selection. *Bmc Genomics,* **15**.

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

Larance, M., Ramm, G., Stöckli, J., van Dam, E.M., Winata, S., Wasinger, V., Simpson, F., Graham, M., Junutula, J.R., Guilhaus, M. (2005) Characterization of the role of the Rab GTPase-activating protein AS160 in insulin-regulated GLUT4 trafficking. *Journal of Biological Chemistry,* **280,** 37803-37813.

Larson, G., Albarella, U., Dobney, K., Rowley-Conwy, P., Schibler, J., Tresset, A., Vigne, J.-D., Edwards, C.J., Schlumbaum, A., Dinu, A. (2007) Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proceedings of the National Academy of Sciences,* **104,** 15276-15281.

Larson, G., Burger, J. (2013) A population genetics view of animal domestication. *Trends in Genetics,* **29,** 197-205.

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

Loftus, S.K., Larson, D.M., Baxter, L.L., Antonellis, A., Chen, Y., Wu, X., Jiang, Y., Bittner, M., Hammer, J.A., Pavan, W.J. (2002) Mutation of melanosome protein RAB38 in chocolate mice. *Proceedings of the National Academy of Sciences,* **99,** 4471-4476.

Maine, G.N., Burstein, E. (2007) COMMD proteins: COMMing to the scene. *Cellular and molecular life sciences,* **64,** 1997-2005.

Megens, H.-J., Crooijmans, R., San Cristobal, M., Hui, X., Li, N., Groenen, M. (2008) Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genet. Sel. Evol.,* **40,** 103-128.

Merks, J., Mathur, P., Knol, E. (2012) New phenotypes for new breeding goals in pigs. *Animal,* **6,** 535-543.

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

Müller, E., Moser, G., Bartenschilager, H., Geldermann, H. (2000) Trait values of growth, carcass and meat quality in Wild Boar, Meishan and Pietrain pigs as well as their crossbred generations. *J. Anim. Breed. Genet.,* **117,** 189-202.

Murgiano, L., D'Alessandro, A., Egidi, M.G., Crisa, A., Prosperini, G., Timperio, A.M., Valentini, A., Zolla, L. (2010) Proteomics and transcriptomics investigation on longissimus muscles in Large White and Casertana pig breeds. *J. Proteome Res.,* **9,** 6450-6466.

Ojeda, A., Huang, L.-S., Ren, J., Angiolillo, A., Cho, I.-C., Soto, H., Lemus-Flores, C., Makuza, S., Folch, J., Perez-Enciso, M. (2008) Selection in the making: a worldwide survey of haplotypic diversity around a causative mutation in porcine IGF2. *Genetics,* **178,** 1639-1652.

Okumura, N., Matsumoto, T., Hamasima, N., Awata, T. (2008) Single nucleotide polymorphisms of the KIT and KITLG genes in pigs. *Animal Science Journal,* **79,** 303-313.

Paudel, Y., Madsen, O., Megens, H.-J., Frantz, L.A., Bosse, M., Bastiaansen, J.W., Crooijmans, R.P., Groenen, M.A. (2013) Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. *BMC Genomics,* **14,** 449.

Ramírez, O., Quintanilla, R., Varona, L., Gallardo, D., Díaz, I., Pena, R., Amills, M. (2014) DECR1 and ME1 genotypes are associated with lipid composition traits in Duroc pigs. *J. Anim. Breed. Genet.,* **131,** 46-52.

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

Rheindt, F.E., Edwards, S.V. (2011) Genetic introgression: an integral but neglected component of speciation in birds. *The Auk,* **128,** 620-632.

Ropka-Molik, K., Bereta, A., Tyra, M., Różycki, M., Piórkowska, K., Szyndler-Nędza, M., Szmatoła, T. (2014) Association of calpastatin gene polymorphisms and meat quality traits in pig. *Meat Sci,* **97,** 143-150.

Rubin, C.-J., Megens, H.-J., Barrio, A.M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö., Jern, P., Jørgensen, C.B. (2012) Strong signatures of selection in the domestic pig genome. *Proceedings of the National Academy of Sciences,* **109,** 19529-19536.

Sankararaman, S., Mallick, S., Dannemann, M., Prufer, K., Kelso, J., Paabo, S., Patterson, N., Reich, D. (2014) The genomic landscape of Neanderthal ancestry in present-day humans. *Nature,* **507,** 354-+.

Schoenebeck, J.J., Hutchinson, S.A., Byers, A., Beale, H.C., Carrington, B., Faden, D.L., Rimbault, M., Decker, B., Kidd, J.M., Sood, R. (2012) Variation of BMP3 contributes to dog breed skull diversity. *PLoS genetics,* **8,** e1002849.

Tier, B., Meyer, K. (2004) Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. *J. Anim. Breed. Genet.,* **121,** 77-89.

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

Vidal, O., Varona, L., Oliver, M., Noguera, J., Sanchez, A., Amills, M. (2006) Malic enzyme 1 genotype is associated with backfat thickness and meat quality traits in pigs. *Animal* *genetics,* **37,** 28-32.

Weir, B.S., 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.

Wilkinson, S., Lu, Z.H., Megens, H.-J., Archibald, A.L., Haley, C., Jackson, I.J., Groenen, M.A., Crooijmans, R.P., Ogden, R., Wiener, P. (2013) Signatures of diversifying selection in European pig breeds. *PLoS genetics,* **9,** e1003453.

Yong, W.J., Jing, L., Jiugang, Z., Lei, C., Yonggang, L. (2012) A novel porcine gene, POT1, differentially expressed in the longissimus muscle tissues from Wujin and Large White pigs. *Cytokine,* **59,** 22-26.

7.7 Supporting information

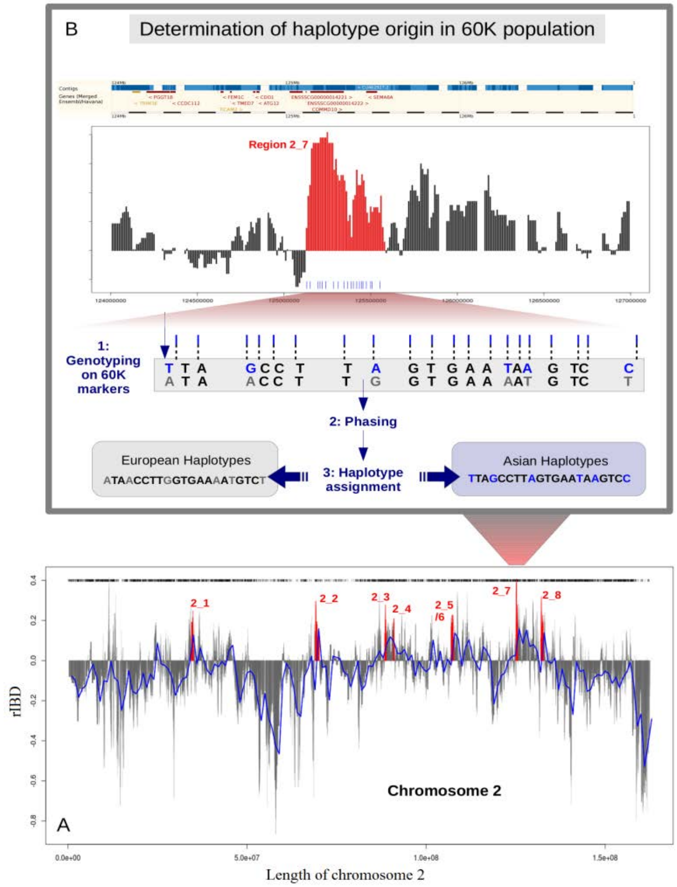

Figure 7.S1 Determination of haplotype origin in introgression regions. Chromosome 2 contains most of the introgression regions, which are indicated in red in **S1A.** The x-axis represents the full length of chromosome 2 and the y-axis contains the rIBD. The blue line displays the smoothed rIBD signal based on 1Mb bins, and the grey bars indicate the raw rIBD values for each 10kb bin. The selected regions of introgression are highlighted in red, and the location of markers on the Illumina porcine 60K beadchip are indicated as vertical bars in black above the chromosome. **S1B**. Determination of the origin of haplotypes in introgressed regions consists of 3 steps: 1) Genotyping of all 9970 individuals for the markers that cover the introgression region; 2) Phasing of haplotypes in the introgression region with the full dataset; 3) Comparison of the haplotypes in the commercial line to the haplotypes in the European population and the haplotypes in the Asian population, and assignment of the haplotypes to one of these two groups. In this example we focus on region 19 (Ssc2: 125.12-125.59 Mb) that covers the full coding sequence of the candidate gene COMMD10, as can be seen in the Sscrofa10.2 Ensembl annotation (V.76.102).

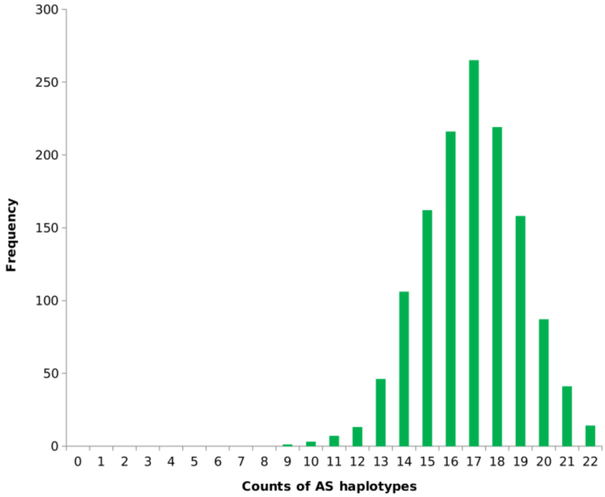

Figure 7.S2 Frequency of Asian haplotypes in introgressed regions per individual. The x-axis displays the sum of "C" alleles (=number of Asian haplotypes) that are observed for an individual, summed over all 11 regions of introgression. The y-axis contains the frequency of individuals in the Large White population that are observed to have the associated number of Asian haplotypes.

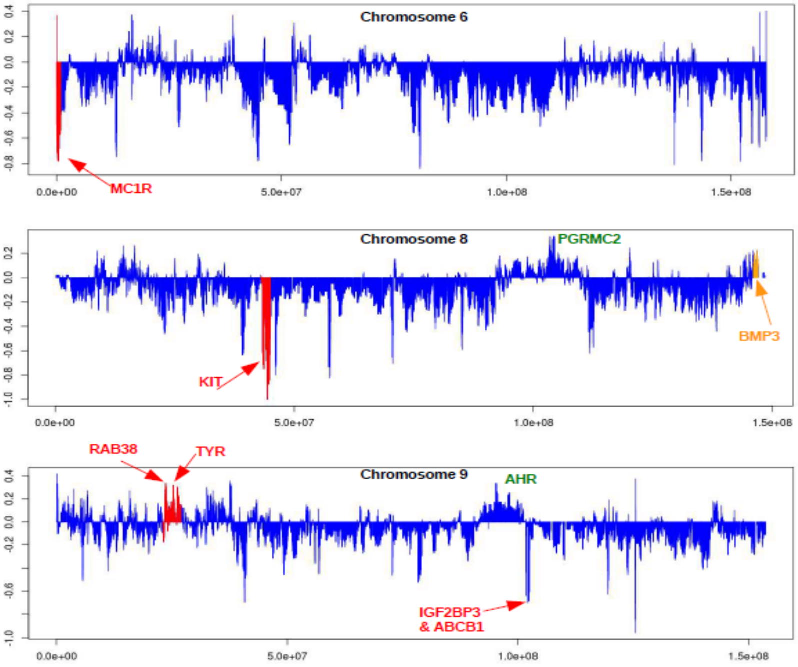

Figure 7.S3 Detailed introgression signals of chromosome 6, 8 and 9.

Table 7.S1 Regions with average introgression signal ZrIBD >2.

Region	Chr	startCore	stopCore	startEx	stopEx	lengthCore	lengthEx	nr_SNPs	p	q	o_EE	o_EC	o_CC	e_EE	e_EC	e_CC	chi_sqr
1_1	Ssc10_2_1	32.02	32.31	31.7	32.55	0.29	0.85	29	0.18	0.82	0	631	1132	56	518	1188	84
1_2	Ssc10_2_1	52.11	52.28	52.05	52.35	0.17	0.3	8	0.25	0.75	91	949	1261	139	853	1309	29
1_3	Ssc10_2_1	53.39	53.76	53.15	54.01	0.37	0.86	21	0.24	0.76	29	1016	1224	127	819	1322	130
1_4	Ssc10_2_1	84.55	84.6	84.49	84.65	0.05	0.16	7	0.46	0.54	447	1249	615	497	1149	665	18
1_5	Ssc10_2_1	86.14	86.91	85.97	87.17	0.77	1.2	23	0.35	0.65	131	1182	775	249	944	894	132
1_6	Ssc10_2_1	92.9	93.42	92.36	94.4	0.52	2.04	41	0.60	0.40	448	1055	125	585	781	261	199
15_1	**Ssc10_2_15**	**30.01**	**30.23**	**29.97**	**30.4**	**0.22**	**0.43**	**5**	**0.11**	**0.89**	**13**	**471**	**1792**	**27**	**442**	**1806**	**10**
15_2	Ssc10_2_15	59.95	60.17	59.88	60.25	0.22	0.37	6	0.51	0.49	582	1202	525	606	1154	549	4
15_3	Ssc10_2_15	97.9	98.38	97.39	98.45	0.48	1.06	13	0.41	0.59	139	1582	562	379	1102	802	433
15_4	Ssc10_2_15	124.04	124.2	123.93	124.25	0.16	0.32	7	0.30	0.70	130	1107	1073	202	962	1145	52
18_1	**Ssc10_2_18**	**23.95**	**24.21**	**23.78**	**24.56**	**0.26**	**0.78**	**5**	**0.39**	**0.61**	**329**	**1085**	**830**	**338**	**1066**	**839**	**1**
2_1	Ssc10_2_2	34.55	34.93	34.47	35.05	0.38	0.58	11	0.41	0.59	322	1188	729	375	1082	782	21
2_2	**Ssc10_2_2**	**69.11**	**69.36**	**69.04**	**70.13**	**0.25**	**1.09**	**16**	**0.44**	**0.56**	**386**	**1192**	**676**	**428**	**1108**	**718**	**13**
2_3	Ssc10_2_2	88.57	88.78	88.44	88.92	0.21	0.48	14	0.56	0.44	640	1019	392	644	1011	396	0
2_4	**Ssc10_2_2**	**90.89**	**91.02**	**90.85**	**91.07**	**0.13**	**0.22**	**7**	**0.20**	**0.80**	**74**	**779**	**1447**	**93**	**740**	**1466**	**7**
2_5	*Ssc10_2_2*	*107.03*	*107.33*	*106.97*	*107.4*	*0.3*	*0.43*	*15*	*0.24*	*0.76*	*124*	*877*	*1305*	*137*	*850*	*1318*	*2*
2_6	**Ssc10_2_2**	**107.37**	**107.44**	**107.11**	**107.71**	**0.07**	**0.6**	**19**	**0.32**	**0.68**	**230**	**1027**	**1049**	**240**	**1007**	**1058**	**1**
2_7	**Ssc10_2_2**	**125.13**	**125.36**	**125.12**	**125.59**	**0.23**	**0.47**	**15**	**0.38**	**0.62**	**276**	**1054**	**763**	**308**	**989**	**795**	**9**
2_8	Ssc10_2_2	132.34	132.58	131.74	132.86	0.24	1.12	18	0.41	0.59	157	1329	495	340	961	678	290
3_1	**Ssc10_2_3**	**31.66**	**32.11**	**31.6**	**32.18**	**0.45**	**0.58**	**15**	**0.25**	**0.75**	**130**	**798**	**1189**	**132**	**794**	**1191**	**0**
3_2	Ssc10_2_3	120.4	120.6	120.4	120.66	0.2	0.26	6	0.31	0.69	48	1300	914	215	965	1081	273
6_1	**Ssc10_2_6**	**16.05**	**16.19**	**15.97**	**16.28**	**0.14**	**0.31**	**7**	**0.01**	**0.99**	**0**	**44**	**2198**	**0**	**43**	**2198**	**0**
7_1	Ssc10_2_7	24	24.26	23.36	24.74	0.26	1.38	22	0.60	0.40	734	1027	307	752	989	325	3
7_2	Ssc10_2_7	26.67	26.93	26.62	27	0.26	0.38	10	0.24	0.76	0	1122	1185	136	849	1321	238
8_1	Ssc10_2_8	103.3	103.85	102.63	106.39	0.55	3.76	49	0.52	0.48	76	1946	7	542	1013	473	1722
9_1	**Ssc10_2_9**	**23.45**	**23.75**	**23.4**	**24.07**	**0.3**	**0.67**	**14**	**0.03**	**0.97**	**2**	**105**	**1934**	**1**	**105**	**1933**	**0**
9_2	Ssc10_2_9	37.51	37.86	37.44	37.89	0.35	0.45	62	0.57	0.43	722	1175	389	750	1118	417	6
9_3	**Ssc10_2_9**	**95.15**	**95.57**	**94.47**	**96.26**	**0.42**	**1.79**	**30**	**0.26**	**0.74**	**134**	**818**	**1101**	**143**	**798**	**1110**	**1**
9_4	*Ssc10_2_9*	*96.22*	*96.38*	*96.17*	*96.44*	*0.16*	*0.27*	*5*	*0.28*	*0.72*	*196*	*924*	*1190*	*187*	*941*	*1181*	*1*
9_5	**Ssc10_2_9**	**97.89**	**98.21**	**97.69**	**98.42**	**0.32**	**0.73**	**6**	**0.28**	**0.72**	**197**	**918**	**1189**	**187**	**938**	**1179**	**1**

"E" allele stands for a European haplotype and "C" allele stands for an Asian (Chinese) haplotype. Regions indicated in red are discarded based on their allele frequencies. Italicized regions are merged with the neighboring region and regions in bold passed the Hardy-Weinberg threshold.

Table 7.S2 1% tails of the introgression distribution, based on 1Mb bins over all autosomes.

Lowest 1%						Top 1%					
chr	Mbp	nr genes	rIBD	Region	specialties	chr	Mbp	nr genes	rIBD	Region	specialties
1	134	6	-0.522	*NA*	GLDN, DMXL2	1	54	0	0.141	1_3	*no genes*
1	314	22	-0.417	*NA*	GO: fatty acid synthesis	1	87	11	0.267	1_5	AMD1
2	58	13	-0.426	*NA*	ZNF496, NLRP3	1	94	6	0.234	1_6	ME1
2	59	21	-0.465	*NA*	ZNF496, NLRP3	2	35	1	0.126	2_1	KIF18A; Rubin 2012**
2	161	5	-0.528	*NA*	Groenen 2012*	2	70	42	0.159	**2_2**	MANY genes
2	162	14	-0.415	*NA*	Groenen 2012*	2	90	7	0.118	*(2_3, 2_4)*	Wilkinson 2013***
3	41	43	-0.563	*NA*	Groenen 2012*	2	108	6	0.137	**2_5**	IRAP
3	42	58	-0.868	*NA*	Groenen 2012*	2	126	3	0.159	**2_6**	COMMD10
3	43	44	-0.774	*NA*	Groenen 2012*	2	128	0	0.152	*NA*	*no genes*
4	49	0	-0.483	*NA*	*no genes*	2	133	1	0.137	2_7	CSNK1G3,CEP120
4	50	3	-0.445	*NA*		3	32	1	0.126	**3_1**	SNX29
6	1	22	-0.561	*NA*	MC1R	7	24	128	0.142	7_1	MHC
6	45	30	-0.510	*NA*	Many genes (CYP2 (3x))	7	25	32	0.144	*NA*	MHC
6	52	39	-0.440	*NA*	Many genes	8	104	7	0.173	8_1	PGRMC2
7	86	39	-0.441	*NA*	olfactory?	8	105	0	0.204	8_1	*no genes*
8	45	10	-0.669	*NA*	KIT close	8	106	0	0.123	8_1	*no genes*
13	101	2	-0.655	*NA*	MBNL1	8	147	10	0.127	*NA*	ANTXR2,FGF5,BMP3; Wilkinson 2013***
13	102	1	-0.518	*NA*	MBNL1	9	26	13	0.125	*(9_1, 9_2)*	TYR, NOX4,RAB38 ; Wilkinson 2013***
13	127	10	-0.415	*NA*	PIK3Ca	9	27	0	0.122	*NA*	*no genes*
14	57	0	-0.478	*NA*	*no genes*	9	96	3	0.188	**9_3**	AHR,SNX13
15	37	7	-0.462	*NA*	PTPN4	9	97	1	0.133	**9_4**	AHR,SNX13
15	45	5	-0.420	*NA*	ASB5 , spata4	9	99	3	0.136	**9_5**	TMEM196, TWISTNB
15	65	0	-0.498	*NA*	*no genes*	12	2	21	0.170	*NA*	Many (FASN)
17	15	10	-0.469	*NA*	SLC23A2, ANKRD26	15	31	1	0.138	**15_1**	ELfe3
18	1	13	-0.441	*NA*	VIPR2	18	25	0	0.157	**18_1**	POT1

*Groenen et al. 2012 indicate that this region has also been found to be selected in European wild boar. ** Rubin et al. 2012 report this region as being under selection in European domestic pigs. *** This region was found to be introgressed and selected in Wilkinson et al. 2013.

Supplementary Text 7.1

Introgression mapping

The analyses in the paper are based on previously identified regions in the genome of 9 Large White pigs that show a high proportion of Asian introgression, compared to the rest of their genomes. Although the original identification of these regions and the corresponding dataset is described in Bosse *et al.* (2014a), here we explain the methodology and rationale of the identification of these introgressed regions. Wild boars from different locations in Europe were used to represent the source of domestication, and pigs from three different Asian breeds were used to represent the pool of putative introgressed haplotypes.

Dataset and sample background

A total of 70 individual wild and domesticated pigs (*Sus scrofa)* were re-sequenced with the Illumina paired-end sequencing technology (Illumina Inc.) to ~10x depth of coverage. These individuals were divided into four functional and geographical groups; Asian wild boars (ASWB, n=8), Asian domesticated pigs (ASDom, n=13), European wild boars (EUWB, n=18) and European domesticated pigs (EUDom, n=29). Two wild boars from Sumatra were used as an outgroup. Reads were trimmed to a minimum phred quality score > 20 over three consecutive bases, with each mate having a minimal size of 45 bp after trimming. Reads were uniquely aligned to the Porcine reference genome build 10.2 with Mosaik Aligner (V. 1.1.0017). SNPs were called using Samtools mpileup 0.1.12a (r862) and filtered with VCFtools for a genotype quality of >20, a minimum read depth of 7x and a maximum read depth of twice the average read depth. Sites that were variable in at least one individual and covered >3x in all 70 pigs were extracted, resulting in 2.377.607 informative markers.

Identification of introgressed haplotypes

To identify haplotypes in the 9 Large White pigs that were shared with either European wild boars or Asian domestic pigs, we performed IBD detection using the full dataset of 70 pigs to increase phasing accuracy. Using all 2,377,607 markers, we phased each chromosome separately with Beagle fastPhase (V. 3.3.2) and used Beagle fastIBD for identifying shared haplotypes in 10 independent cycles with a IBD threshold of 5.0^{-6} (Browning and Browning (2011). We divided the haplotypes in the 9 Large White pigs into 2 groups: 1) haplotypes that were found to be IBD with European wild boar; 2) haplotypes that were IBD with Asian commercial pigs.

Introgression mapping

The shared haplotypes were mapped onto the pig reference genome as described in figure 1. The genome was divided into bins of 10 kb, and for each bin we calculated the total number of haplotypes shared with Asian commercial pigs and the number of haplotypes shared with European wild boar. To estimate the frequency of Asian haplotypes (IBDASDom) relative to European haplotypes (IBDEUWB), we normalized the total numbers of shared haplotypes with each background (ranging from 0 to 1) with a score of 1 signifying that all Large White pigs contained a haplotype with that particular background. We then calculated the relative IBD (rIBD) for the segment of 10Kb as follows:

Relative IBD between two pig groups: (rIBD = IBDASDom – IBDEUWB)

This way, the rIBD can range from -1 (all Large White haplotypes in the bin have a European background, and none have an Asian background) to 1 (all Large White haplotypes have an Asian background, and none have a European background), see also Figure 1 for graphical display.

Regions of strongest introgression

Those regions with the strongest introgression signal in the Large White population were used for further analyses. The genome-wide rIBD values were Z-transformed (ZrIBD = (rIBD – μ)/ σrIBD) and those bins exceeding ZrIBD=2 were extracted. Consecutive bins with ZrIBD>2 were merged into longer regions of introgression, so that we ended up with a total of 400 regions with a strong introgression signal.

Supplementary Text 7.2

Filtering of introgressed regions

Each region was phased independently and the origin of haplotypes (i.e. Asian or European) was determined for all individuals (see Figure 1B and methods for details). Haplotypes that were not identical to either the haplotypes found in the European group or the haplotypes found in the Asian group, were screened for their similarity to any of those haplotypes. If both groups contained closely related haplotypes in the sense that they differed the same number of nucleotides compared to the test haplotype, it received the code 'Both'. If, however, one of the two groups contained a haplotype that was more closely related, the haplotype was assigned to the closest group. The total number of haplotypes without a perfect match but with a closer relationship to either Asian or European haplotypes was 6 over all 33 regions, and the total number of individuals containing such unknown haplotype was 9. We therefore conclude that the overall frequency of these unknown haplotypes is very low and does not influence our analyses much. However, the number of unknown haplotypes with the best match in both groups is substantial, just as the number of haplotypes that had a perfect match in both groups.

Further, when the corrected number of observed haplotypes in the European and Asian reference groups differed at least by a factor 4, the haplotype was assigned to the group in which it was observed most. If not, it was assigned to the group for which both backgrounds were considered (“Both”). As it is often the case with these types of thresholds, this factor 4 is relatively arbitrary. We could have picked a lower one but decided to be relatively strict with a factor 4 difference in haplotype frequency to be confident enough about the background of the haplotype. It needs to be stated that in most cases the difference was much higher than a factor 4, and therefore the exact cut-off will not influence our analyses much (data not shown). Especially the high-frequency haplotypes were clearly polarized towards one background. Having a lower threshold will reduce the number of 'both' haplotypes in the test, and a higher threshold would have increased the proportion of 'both' haplotypes. Since miss-assignment of the background of haplotypes will most probably decrease the significance of the effect of a particular haplotype background on backfat thickness, we are confident our signal is genuine.

In 30 regions, enough markers (>2) were available to assign the origin of the haplotypes (Table S1). Two regions, 2_5 and 9_4, were in strong LD and partially overlapped with a neighboring region. In order to only have independent regions in the analysis, these regions were merged with their neighboring region. The 28 remaining regions of introgression were filtered for 1) Asian haplotype frequency and 2) Hardy-Weinberg equilibrium (HWE).

The frequency of European haplotypes was higher than 0.40 in four regions, leading to doubts about the strength of the introgression in these regions. Therefore, the four regions were removed. These high European haplotype frequencies could be due to the relatively low sample size of only 18 haplotypes that were used to identify the Asian introgression in the re-sequence data. In total, 14 regions deviated from HWE ($P<0.00001$) due to an excess of heterozygotes with one Asian and one European haplotype (Table S1). We can only speculate about the reasons for this high proportion of non-HWE regions. In a commercial pig population, we might expect deviations from HWE for particular traits due to selection. However, we might also have technical difficulties in the phasing step for these regions, assigning composite haplotypes to either groups because the region spans a longer stretch of DNA than some haplotypes. Another potential reason for an excess of heterozygotes is that the region is copy number variable or is duplicated, leading to false heterozygous signals. Therefore, these regions were excluded from the association analysis, even though the signal could be genuine. Finally, two regions were removed because they were fixed for the Asian haplotypes. Even though these regions are biologically interesting, they could not be used for the association analysis because segregation of haplotypes is required. In total, 11 regions passed both thresholds (Asian haplotype frequency and HWE) and were further used in the association analyses (see Table S1 and Methods for details).

Supplementary Text 7.3

Non-synonymous mutations in introgressed regions

6_1 ME1 - Looking at the genotypes in our re-sequence based matrix for region 1_6, we see that especially in the 3' region of the *ME1* gene, the European domestic pigs have alleles not found in the European wild boar, but present in the Asian breeds. The 3'UTR region contains an SNP that appears fixed in the European wild population (T) but the European domestic population has an alternative variant 2:c.*753T>C. Also, a non-synonymous mutation (H/N; rs80816302) can be found at position aa347 in the protein that is the derived allele in Asia and is in moderate frequency in the Large White population, but absent in the European wild population.

2_6 *LNPEP and CAST -* Two of these genes, *LNPEP* (rs344711695) and *CAST* (rs333184969 and rs45432488) had non-synonymous mutations of putative Asian origin in multiple Large White pigs. The S/L mutation at aa334 in *LNPEP* is predicted as benign. Both mutations in *CAST*, R/S at aa728 and K/R at aa 249 or 339 in the protein were predicted to have no deleterious effect on the protein.

2_7 *COMMD10 -* We found a benign non-synonymous mutation at aa20 in the protein (S/L, variation name rs325242824) of Asian origin in the second exon that segregates in the European commercial pigs, but is absent in the European wild boar and closely related species.

Supplementary Text 7.4

Bias in the selection of the regions

However, it is important to notice that these 11 regions cover a small portion of the full genome, and therefore the genes or other regulatory elements that are present in these regions may not be a representative sample for the full genome. Therefore, of biggest concern here is whether a potential bias exists in our selection criteria that may have promoted the selection of regions that have an effect on BF, other than the background of the haplotypes that are present. A potential bias can be introduced due to recombination frequency or gene content. As we have shown in Figure 2, they have only a very modest effect on the overall introgression signal. The criteria we used to select these regions were solely lead by the strength of introgression and the usability of the regions for running our statistical test. We have used rather stringent criteria that are usually used for single SNPs rather than haplotypes. The effect of these regions on BF is, therefore, an independent factor that we have only tested after the filtering. However, we cannot completely rule out the possibility that the selection criteria are in any way linked to the contribution of a region to BF. Although unlikely, the number of SNPs in a certain part of the genome might be higher in regions with a relatively strong potential effect on BF, because these SNPs have been identified in previous BF-related studies and therefore have been included on the 60K SNPchip. But during the design of this chip, an even distribution of SNPs over genomic regions has been one of the priorities, and therefore we are confident that such bias is minor, and will certainly not be in the range of the significance we have demonstrated in our statistical tests.

Supplementary Text 7.5

Theoretical exercise

We have opted for not to perform a false discovery rate or apply a Bonferroni test to validate our association results because we do not have a large number of tests (we performed the association analysis on a limited number of introgressed regions). However, we believe that an additive effect of 0.09 mm of backfat (when analyzing all 11 regions together) is more than expected by chance. To empirically generate the null distribution (H_0: frequency of introduced AS haplotypes have increased solely due to drift) we would need to take a random set of 11 regions of the genome and test the association of the AS haplotypes with increased BF. The problem is that we should do this in the original hybrid population, before drift and selection have taken place. If we take a random set of regions in the current dataset, for most of the regions we cannot do an association analysis because we do not have AS haplotypes in those regions. What we need is the distribution of allele substitution effects (EU *vs* AS allele) across all positions of the genome. This value will be larger than 0, but presumably, it will be smaller than 0.09 mm. The problem is that we can only measure this substitution effect in the introgressed regions. To get an indication of how likely our result is to be obtained by chance, we performed a theoretical exercise. Assume that the BF QTL are evenly distributed over the genome. We estimate the probability of sampling six real QTL if we randomly choose 11 regions across the genome. This probability was estimated as follows:

$$\text{probability=} \left(\frac{nr_{QTL}}{nr_{regions}}\right)^6 * \left(\frac{nr_{regions}-nr_{QTL}}{nr_{regions}}\right)^5 * \frac{N!}{(N-k)!\,k!}$$

$$nr_{QTL} = \frac{BF_{asian} - BF_{european}}{2 * \alpha}$$

$$nr_{regions} = \frac{L_{genome}}{\bar{L}_{regions}}$$

where nr_{QTL} is the expected number of QTL that affect the trait, $nr_{regions}$ is the number of regions of the whole genome with the average length of the introgression regions, *N* is the number of regions randomly chosen (*N*=11), and *k* is the number of QTL that are included in the *N* random regions (*k*=6). Further, BF_{asian} is the average BF observed in Asian breeds, $BF_{european}$ is the average BF observed in European breeds ($BF_{asian} - BF_{european} = 19\,mm$, Müller *et al.* 2000), α is the allele substitution effect (0.09 mm from our study), L_{genome} is the length of the genome (2,808,525,991 bp) and $\bar{L}_{regions}$ is the average length of the introgressed regions (697,273 bp from our study). Therefore, if we choose randomly 11 regions of 697,273 bp out of the total of 4,028 regions, the probability that six of the 106 expected QTLs are included is 1.34 x 10^{-7}. With such a low probability, we expect that the associations found in this study are true association rather than association by chance.

8

General discussion

8.1 Introduction

The recent and fast developments of genomics regarding dense Single Nucleotide Polymorphisms (SNPs) panels and sophisticated statistical models based on genomic information have enabled the inclusion of genomic information in the genetic evaluations of several livestock species (Hayes *et al.* 2009; Sonesson & Meuwissen 2009; Lillehammer *et al.* 2013). With the implementation of genomic selection in pig breeding, the genetic progress per year in purebred populations is expected to increase up to 55% compared to traditional pedigree-based selection (Lillehammer *et al.* 2013). However, as most animals in the pork production system are crossbreds, the increase in genetic progress in purebreds will only be observed on production farms if this progress is expressed in the performance of crossbreds.

The main goal of the research reported in this thesis was to evaluate different models based on genomic information that can contribute to improve the performance of crossbred animals. Another aim was to gain insight into the genetic architecture of the evaluated (complex) traits and to investigate how selection history influenced haplotype patterns of current commercial pigs. The results reported in this thesis showed that: **chapter 2**) dominance effects account for a considerable proportion of the phenotypic variance in several traits while the contribution of imprinting effects is limited; **chapter 3**) a model that accounts for both additive and dominance effects simultaneously increases the prediction accuracy of phenotypes; **chapter 4**) genome-wide association studies (GWAS) can identify QTL with dominance effects in addition to QTL with additive effects; **chapter 5**) a model that accounts for GWAS findings in the genetic evaluation improves the prediction of phenotypes; **chapter 6**) breed-specific effects may play a role in the genetic variation of crossbred performance; when predicting crossbred performance, a model that accounts for breed-specific effects results in similar or higher prediction accuracies compared to predictions based on traditional genomic selection models; and **chapter 7**) the majority of the introgressed Asian haplotypes found in a modern European pig breed were associated with backfat, which indicates that human-driven introgression and selection may have shaped the genomic composition of commercial pig breeds.

In this general discussion, I have discussed the results of the previous chapters in a broader perspective. I focused on the practical application of these results in pig breeding programs with the specific emphasis on the improvement of crossbred performance. I have also discussed issues related to the interpretation of results from current models based on genomic information and the challenges and opportunities for the future of pig breeding.

8.2 Dominance effects

The results presented in this thesis showed that dominance effects account for a large proportion of the total genetic variance (up to 44%) for several traits in distinct pig populations. The proportion of dominance variance relative to additive variance varied considerably across traits and when evaluating the same trait in different populations (**chapters 2** and **3**). The accuracy of predicting phenotypes using total genetic values from a model that account for both additive and dominance effects (MAD) was higher compared to using breeding values from a model that accounts only for additive effects (MA), especially in populations where high dominance variance was detected (**chapter 3**). The results presented in this thesis show the relevance of dominance effects, give insight into the genetic architecture of the evaluated traits, and also allow a better prediction of phenotypes. However, dominance estimates need to be carefully interpreted depending on the model used in the analysis (genotypic model vs breeding model) because their estimated parameters require a different interpretation.

8.2.1 Breeding model *vs* genotypic model

The partitioning of the genetic variance was performed using a genotypic model in **chapter 2** and a breeding model in **chapter 3**. Although the genotypic model and the breeding model are equivalent models, their estimated parameters require a different interpretation (Vitezica *et al.* 2013). Using the genotypic model, the analyses were performed by recoding the genotypes (GG, GC, CC) as (-1, 0, 1) for the additive component, and (0, 1, 0) for the dominance component. With this, the estimated SNP effects are the additive effects (a) and dominance effects (d), as described by Falconer and Mackay (1996). Thus, when we evaluate *var*($\mathbf{A}\boldsymbol{a}$) at each iteration of the Gibbs sampler for obtaining the total variance of the additive component, $\mathbf{A}$ being the design matrix for the additive effects and $\boldsymbol{a}$ the vector of additive effects (average difference between homozygotes), it implies that the additive genetic variance (σ_a^2) of a single SNP is estimated as $\sigma_a^2 = 2pqa^2$, where p and q are the allele frequencies of the evaluated SNP. When the breeding model was applied, the transformation proposed by Vitezica *et al.* (2013) was followed and therefore the genotypes (GG, GC, CC) were recoded as ($0-2p$, $1-2p$, $2-2p$) for the additive component, and ($0-2p^2$, $2pq$, $2-2q^2$) for the dominance component. With this, the estimated SNP effects are the allele substitution effects (α) and dominance deviations (δ). In this scenario, when we evaluate *var*($\mathbf{A}\boldsymbol{\alpha}$) at each iteration of the Gibbs sampler for obtaining the total variance of the additive component, $\boldsymbol{\alpha}$ being a vector of allele substitution effects, it implies that σ_a^2 of a single SNP is estimated as $\sigma_a^2 = 2pq\alpha^2$, where $\alpha = a + d(q - p)$, as described by Falconer and Mackay (1996).

Therefore, the main practical difference between genotypic and breeding model is that in the first, dominance effects (d) do not contribute to the additive genetic variance ($\sigma_a^2 = 2pqa^2$), while in the later they contribute partly ($\sigma_a^2 = 2pq[a + d(q - p)]^2$). This explains why, in **chapter 2**, a decrease in the total additive genetic variance was observed when the MA model was replaced by the MAD model. In other words, moving from MA to MAD, applying a genotypic model, the dominance variance that is "heritable" (contributes to the additive variances) shifts to the dominance variance. Because of this shift, the variance estimates from genotypic models are not directly comparable to pedigree-based estimates. Applying the breeding model, moving from MA to MAD does not influence the additive variance estimates (as shown in **chapter 3**) because the mentioned shift of "heritable" dominance variance does not occur. Therefore, if the aim is to estimate breeding values and dominance deviations, the breeding model should be applied. However, if the aim is to quantify the contribution of dominance effects to the additive variance (measured as the decrease of additive variance when moving from MA to MAD), the genotypic model should be applied.

Based on the different interpretation of the estimates from the different models, I conclude that the results presented in **chapter 2** (genotypic model) contribute to the general knowledge about the contribution of additive and dominance effects to the total genetic variance of several traits. These results improve our understanding of the genetic architecture of the evaluated quantitative traits and should be considered when evaluating the relevance of dominance effects for the traits included in the breeding goal. On the other hand, the results presented on **chapter 3** (breeding model) contributes to a better understanding of the potential use of dominance effects in breeding programs. These results illustrate how much gain in prediction accuracy can be achieved when phenotypes are predicted using total genetic values (sum of breeding value and dominance deviations) compared to using only breeding values.

8.2.2 Practical application of dominance models

The results presented in this thesis showed that dominance effects are relevant for pig breeding, and that accounting for dominance effects yields more accurate prediction of phenotypes, especially for traits with high dominance variance. However, as these results also showed that the amount of dominance variance varies considerably across traits and populations, the MAD model will not be required for all traits in all populations. This is because for traits with low or no dominance variance, the application of the MAD model will not improve the prediction of phenotypes, although it would also not hamper the predictions (Toro & Varona 2010; Zeng *et al.* 2013).

Based on the state-of-the-art of current genetic evaluation in pig breeding, replacing MA by MAD would be challenging. Currently, the genetic evaluations in the major pig breeding companies are based on the so-called "single-step" approach (Legarra *et al.* 2009; Misztal *et al.* 2009; Christensen & Lund 2010). With the single-step, the additive genetic relationship between relatives is accounted for via an **H** relationship matrix, which combines genomic-based and pedigree-based additive genetic relationships between genotyped and non-genotyped animals. This combination of pedigree-based and genomic-based relationships would be a problem when it comes to dominance relationships. Using dominance relationships estimated from pedigree information, the estimation of dominance variance and dominance deviations is not performed accurately (Vitezica *et al.* 2013; Zhu *et al.* 2015). Therefore, applying a "dominance **H** matrix" in the genetic evaluations would be of low added values if a large proportion of the evaluated population is not genotyped.

Another challenge for the application of MAD models would be the increase in running time of the genetic evaluations. In pig breeding, this would be especially important because genetic evaluations are typically performed at a daily basis. Thus, any increase in running time needs to be evaluated carefully to not compromise the schedule that needs to be followed from the start to the end of the genetic evaluations. Therefore, if dominance is relevant only for a few traits and populations, the benefits of applying MAD models may not be worth the increase in running time or the investments in computer capacity needed to overcome it.

Further, as shown in **chapter 3**, the application of MAD was only beneficial for predictions of phenotypes using total genetic values. Predicting phenotypes using the breeding values from both MA and MAD models resulted in the same prediction accuracy. As discussed in the previous section, in breeding models the dominance effects that contribute to the allele substitution effects are already accounted for even when the MA model is applied. Therefore, using the MAD model does not contribute to a better estimation of the breeding values. This means that using the MAD model for estimating breeding values for crossbred performance will not bring an advantage compared to traditional models.

In my opinion, the benefits of using dominance models will come from mate allocation. In this scenario, additive and dominance effects preferably estimated in a crossbred training population would be used to estimate the total genetic values of the purebred breeding animals. Then, after applying a traditional selection based on breeding values, the average total genetic value of purebred parents would be used to predict the average performance of their crossbred offspring. In this scenario, we will be looking for the best sire-dam combinations which result in the best-performing litters. Using simulated data, Toro and Varona (2010) predicted that accounting for dominance effects using mate allocation in addition to traditional selection based on breeding values increased the genetic level up to 22%.

8.2.3 QTL with dominance effect

In **chapter 4**, a GWAS for additive and dominance effects on number of teats was performed to identify genomic regions which show dominance variance and to investigate the importance of dominance using a high-density SNP panel in a Landrace-based population of pigs. The results presented in **chapter 4** showed that it is possible to identify dominant QTL fitting the SNP in the model as a class effect instead of regression, as it is usually done when searching only for additive QTL. The next step would be to apply the same methodology for other traits and populations (including crossbred populations). If QTLs with large dominance effects are found, they could be used in the marker-assisted BLUP and GBLUP models (MA-BLUP and MA-GBLUP, respectively) as presented in **chapter 5** for QTLs with large additive effects. Using QTLs with dominance effects under MA-(G)BLUP models would be especially interesting in situations where the phenotype in addition to breeding value is relevant for animals located in any part of the breeding pyramid, such as disease resistance related traits.

8.3 Contribution of single genes to prediction accuracy

The results presented in **chapter 5** showed that accounting for GWAS findings in marker-assisted models (MA-BLUP and MA-GBLUP) results in increased prediction accuracies compared to traditional models (BLUP and GBLUP). This approach benefits from the linkage disequilibrium (LD) between SNPs and the QTL as well as from the realized family relationship from GBLUP.

In the near future, I strongly believe that marker-assisted models (especially MA-GBLUP) will be extensively exploited in breeding programs. Marker-assisted models will become especially interesting when GWAS based on whole-genome sequence data start to be performed, which may lead to a more accurate identification of QTLs. The use of GWAS findings based on whole-genome sequence data is promising because in this case some markers might be in full LD with the causal mutation or they are even the causal mutations themselves. In fact, the use of the GWAS findings in marker-assisted models might be the major way of benefiting from whole-genome sequence data in genomic predictions. Recent studies have shown that using sequence data through GBLUP (i.e. to build the **G** matrix) does not improve the prediction accuracies compared to high-density SNP panels (Pérez-Enciso *et al.* 2015, van Binsbergen *et al.* 2015). Although these results are disappointing, this is in line with other studies which showed that using about 3,000 markers is already enough for building an accurate **G** matrix to be used in GBLUP (Rolf *et al.* 2010; Lopes *et al.* 2013).

For practical application, I envision that genetic evaluations in a near future can potentially be performed using customized SNP chips that will include SNPs from commonly-used SNP chips, such as the 60K SNP chip (Ramos *et al.* 2009), and significant SNPs identified in GWAS using whole-genome sequence data. In this scenario, MA-GBLUP will be applied using a **G** matrix built using all SNPs from the customized SNP chip and including the SNPs associated with the target traits as fixed effects in the model, as described in **chapter 5**. A similar approach has been recently described by Brøndum *et al.* (2015) for dairy cattle. These authors performed GWAS using whole-genome sequence data and selected between three and five SNPs per QTL region per trait. Further, they included all selected markers (N=1,623) in a low-density SNP chip that was used in combination with the commonly-used 54K SNP chip to estimate breeding values. In the genomic evaluations, they applied a model including two **G** matrices: one based on the markers from the 54K SNP chip, and the other based on the significant markers from the whole-genome sequence data. As a result, Brøndum *et al.* (2015) reported that the reliability of the breeding values increased up to five percentage points (using two **G** matrices) compared to traditional GBLUP (using only one **G** matrix).

Although the approach proposed by Brøndum *et al.* (2015) is interesting and showed an increase in the reliability of the breeding values, I expect that selecting only the most significant SNP per QTL region and including this SNP as a fixed effect in the model would result in higher prediction accuracies compared to using a second **G** matrix. This is because when including SNPs as fixed effects in the model a specific set of SNPs will be used per trait, which gives a higher weight to each marker with large effect. On the other hand, building a second **G** matrix implies that all SNPs for all traits are analyzed together under the assumption that all markers (including those not associated with the target trait) will explain the same proportion of the variance of all traits, which may limit the effect of markers associated with the target trait.

The approach described above will only be possible when thousands of animals are sequenced. Hickey (2013) suggested that within five years the major breeding companies would have sequence information available on a million of animals. I think this prediction is too optimistic because to date availability of sequence data is still limited in most livestock species (especially in pigs). While waiting for sequence data on at least a few thousands of pigs, the benefits from marker-assisted models will come from using markers linked to QTL identified using the commonly-used 60K SNP chip. However, as already discussed in **chapter 5**, SNPs with large effect will not be identified for all traits. In situations like this, the application of traditional GBLUP is likely to be sufficient to obtain most of the advantages from genomic data for prediction.

8.3.1 Predicting crossbred phenotypes using markers with large effect

In **chapter 5**, using breeding values from marker-assisted models (MA-BLUP and MA-GBLUP) for predicting phenotypes of purebred animals resulted in higher prediction accuracy compared to using breeding values from traditional models (BLUP and GBLUP). As the QTL regions used in these analyses showed quite a large effect (accounted for up to 6% of the total phenotypic variance), I also expected that prediction of crossbred phenotypes using the breeding values of their purebred parents would yield similar results. To test this hypothesis, I evaluated a dataset with information on number of teats of crossbred animals (F1) obtained from the cross between the Large White and the Dutch Landrace population evaluated in **chapter 5**. This dataset consisted of 51,423 crossbred animals that were born between 2013 and 2015 (Table 8.1). The prediction accuracy was measured as the correlation between the breeding value of the purebred parents and the average phenotype of their F1 offspring. When both parents of the F1 were genotyped, the correlation was estimated between the average breeding value of the parents and the average phenotype of their litter. The phenotypes of the F1 animals used in these analyses were pre-adjusted phenotypes that were obtained in the same way as described for the purebred populations in **chapter 5**. The breeding values of the purebred animals from the marker-assisted and traditional models were obtained from the analyses from **chapter 5**.

As shown in Table 8.1, the lowest prediction accuracies were obtained when using the breeding values from BLUP and the highest when using the breeding values from MA-GBLUP. This is in line with the results presented for the purebred populations. Therefore, if markers with large effect exist and are identified, the use of these markers in MA-BLUP or MA-GBLUP has potential to improve the accuracy for selecting purebreds for crossbred performance, compared to using traditional models.

Table 8.1 Correlation between breeding values of purebred (PB) parents and the average phenotype of their crossbred (CB) offspring.

Population	Nr. PB parents	Nr. CB offspring	BLUP	MA-BLUP	GBLUP	MA-GBLUP
Large White	197	34,869	0.284	0.318	0.380	0.418
Dutch Landrace	362	11,352	0.289	0.296	0.307	0.327
Parental average	406*	5,202	0.380	0.408	0.398	0.438

*Number of unique sire-dam combination.

8.4 GWAS methodologies and their interpretation

The results from **chapter 5** showed that the prediction of phenotypes was more accurate when accounting for GWAS findings in marker-assisted models (MA-BLUP and MA-GBLUP) compared to traditional models (BLUP and GBLUP). For a successful application of the marker-assisted models, the identification of SNPs truly associated with the target phenotype is required. Thus, the use of an appropriate GWAS methodology is essential.

In **chapter 5**, a single-SNP GWAS was performed. This GWAS methodology consists of fitting one SNP at a time in the model (normally as a fixed effect) to obtain a *P* value for each SNP, which will be used to measure the strength of evidence for association. The use of *P* values has been described as a limitation of this GWAS methodology (Sham & Purcell 2014). The interpretation of these P values is difficult, as we need to take into consideration the statistical power of the analysis (e.g. minor allele frequency and sample size), as an insignificant test can indeed indicate the absence of an effect or an inadequate statistical power (Sham & Purcell 2014). However, even with its limitations, single-SNP GWAS has been widely used and has resulted in the successful identification of QTL regions (Stephens & Balding 2009; Li *et al.* 2011).

Bayesian GWAS is an alternative methodology to identify QTL regions in which all SNPs are included in the model simultaneously (not implying that all SNPs should necessarily have an effect). This methodology is expected to alleviate the limitations of *P* values at the cost of additional modeling assumptions (Stephens & Balding 2009). Bayesian GWAS requires explicit assumptions about effect sizes of the associated SNPs, which is not straightforward. These assumptions will vary considerably according to the Bayesian model chosen out of the wide Bayesian alphabet range (Gianola *et al.* 2009; Habier *et al.* 2011). In Bayesian GWAS, the strength of evidence for association are frequently measured in terms of Bayes factors (BF, e.g. Duijvesteijn *et al.* 2014 and Legarra *et al.* 2015), but QTL regions have also been pointed out based on the variance explained by a group of *n* consecutive SNPs (SNP window, e.g. Onteru *et al.* 2012). The BF are defined as the ratio of the probability of the data under the null hypothesis and the alternative hypothesis (Wakefield 2009). The variance explained by SNP windows is a measure of the variation explained by chromosome fragments as a proportion of total genetic variance. Evaluating the variance explained by an SNP window instead of the effect of a single SNP was described to be more efficient for pointing out QTL regions because the effect of a QTL may be distributed across many SNPs in LD with the QTL (Onteru *et al.* 2012; Hayes *et al.* 2013). With this, the effect of an individual SNP will tend to underestimate the real effect of the QTL (Onteru *et al.* 2012) and no clearly visible peak will appear in the GWAS plot.

Another alternative for identifying QTL regions is to estimate SNP effects from the solutions of GBLUP, hereafter called "backsolving GWAS". With this methodology, genomic breeding values from GBLUP are backsolved to SNP effects, as described by Wang *et al.* (2012) and applied in **chapter 6** as well. When this GWAS methodology is applied, no strength of evidence for association, such as *P* values and BF, is given. The QTL regions are pointed out according to the variance explained per SNP. Wang *et al.* (2012) have described this methodology as ssGWAS because they estimated SNP effects from breeding values from a single-step GBLUP (ssGBLUP). The advantage of backsolving GWAS is the ability to combine genomic evaluations and GWAS. Right after the estimation of the breeding values, the SNP effects can be estimated without the need of estimating pseudo-phenotypes that in a second step will be included in the association analyses (single-SNP or Bayesian GWAS, e.g. Diniz *et al.* 2014; Duijvesteijn *et al.* 2014; Sevillano *et al.* 2015). In the backsolving GWAS, all SNPs are analyzed simultaneously. However, it assumes that all SNPs explain the same amount of the phenotypic variance.

In this section, my aim was to compare the results of the GWAS performed in **chapter 5** (single-SNP GWAS) with the results of a Bayesian GWAS and a backsolving GWAS. For these comparisons, I performed the two alternative GWAS using the data from the Norwegian Landrace population (6,072 animals genotyped for 38,085 SNPs).

8.4.1 Bayesian GWAS

The Bayesian GWAS was performed fitting all SNPs simultaneously in a Bayesian variable selection model (George & McCulloch 1993):

$$\boldsymbol{y} = \mathbf{1}\mu + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector of pre-corrected phenotypes (see more detail in **chapter 5**), μ is the mean number of teats, $\mathbf{Z}$ is a design matrix with SNP genotypes coded as 0, 1, or 2 copies of a given allele, $\boldsymbol{\beta}$ is a vector of unknown SNP effects, $\boldsymbol{e}$ is a vector of random residual effects assumed to be normally distributed $N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance and $\mathbf{I}$ is an identity matrix. A Bernoulli distribution was applied on the allele substitution effects:

$$\boldsymbol{\beta} \sim \begin{cases} N(\mathbf{0}, \mathbf{I}\sigma_{\mathrm{g0}}^2) \text{ with probability: } \pi_0 \\ N(\mathbf{0}, \mathbf{I}\sigma_{\mathrm{g1}}^2) \text{ with probability: } \pi_1 = 1 - \pi_0 \end{cases}$$

where the first distribution is the null distribution which contains SNPs that are expected to explain a small proportion of variance (σ_{g0}^2), and the second distribution contains SNPs that are expected to explain a large proportion of variance (σ_{g1}^2) of the trait. The probability to be in the null distribution (π_0) was set to 0.999, meaning only one in every 1,000 SNPs will be in the second distribution, which is on average 38 SNPs per cycle. The Bayesian variable selection model was implemented in the program Bayz (http://bayz.biz/). A total of 250,000 MCMC chains with a burn-in of 50,000 cycles were run and a Metropolis-Hastings sampler was applied to obtain good convergence. The level of significance of the SNPs was determined by evaluating the BF of each SNP and the variance of SNP windows (five consecutive SNPs). The BF was calculated as an odds ratio:

$$BF = \frac{\hat{b}_i/(1 - \hat{b}_i)}{\pi_1/\pi_0}$$

where π_0 and π_1 were the prior probability and $\hat{b}_i$ the posterior probability. The QTL regions were identified by visual inspection of GWAS plots. The BF was plotted against the physical position of the markers in the genome. The variance of the SNP windows was plotted against the average physical position of the markers included in the SNP window.

8.4.2 Backsolving GWAS

The breeding values used in the backsolving GWAS were estimated applying the following GBLUP model in ASReml (Gilmour *et al.* 2009):

$$\boldsymbol{y} = \mathbf{1}\mu + \mathbf{X}\boldsymbol{a} + \boldsymbol{e}$$

where $\boldsymbol{y}$, $\mathbf{1}$, μ and $\boldsymbol{e}$ were as previously defined, $\mathbf{X}$ is an incidence matrix for additive genetic effects, $\boldsymbol{a}$ is a vector of additive effects assumed to be distributed as ~$N(\mathbf{0},\mathbf{G}\sigma_a^2)$, where σ_a^2 is the additive genetic variance and $\mathbf{G}$ is a genomic additive relationship matrix accounting for the (co)variances between animals due to relationships. The $\mathbf{G}$ matrix was built according to VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{MM}'}{\sum_{i=1}^{n} 2p_i q_i}$$

where $\mathbf{M}$ is a matrix of centered genotypes, and p and q are the allele frequencies of the SNPs. The backsolving GWAS was performed as described by Wang *et al.* (2012) by estimating SNP effects $\widehat{\boldsymbol{a}}$ as follows:

$$\widehat{\boldsymbol{a}} = \frac{\mathbf{M}'\mathbf{G}^{-1}\widehat{\boldsymbol{u}}}{\sum_{i=1}^{n} 2p_i q_i}$$

where $\widehat{\boldsymbol{u}}$ is a vector of estimated breeding values. After estimating the SNP effects, the variance of each SNP was estimated ($\hat{\sigma}_{a_i}^2 = \hat{a}_i^2 2p_i q_i$) and plotted against the physical position of the markers in the genome.

8.4.3 Comparison of GWAS methodologies

Figure 8.1 shows the GWAS plots obtained from all GWAS methodologies. As already described in **chapter 5** (single-SNP GWAS), the most significant SNP (lowest *P* value) was observed on chromosome 7 (Figure 8.1 A). For the Bayesian GWAS, the most significant signals varied depending on how the results were reported. When BF for each SNP was plotted (Figure 8.1 B), the most significant SNP (highest BF) was observed on chromosome 12, while the highest SNP window variance (Figure 8.1 C) was observed on chromosome 7. For the backsolving GWAS (Figure 8.1 D), the SNP that explained the highest proportion of phenotypic variance was observed on chromosome 7 as well.

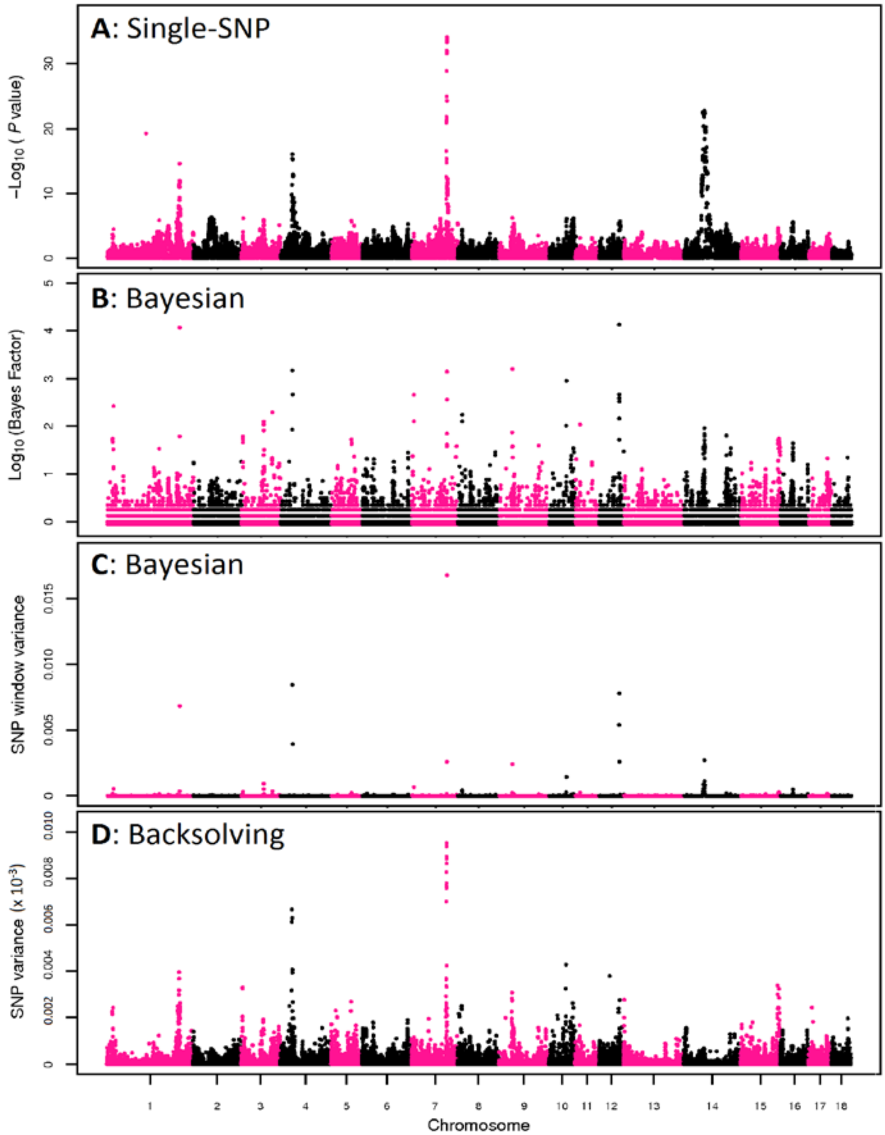

Figure 8.1 GWAS plot from different GWAS different methodologies. The location of an SNP window is given as the average position of the SNPs in this window.

From all GWAS plots, the least clear one was obtained when BF were used to report the Bayesian GWAS. From the same analysis (Bayesian GWAS) quite different interpretations can be made depending on how the results are represented (BF or SNP window). While the plots of BF was not clear, the plots of

SNP window variance showed clear peaks and in most cases these peaks were in concordance with those from single-SNP and backsolving GWAS. As previously described, the advantage of reporting variances of SNP windows instead of reporting effects of single SNPs is because the effect of a QTL is distributed across many SNPs in LD when all SNPs are included in the model simultaneously (Onteru *et al.* 2012). Therefore, with SNP windows, the effect of QTL becomes more visible because effects of SNPs that are linked to the same QTL are put together.

Except the plot of BF, all other plots clearly show the QTL region on chromosome 7, which was expected due to the effect of this QTL on number of teats (as discussed in **chapter 5**). The major difference between GWAS methodologies was regarding the QTL region observed on chromosome 14 from the single-SNP analysis, and the QTL region on chromosome 12 from the Bayesian GWAS. Within the QTL region of chromosome 14, gene EDAR-associated death domain (*EDARADD*) is annotated. This gene mediates signaling of the ectodysplasin receptor (Morlon *et al.* 2005), which is required for the normal development of mammary glands (Thesleff & Mikkola 2002). Therefore, the presence of candidate gene with a biological function related to the evaluated trait within the QTL region observed on chromosome 14, gives some confidence that this is a real QTL rather than a false positive. Then, the question is why the peak on chromosome 14 is not pronounced in this region when the alternative GWAS methodologies were applied (fitting all SNP in the model simultaneously).

My hypothesis is that the QTL region on chromosome 14 was not observed because of the extent of LD in this region. This is because SNPs with a -$\log_{10}$(*P* value) >10 (e.g. highly significant) were spread from 44 to 62 Mb. Therefore, when all SNPs are fitted in the model simultaneously (Bayesian and backsolving GWAS) the effect of the putative QTL in this region is spread across the SNP in this large region, not being clearly detected in the GWAS plot. To test this hypothesis, I proposed to plot the variance per LD block instead of per SNP window (Bayesian GWAS) or per SNP (backsolving GWAS).

The LD blocks were estimated using Haploview (Barrett *et al.* 2005) as described by Veroneze *et al.* (2013). Following my expectation, in the QTL region on chromosome 14, the three largest LD blocks of this population were identified with >125 SNPs each. With the knowledge on the LD blocks across the genome, I could then estimate the variance for each LD block based on the results from the Bayesian GWAS and the backsolving GWAS. Then, the variance of the LD block was plotted against the average physical position of the markers included in this LD block (Figure 8.2). With these new plots, a pronounced peak on chromosome 14 could also be observed from the Bayesian GWAS (Figure 8.2 A) and backsolving

GWAS (Figure 8.2 B), indicating that this QTL might be a true association. In addition, it confirms my expectation that this QTL was not clearly observed in Figure 8.1 (B, C, and D) because of the LD patterns in this region. Further, with the plots of LD block variance, the results of the Bayesian and backsolving GWAS were quite similar, except for the peak on chromosome 12, which was only clearly pronounced from the Bayesian GWAS. Why this QTL region on chromosome 12 is only pronounced when the Bayesian GWAS is applied is still unclear. One reason could be due to the difference in detection power of the three GWAS methodologies, but it still needs to be further evaluated for a more conclusive answer.

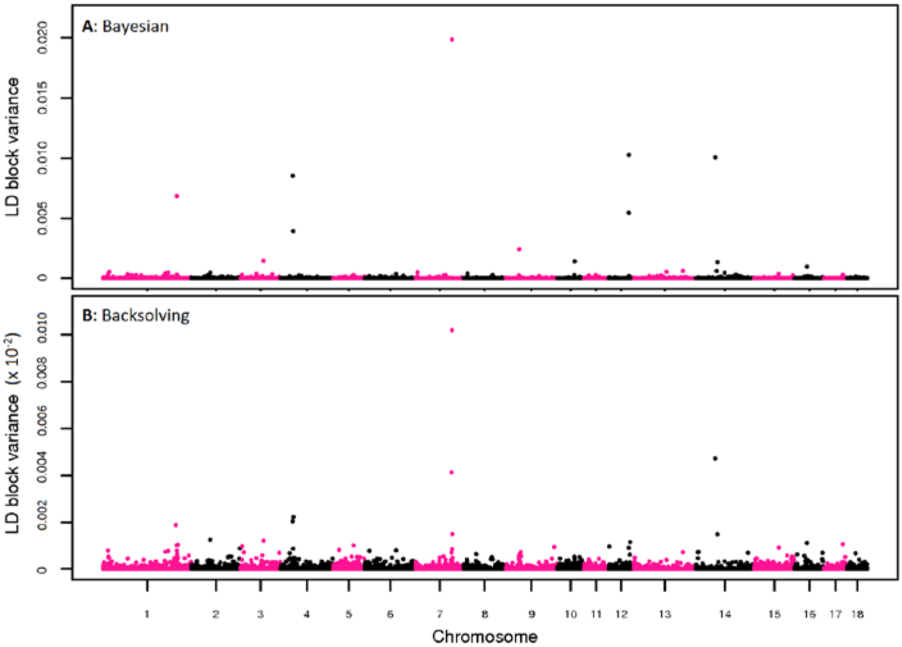

Figure 8.2 GWAS plot from the Bayesian (A) and backsolving (B) GWAS plot of the LD block variance against the physical position of the linkage disequilibrium (LD) block. The location of an LD block is given as the average position of the SNPs in this LD block.

Based on the results and evidence here presented, I conclude that all GWAS methodologies, in general, give the same answers depending on how the results are presented. These results indicate as well that reporting Bayesian GWAS in terms of BF is not the best option. Further, when the LD extent in the evaluate population is large and Bayesian GWAS or backsolving GWAS are applied, the results should be reported in terms of LD blocks variance.

8.5 Origin of alleles

The genome of crossbred animals is formed by the combination of alleles inherited from the parental breeds that may be expressed differently according to their parental or breed origin. Thus, for making an accurate prediction of crossbred performance, a better understanding of such combination of alleles is essential. If the target traits are controlled by imprinting or breed-specific effects and these effects are not accounted for, the prediction of crossbred performance will not be optimal because an accurate differentiation between the genotypic effect of the heterozygotes (AB and BA) will not be possible.

8.5.1 Imprinting effects

In **chapter 2**, a method for accounting for imprinting effects was described. This method was validated in simulated data and applied to three traits in three purebred pig populations. The results of that chapter showed that the contribution of imprinting effects to the total phenotypic variance of the evaluated traits was relatively small (1-3%). Based on these results one could argue that imprinting effects are not relevant for traits and populations evaluated. However, the analyses were performed based on a limited number of records. Therefore, it may be too early to draw a general conclusion about the relevance of imprinting effect for pig breeding.

In **chapter 2,** on average 1,400 animals were analyzed per trait per population. When looking at imprinting effects, we are comparing the effect of AB and BA heterozygous genotypes. Considering a marker with minor allele frequency (MAF) of 0.5, under HWE, the number of heterozygous animals will be 700. Therefore, on average, we will have 350 animals with the AB genotype and 350 with the BA genotype. Considering that most markers will have MAF<0.50, the number of animals per heterozygous genotype will be smaller. Therefore, the data used in the estimation of imprinting variance was quite limited, which limits the conclusions. Nevertheless, the method applied in **chapter 2** was shown to be efficient to evaluate the simulated data and can, therefore, be used in future analyses on larger datasets.

8.5.2 Breed-specific effects

The results presented in **chapter 6** indicate that breed-specific effects may play a role on the genetic variation of traits measured on crossbreds and that predicting crossbred performance using a model that accounts for breed-specific effects results in similar or higher prediction accuracies compared to using a traditional genomic selection model. In **chapter 7**, breed-specific effects were also described by showing that the majority of the introgressed Asian haplotypes found in a modern European pig breed were associated with backfat. Asian haplotypes were related to more backfat than European haplotypes. The results presented in **chapter 6** indicate breed-specific effects due to current hybridization, which has a direct application in current breeding programs for improving crossbred performance. The results presented in **chapter 7** show breed-specific effects due to hybridization that took place hundreds of years ago and provide a unique insight into the process how human-driven introgression and selection may have shaped the genomic composition of commercial pig breeds.

8.5.3 Practical application

For practical application of imprinting and breed-specific effects, it is also necessary to take into account how they would fit in the current structure of the genetic evaluations (e.g. single-step approach). Accounting for breed-specific effects in the single-step approach would become possible applying the approach described by Christensen *et al.* (2014), as already discussed in **chapter 6**. For imprinting effects, we would face the same challenges as already described for dominance effects because using pedigree information it is not possible to estimate imprinting-based relationships. Therefore, applying an “imprinting **H** matrix” would not yield accurate results. Thus, if the imprinting effects need to be considered in breeding programs, their benefit would also become effective through mate allocation techniques. Imprinting effects would be used then to determine the best sire-dam combination aiming to maximize the performance of their offspring, as I already discussed for dominance effects.

Finally, one could wonder how the association between introgressed Asian haplotypes with backfat could benefit selection. In my opinion, the practical application of the associations presented in **chapter 7** would be difficult because the determination of the haplotype origin is not so straightforward. However, these results are of high relevance for the general knowledge that they provide. These results provide an example of how humans have influenced the genomic composition of European commercial pigs by introducing DNA from Asian pigs. These results demonstrate how selection and introgression by humans contributed to the hybrid nature of the genomes of commercial pigs. Further, these results also remind us that the animals that we know today as purebreds are somehow crossbreds.

8.6 Future of pig breeding

In the last years, genomic selection has been implemented in the major pig breeding companies with the expectation that genetic progress will increase up to 55% (Lillehammer *et al.* 2013). For the future, one of the challenges for pig breeders will be to accelerate the rate of genetic progress even further. In my opinion, further improvements can potentially be achieved with the application of new phenotyping techniques for refining current phenotypes; the use of single genes in marker-assisted models or genome editing; and the use of advanced reproduction technologies. Another challenge will be to keep accelerating genetic progress while taking demands of society regarding animal welfare and sustainability into account. Breeding programs will have to provide farmers with highly productive and efficient animals that will be raised under acceptable welfare standards and will leave a reduced amount of pollutants in the environment. Future of pig breeding will be based on the use of advanced technology aimed at efficient production while living up to the increasing expectations of the society. Therefore, the future of pig breeding contains plenty of opportunities and challenges and in this section I will briefly discuss two of them: 1) new phenotyping techniques and 2) genome editing.

8.6.1 New phenotyping techniques

The introduction of new phenotyping techniques has the potential to be one of the major factors that will enable to accelerate genetic progress in pig breeding. Examples of such new phenotyping techniques are X-ray computed tomography (CT scanning), and automatic feed intake recording system. Using CT scanning, carcass composition can be accurately and precisely measured (Scholz *et al.* 2015) and, therefore, this technique will be highly relevant for performance testing. With CT scanning, we will be, for example, no longer evaluating average growth as a whole. Instead, we will have the opportunity to evaluate the average gain of fat, bones and high-quality pork. The use of automatic feed intake recording system enables accurate measurement of feed intake of group-housed animals. Therefore, this technique allows a more accurate selection for feed efficiency while providing more welfare to animals via group-housing. With automatic feed intake recording systems, selecting for feed efficiency could be done from birth to slaughter, but this phenotype can also be refined and evaluated across all growth phases.

I expect that refining existing phenotypes will increase the power of GWAS. In today's situation, when we perform a GWAS for average daily gain, for example, we may be identifying QTL regions that are associated simultaneously with fat, bones and protein deposition. On the other hand, QTL regions that are associated with only one of these refined traits may not be identified because the phenotype used in the GWAS is not appropriate. If this hypothesis holds, and therefore the power to identify QTLs increases, the potential identified QTL could be used in marker-assisted models (as discussed in the "8.3 Contribution of single genes to prediction accuracy" section) to improve prediction of both purebred and crossbred performance.

Refined phenotypes, however, will also impose some challenges regarding their practical application in breeding programs. How to accommodate all refined phenotypes (and the potentially new ones that could be developed) in the selection indexes will be one of the challenges. Further, deciding on which animals to phenotype will be difficult as well, especially due to the costs of the new phenotyping techniques. Until recently, a key question was: "which animals shall we genotype?". With the fast developments of the genomic tools, the genotyping costs have been reduced considerably. Consequently, genotyping of young selection candidates has become routine in breeding programs. The costs of phenotyping, however, is moving in the opposite direction due to the increasing labor costs and the high investments required for applying new phenotyping techniques. Therefore, the main current (and future) key question is (will be): "which animals shall we phenotype?".

When the aim is to predict crossbred performance, training on crossbred data is more accurate compared to purebred data, as shown in **chapter 6** and in previous studies (Esfandyari *et al.* 2015; Hidalgo *et al.* 2015). Therefore, the use of new phenotyping techniques for evaluating traits measured on crossbreds is desired. However, this may not become feasible in the next couple of years due to the costs of such techniques.

8.6.2 Genome editing

Genome editing is a technique with which nucleotides in specific regions of the genome (causal mutations) are inserted, deleted, or replaced using artificially engineered nucleases (Kim & Kim 2014). When these changes are made in the genome of germ cells, they can be transmitted to future generations (Gordon & Ruddle 1981). Thus, editing the genome of a purebred sire will have an effect on the performance of its purebred and crossbred offspring. In addition, genome editing could be applied to causal mutations with an additive effect, but also with other effects, such as dominance and imprinting. For causal mutations with dominance effects, genome editing could be used to fix opposite alleles in dam and sire lines in order to maximize the performance of their crossbred offspring. For causal mutations with paternally-expressed imprinting effects, for example, genome editing could be used for fixing the favorable allele in the sire line and therefore also contribute to maximize crossbred performance. With this technique, it would be possible to artificially change the genome of pigs in order to obtain animals with the most favorable allelic combination.

Several methods for genome editing have been described and successful *in vivo* experiments have been performed (Carlson *et al.* 2012; Lillico *et al.* 2013; Kim & Kim 2014). In a simulation study, Jenko *et al.* (2015) showed that with a combination of genomic selection with genome editing (assuming additive effects) the response to selection was two times higher compared to the scenario where only genomic selection was applied, showing the evidence that genome editing can greatly improve genetic progress. However, genome editing has not yet been implemented in breeding programs. The first limitation is that such methods are not allowed in non-experimental animals in many countries. Further, although in the last years many GWAS have been performed and thousands of QTL regions have been identified, the successful discovery of the causal mutations underlying these QTL regions are still limited. Without knowledge of the causal mutations underlying the target QTL, successful application of genome editing will not be achieved (Jenko *et al.* 2015). Another limitation to the application of genome editing is the fact that the overall effect on the total performance of an animal by editing one specific gene is still unclear. Finally, in my opinion, the biggest limitation is whether or not society will accept it. The general public may associate an animal that had its genome edited to a transgenic animal (e.g. an animal that had incorporated in its genome fragments of DNA from other species) and fearing an impact on human health they may reject such methods.

In conclusion, the genome editing technology is available and shows promise to improve both purebred and crossbred performance, however, whether it will be put in practice is still doubtful.

8.7 Concluding remarks

This thesis provides important insights into the genetic architecture of the evaluated (complex) traits and also shows evidence that human-driven introgression and selection have shaped the genomic composition of current commercial pig breeds. For practical application, this thesis shows that by going beyond traditional genomic selection models, phenotypes can be predicted more accurately. Therefore, these improved models should be considered to improve crossbred performance. Although the research presented in this thesis was performed using data from pigs and the discussion on the practical application of results was focused on pig breeding, these results are also relevant for other livestock species where crossbreeding is applied.

8.8 References

Barrett, J.C., Fry, B., Maller, J., Daly, M. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics,* **21,** 263-265.

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

Carlson, D.F., Tan, W., Lillico, S.G., Stverakova, D., Proudfoot, C., Christian, M., Voytas, D.F., Long, C.R., Whitelaw, C.B.A., Fahrenkrug, S.C. (2012) Efficient TALEN-mediated gene knockout in livestock. *Proceedings of the National Academy of Sciences,* **109,** 17382-17387.

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

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

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

Duijvesteijn, N., Veltmaat, J.M., Knol, E.F., Harlizius, B. (2014) High-resolution association mapping of number of teats in pigs reveals regions controlling vertebral development. *BMC Genomics,* **15,** 542.

Esfandyari, H., Sørensen, A.C., Bijma, P. (2015) A crossbred reference population can improve the response to genomic selection for crossbred performance. *Genet. Sel. Evol.,* **47,** 1-12.

Falconer, D.S., Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics,* 4th edn. Longmans Green, Harlow.

George, E.I., McCulloch, R.E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association,* **88,** 881-889.

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

Gordon, J.W., Ruddle, F.H. (1981) Integration and stable germ line transmission of genes injected into mouse pronuclei. *Science,* **214,** 1244-1246.

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

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

Hayes, B.J., Lewin, H.A., Goddard, M.E. (2013) The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. *Trends Genet.,* **29,** 206-214.

Hickey, J. (2013) Sequencing millions of animals for genomic selection 2.0. *J. Anim. Breed. Genet.,* **130,** 331-332.

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

Jenko, J., Gorjanc, G., Cleveland, M.A., Varshney, R.K., Whitelaw, C.B.A., Woolliams, J.A., Hickey, J.M. (2015) Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. *Genet. Sel. Evol.,* **47,** 1-14.

Kim, H., Kim, J.-S. (2014) A guide to genome engineering with programmable nucleases. *Nat. Rev. Genet.,* **15,** 321-334.

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

Legarra, A., Croiseau, P., Sanchez, M.P., Teyssèdre, S., Sallé, G., Allais, S., Fritz, S., Moreno, C.R., Ricard, A., Elsen, J.-M. (2015) A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. *Genet. Sel. Evol.,* **47,** 6.

Li, J., Das, K., Fu, G., Li, R., Wu, R. (2011) The Bayesian lasso for genome-wide association studies. *Bioinformatics,* **27,** 516-523.

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

Lillico, S.G., Proudfoot, C., Carlson, D.F., Stverakova, D., Neil, C., Blain, C., King, T.J., Ritchie, W.A., Tan, W., Mileham, A.J. (2013) Live pigs produced from genome edited zygotes. *Sci. Rep.,* **3**.

Lopes, M.S., Silva, F.F., Harlizius, B., Duijvesteijn, N., Lopes, P.S., Guimarães, S.E., Knol, E.F. (2013) Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genet.,* **14,** 92.

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

Morlon, A., Munnich, A., Smahi, A. (2005) TAB2, TRAF6 and TAK1 are involved in NF-κB activation induced by the TNF-receptor, Edar and its adaptator Edaradd. *Hum. Mol. Genet.,* **14,** 3751-3757.

Onteru, S., Fan, B., Du, Z.Q., Garrick, D., Stalder, K., Rothschild, M. (2012) A whole-genome association study for pig reproductive traits. *Anim. Genet.,* **43,** 18-26.

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

Rolf, M.M., Taylor, J.F., Schnabel, R.D., McKay, S.D., McClure, M.C., Northcutt, S.L., Kerley, M.S., Weaber, R.L. (2010) Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. *BMC Genet.,* **11,** 24.

Scholz, A., Bünger, L., Kongsro, J., Baulain, U., Mitchell, A. (2015) Non-invasive methods for the determination of body and carcass composition in livestock: dual-energy X-ray absorptiometry, computed tomography, magnetic resonance imaging and ultrasound: invited review. *animal***,** 1-15.

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

Sham, P.C., Purcell, S.M. (2014) Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.,* **15,** 335-346.

Sonesson, A.K., Meuwissen, T. (2009) Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol,* **41,** 37.

Stephens, M., Balding, D.J. (2009) Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.,* **10,** 681-690.

Thesleff, I., Mikkola, M.L. (2002) Death receptor signaling giving life to ectodermal organs. *Science Signaling,* **2002,** pe22.

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

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

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

Vitezica, Z.G., Varona, L., Legarra, A. (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics,* **195,** 1223-1230.

Wakefield, J. (2009) Bayes factors for genome-wide association studies: comparison with P-values. *Genet. Epidemiol.,* **33,** 79-86.

Wang, H., Misztal, I., Aguilar, I., Legarra, A., Muir, W. (2012) Genome-wide association mapping including phenotypes from relatives without genotypes. *Gent. Res.,* **94,** 73-83.

Zeng, J., Toosi, A., Fernando, R.L., Dekkers, J.C., Garrick, D.J. (2013) Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genet. Sel. Evol.,* **45**.

Zhu, Z., Bakshi, A., Vinkhuyzen, A.A., Hemani, G., Lee, S.H., Nolte, I.M., van Vliet-Ostaptchouk, J.V., Snieder, H., Esko, T., Milani, L. (2015) Dominance genetic variation contributes little to the missing heritability for human complex traits. *The American Journal of Human Genetics,* **96,** 377-385.

Summary

With the implementation of genomic selection in pig breeding, the genetic progress per year in purebred populations is expected to increase up to 55% compared to traditional selection based on pedigree information. However, as most of the animals in the pork production system are crossbreds, the increase in genetic progress in purebreds will only be observed on production farms if this progress is expressed in the performance of crossbreds. The main goal of this thesis was to evaluate different models based on genomic information that can contribute to the improvement of crossbred performance. Another aim was to gain insight into the genetic architecture of the evaluated (complex) traits and to investigate how selection history has influenced haplotype patterns of current commercial pigs. In **chapter 2**, an SNP regression method was applied to estimate the contribution of additive, dominance, and imprinting effects to the phenotypic variation. The SNP regression method was validated in simulated data and applied to three traits in three purebred pig populations. I showed that dominance effects account for a considerable proportion of the phenotypic variance of several traits. The contribution of imprinting effects for the phenotypic variance of several traits was limited. In **chapter 3**, the accuracy of predicting performance with a model that accounts for dominance effects, in addition to additive effects was compared to a model that accounts for only additive effects. This analysis was performed with records on lifetime daily gain from three purebred pig populations and showed that accounting for both additive and dominance effects increases the prediction accuracy of phenotypes. In **chapter 4**, I investigated the importance of dominance using a high-density SNP panel in pigs and showed that GWAS are also able to identify QTL with dominance effects in addition to QTL with additive effects. In a GWAS study for number of teats the dominant QTL explained one-fourth of the variance explained by additive QTL. In **chapter 5**, the utility of GWAS findings for the prediction of phenotypes was investigated. Individual SNPs were incorporated in the traditional methods (BLUP and GBLUP) resulting in marker-assisted BLUP (MA-BLUP) and marker-assisted GBLUP (MA-GBLUP). This chapter showed that accounting for GWAS findings in the genetic evaluation improves the prediction of phenotypes. In **chapter 6**, the existence of breed-specific effects was investigated. First, the separate contribution of each purebred line to the genetic variance of traits observed in (two-way) crossbred animals was estimated. Second, the prediction accuracy of crossbred performance was compared using a traditional model and a model that accounts for breed-specific effects. Breed-specific effects were shown to play a role in the genetic variation of crossbred performance, and predicting crossbred performance using a model that accounts for breed-specific effects resulted in the same or higher prediction accuracies compared to traditional genomic selection models. In **chapter 7**, the hypothesis that the introgression landscape in commercial breeds is shaped mostly by artificial selection was tested by searching for an association between introgressed Asian haplotypes and commercial trait phenotypes. The majority of the introgressed Asian haplotypes in a modern European pig breed were found to be associated with backfat, which indicates that human-driven introgression and selection may have shaped the genomic composition of commercial pig breeds. In the general discussion (**Chapter 8**), the results of this thesis were discussed in a broader perspective. The main focus of this chapter was on the practical application of the results of this thesis in pig breeding programs with the specific emphasis on the improvement of crossbred performance. Further, the interpretation of results from current models based on genomic information and the challenges and opportunities for the future of pig breeding were also discussed.

Training and supervision plan

The Basic Package (3 credits)	year	credits
WIAS Introduction Course	2012	1.5
Course on philosophy of science and/or ethics	2013	1.5
Scientific Exposure (22 credits)	year	credits
International conference		
15th QTL-MAS Workshop, France	2011	0.6
4th ICQG, Scotland	2012	1.8
64th EAAP Annual Meeting, France	2013	1.5
VI Congresso da Sociedade Científica de Suinicultura, Portugal	2013	0.9
10th WCGALP, Canada	2014	1.8
Workshop phenotyping and genotyping in animal breeding, Brazil	2014	0.9
V Simpósio internacional de genética e melhoramento, Brazil	2014	0.9
I GENORDE - I encontro da rede genômica Nordeste, Brazil	2015	0.6
66th EAAP Annual Meeting, Poland	2015	1.5
XVII Congresso ABRAVES, Brazil	2015	1.2
Seminars and workshops		
WIAS Science Day, Wageningen	2013-1015	0.9
F&G Connection Days, Vught	2012, 2014	0.6
Presentations		
4th ICQG, Scotland, POSTER	2012	1.0
64th EAAP Annual Meeting, France, POSTER	2013	1.0
64th EAAP Annual Meeting, France, ORAL	2013	1.0
VI Congresso da Sociedade Científica de Suinicultura, Portugal, INVITED SPEAKER	2013	1.0
10th WCGALP, Canada, ORAL	2014	1.0
Workshop phenotyping and genotyping in animal breeding, Brazil, INVITED SPEAKER	2014	1.0
I GENORDE, Brazil, INVITED SPEAKER	2015	1.0
66th EAAP Annual Meeting, Poland, ORAL	2015	1.0
XVII Congresso ABRAVES, Brazil, INVITED SPEAKER	2015	1.0

In-Depth Studies (7 credits)	year	credits
Genomic Selection in Livestock	2011	1.5
Genomic Prediction	2012	0.9
Advanced methods and algorithms in animal breeding with focus on genomic selection	2012	1.5
Innovagen Winter School II: Use of SNP whole-genome sequencing data in livestock genetic improvement programs	2013	1.5
Genetic analysis using ASReml4.0	2014	1.5
Professional Skills Support Courses (5 credits)	year	credits
Project management	2012	0.9
Dutch Plus	2012	2.4
Techniques for writing and presenting a scientific paper	2014	1.2
Research Skills Training (8 credits)	year	credits
Preparing own PhD research proposal	2012	6.0
Getting started with ASReml	2014	0.3
External training period at NMBU, Norway	2015	2.0
Didactic Skills Training (2 credits)	year	credits
Lecturing		
Genomics in pig breeding, Toegepaste Biologie, HAS Hogeschool	2015	0.6
Supervising theses		
MSc thesis	2013	1.5
Management Skills Training (6 credits)	year	credits
WAPS Council and Education Committee	2013-2014	6.0
Education and Training Total (minimum 30, maximum 60 credits)		**53**

One credit equals a study load of approximately 28 hours.

Curriculum vitae

About the author

Marcos Soares Lopes is Brazilian. He was born on the 2nd of November 1985 in Caratinga and was raised in the coffee farm of his parents in Bom Jesus do Galho, both in the state of Minas Gerais. In 2003, he finished high school at E.E. Padre Dionísio Homem de Faria and in 2004 he started his BSc in Animal Science at the Universidade Federal de Lavras. In 2005, he decided to move to Viçosa to continue his studies at the Universidade Federal de Viçosa. He obtained his BSc diploma in 2009 with the thesis entitled "QTL fine-mapping on SSC4q16-q25 for carcass traits". In 2011, he obtained his MSc diploma with thesis entitled "Number of SNP markers for parental identification, kinship and inbreeding estimation in pigs". Both his BSc and MSc theses were supervised by Prof. Dr. Simone Guimarães and co-supervised by Dr. Egbert Knol, Prof. Dr. Paulo Sávio Lopes and Prof. Dr. Fabyano Silva. During both his BSc and MSc studies, Marcos did a three-month internship at Topigs Norsvin Research Center where he developed part of the research presented in his BSc and MSc theses. In 2011, he moved to the Netherlands and started working as a researcher at Topigs Norsvin Research Center. Since then, he has been working on the implementation of genomic information in the breeding program of Topigs Norsvin. In 2012, he accepted an additional challenge and combined his work at Topigs Norsvin Research Center with a PhD project at the Animal Breeding and Genomics Center of Wageningen University. In 2015, he received a WIAS PhD fellowship to develop part of his PhD project at the Norwegian University of Life Sciences where he worked together with Prof. Dr. Theo Meuwissen and the team of scientists of Norsvin. The results of his PhD project are presented in this thesis entitled "Genomic selection for improved crossbred performance".

Peer-reviewed publications

2015

Lopes M.S., Bastiaansen J.W.M., Janss L., Knol E.F. & Bovenhuis H. (2015) Estimation of additive, dominance, and imprinting genetic variance using genomic data. *G3: Genes|Genomes|Genetics* 5 2629-37.

Lopes M.S., Bastiaansen J.W.M., Janss L., Knol E.F. & Bovenhuis H. (2015) Genomic prediction of growth in pigs based on a model including additive and dominance effects. *Journal of Animal Breeding and Genetics* (Early online).

Bosse M., **Lopes M.S.,** Madsen O., Megens H.-J., Crooijmans R.P., Frantz L.A., Harlizius B., Bastiaansen J.W.M. &. Groenen M. (2015) Artificial selection on introduced Asian haplotypes shaped the genetic architecture in European commercial pigs. *Proceeding Royal Society B* 282, 20152019.

Hidalgo A.M., **Lopes M.S.**, Harlizius B. & Bastiaansen J.W.M. (2015) Genome-wide association study reveals regions associated with gestation length in two pig populations. *Animal Genetics* (Early online).

Sevillano C.A., **Lopes M.S.**, Harlizius B., Hanenberg E.H., Knol E.F. & 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.

Veroneze R., **Lopes M.S.**, Hidalgo A.M., Guimarães S.F., Silva F.F., Harlizius B., Lopes P.S., Knol E.F., van Arendonk J.M. & Bastiaansen J.W.M. (2015) Accuracy of genome-enabled prediction exploring purebred and crossbred pig populations. *Journal of Animal Science* 93, 4684-91.

Campos C.F., **Lopes M.S.**, e Silva F.F., Veroneze R., Knol E.F., Lopes P.S. & Guimarães S.E. (2015) Genomic selection for boar taint compounds and carcass traits in a commercial pig population. *Livestock Science* 174, 10-7.

Sell-Kubiak E., Duijvesteijn N., **Lopes M.S.**, Janss L.L., Knol E.F., Bijma P. & Mulder H. (2015) Genome-wide association study reveals novel loci for litter size and its variability in a Large White pig population. *BMC Genomics* 16, 1049.

Hidalgo A.M., Bastiaansen J.W.M., **Lopes M.S.**, Harlizius B., Groenen M.A. & de Koning D.-J. (2015) Accuracy of predicted genomic breeding values in purebred and crossbred pigs. *G3: Genes|Genomes|Genetics* 5, 1575-83.

Hidalgo A.M., Bastiaansen J.W.M., **Lopes M.S.**, Veroneze R., Groenen M.A.M. & de Koning D.-J. (2015) Accuracy of genomic prediction using deregressed breeding values estimated from purebred and crossbred offspring phenotypes in pigs. *Journal of Animal Science* 93, 3313-21.

Veroneze R., Lopes P.S., **Lopes M.S.**, Hidalgo A.M., Guimarães S.F., Silva F.F., Harlizius B., Knol E.F., van Arendonk J.M. & Bastiaansen J.W.M. (2015) Accounting for genetic architecture in single- and multi-population genomic prediction using weights from genome wide association studies in pigs. *Journal of Animal Breeding and Genetics* (accepted).

2014

Lopes M.S., Bastiaansen J.W., Harlizius B., Knol E.F. & Bovenhuis H. (2014) A genome-wide association study reveals dominance effects on number of teats in pigs. *PloS One* 9, e105867.

Diniz D., **Lopes M.S.**, Broekhuijse M., Lopes P.S., Harlizius B., Guimarães S.F., Duijvesteijn N., Knol E.F. & Silva F.F. (2014) A genome-wide association study reveals a novel candidate gene for sperm motility in pigs. *Animal Reproduction Science* 151, 201-7.

Silva F.F., Mulder H., Knol E.F., **Lopes M.S.**, Guimarães S.F., Lopes P.S., Mathur P., Viana J. & Bastiaansen J.W.M. (2014) Sire evaluation for total number born in pigs using a genomic reaction norms approach. *Journal of Animal Science* 92, 3825-34.

Azevedo C., Silva F.F., Resende M., **Lopes M.S.**, Duijvesteijn N., Guimarães S.F., Lopes P.S., Kelly M., Viana J. & Knol E.F. (2014) Supervised independent component analysis as an alternative method for genomic selection in pigs. *Journal of Animal Breeding and Genetics* 131, 452-61.

Hidalgo A.M., Bastiaansen J.W.M., Harlizius B., Knol E.F., **Lopes M.S.**, Koning D. & Groenen M. (2014) Asian low-androstenone haplotype on pig chromosome 6 does not unfavorably affect production and reproduction traits. *Animal Genetics* 45, 874-7.

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

2013

Lopes M.S., Silva F.F., Harlizius B., Duijvesteijn N., Lopes P.S., Guimarães S.E. & Knol E.F. (2013) Improved estimation of inbreeding and kinship in pigs using optimized SNP panels. *BMC Genetics* 14, 92.

Veroneze R., Lopes P.S., Guimarães S.E.F., Silva F.F., **Lopes M.S.**, Harlizius B. & Knol E.F. (2013) Linkage disequilibrium and haplotype block structure in six commercial pig lines. *Journal of Animal Science* 91, 3493–501.

2011

Harlizius B., **Lopes M.S.**, Duijvesteijn N., van de Goor L.H.P., van Haeringen W.A., Panneman H., Guimarães S.E.F., Merks J.W.M. & Knol E.F. (2011) A SNP set for paternal identification to reduce the costs of trait recording in commercial pig breeding. *Journal of Animal Science* 89, 1661-8.

Sousa K.R.S., Guimarães S.E.F., Silva Filho M.I., **Lopes M.S.**, Pinto A.P.G., Verardo L.L., Braccini Neto J. & Lopes P.S. (2011) Mapping of quantitative trait loci mapping in chromosomes 5, 7 and 8 of swines. *Brazilian Journal of Animal Science* 40, 115-23.

Sollero B.P., Guimarães S.E.F., Rilington V., Tempelman R., Raney N., Steibel J., Guimarães J., Lopes P.S., **Lopes M.S.** & Ernst C. (2011) Transcriptional profiling during foetal skeletal muscle development of Piau and Yorkshire–Landrace cross-bred pigs. *Animal Genetics* 42, 600-12.

2010

Pinto A.P.G., Lopes P.S., **Lopes M.S.**, Silva Filho M.I., Sousa K.R.S., Carneiro P.L.S. & Guimarães S.E.F. (2010) Detection of quantitative trait loci on chromosomes 9, 10 and 11 of swines. *Brazilian Journal of Animal Science* 39, 2174-81.

2009

Silva K.M., Guimarães S.E.F., Lopes P.S., Nascimento C.S., **Lopes M.S.** & Amaral M.M.D.C. (2009) Mapping of quantitative trait loci for performance on chromosome 4 in pigs. *Brazilian Journal of Animal Science* 38, 474-9.

2008

Paixão D.M., Silva Filho M.I., Pereira M.S., **Lopes M.S.**, Barbosa L., Souza K.R.S., Lopes P.S. & Guimarães S.E.F. (2008) Quantitative trait loci for carcass, internal organ and meat quality traits on porcine chromosomes 16, 17 and 18. *Genetics and Molecular Biology* 31, 898-901.

The complete list of publications which includes publication in conference proceeding can be found in the online CV:

Lattes:	http://lattes.cnpq.br/5989175110323615
Researchgate:	https://www.researchgate.net/profile/Marcos_Lopes5

Acknowledgments

I have clear life goals and I work hard to achieve them. However, without the support of my family and friends, I would be nobody, I would get nowhere. Here (acknowledgments), I will use the Brazilian national passion (football/soccer) to acknowledge most of the people that have somehow helped me to achieve one of my biggest goals: my PhD degree.

In a soccer team, we need a goalkeeper, defenders, midfielders, forwards and the coach. Although these are the ones that really fight on the field, a soccer team will only be completely successful if it has good supporters and a well-prepared stadium for its training and games.

GOALKEEPER: the goalkeeper has a major role in a soccer team. He is the one that do whatever is necessary to avoid the opponents' scores. He is also the one that kicks the ball forward to start the attack. My goalkeeper is God. He protects me and gives me the strength to keep moving forward, especially in the difficult times. Thank you, Lord!

DEFENDERS: the defenders work together with the goalkeeper avoiding the opponents' scores. Even when you have a very good goalkeeper, it is hard to win the game if you do not have good defenders. I am lucky because I have the best defenders on Earth. My defenders are my family members: especially my parents (Raimundo and Maria), my brothers (Manoel, José, Rogério and Renato), my sisters (Rosângela, Rosilene and Rosilei), my girlfriend (Silvia), my uncle (tio Nininho) and my aunt (tia Maria). I have no words to thank my parents for everything they have done for me. They are simply the best parents that somebody could ask for. My brothers and sisters are also great. They have been always there for me, especially when I started my bachelor. I will never forget their help back there. Keeping up a relationship when the one you love lives in the other side of the ocean it is not easy at all. But with Silvia, everything becomes easier and the physical distance is just a detail that soon will not exist anymore. Silvia is my girlfriend, my friend, my best tourism agent, my best company for traveling, my everything. I am also grateful to tio Nininho e tia Maria for all their support, for believing that I could get where I am today, for motivating me to follow my dreams. Meu muito obrigado a todos! Amo vocês!

MIDFIELDERS: Without a solid and strong midfield, it is very difficult for the forwards to score. The midfielders are the ones that prepare the attack and lead the forwards to the scores. In my team, the midfielders are my friends and colleagues from Topigs Norsvin Research Center. First of all, I would like to thank Egbert and Jan Merks for believing in me and offering the job. Egbert: I have learned a lot with you. If I have to say one thing that I have learned the most with you, I would say that it is the simple way of presenting scientific results. Thanks for everything. Dieuwke and Egiel: you guys have no idea about how much I have learned with you. All the (many) times that I have asked you questions about SQL, excel, etc., I have carefully looked how you solved my doubts to learn a bit more about these programs and about the whole structure of the database as well. Thanks for small things like "VLOOKUP" and "create table", thanks for teaching me a lot, even when you did not realize that you were teaching me something. Barbara: how can I forget all your worries about the amount of work that I had to do besides my PhD? Thanks for all the small and fruitful talks. Having you as a colleague at TNRC is great, but getting to know the good person that you are is even greater. The world would be a nicer place if there were more people like you. Roos and Annemieke: thanks for the nice work environment and all the non-work related trips. Pramod: thanks a lot for your "proof-reading help" since my first internship at TNRC. Jascha: he is the craziest office mate that I could ever get. Dude, thanks for our friendship and all the beers that we have drunk together. My dear ladies from TA, thanks a lot for all your help with everything. Special thanks to Sharonne and Gerda for the table football games. Thanks to everybody in Beuningen and in Vught for the nice work environment. We are a great team. Finally, I want to say thanks to two players that are too good to play in only one position. Naomi and Claudia are midfielders but also forwards. They have been my office mates in Beuningen and in Wageningen, and great friends as well. Naomi is the mother of the two little Dutch girls that I like the most. She is really Dutch (sometimes), but she has become (somehow) a bit Brazilian as well. Thanks for helping with my Dutch, for the trips, beers, for our friendship. Claudia is "chata" and vegetarian but her Brazilian heart compensates it all. Thanks for all the good time in Wageningen and for our friendship. Thank you all, bedankt, gracias!

FORWARDS: the forwards have to score to ensure the victory. In my team, I am one of them (the striker), but I need the help of each one of the other forwards to score as much as possible. My forwards are my friends and colleagues from Wageningen University (ABGC). All these years in Wageningen were great for the work I have done and for the friends I have made. André, Katrijn, Gus, Gabriel, Mathieu, Hamed, Ewa, Yvonne, Yogesh, Nancy, Britt, Aniek, Mario, Juanma and many others: thanks for everything. André: he is the kind of friend that is actually like a brother. He was always there for the trips, buying cars, drinking shots, etc.

Valeu, negão! Obrigado por tudo! Katrijn: she is so Dutch! Thanks for the great party weekends in Belgium, dinners, and our friendship! Nancy is my favorite native English speaker. Bro, thanks a lot for all your help. Gus is Gus, that's more than good enough. He just should improve his English to facilitate our communication. Mathieu is French, but nobody is perfect, not even the "golden boy 2". Thanks also to Lisette and Ada for helping out with all the paperwork of the PhD. Thanks to all the ABGC family!

I would like to thank as well other forwards that I had the pleasure to play with during my time in Norway. I would like to thank Cecilie, Oscar, Ina, Øyvind, Maren, the two Eli's, Theo, Dan and all the others at NMBU and Norsvin for the great experience during my three months in Norway. Tusen takk!

COACH: the coach is the one that gives guidance and support to the team. During my PhD project, I had guidance and support from four coaches: Johan, Egbert, Henk and John. Having multiple coaches can be dangerous if there is no match of the different ideas. I was really lucky because the different ideas of my coaches always resulted in one unique and stronger idea. I have learned a lot with all of them but to keep it short I will list the major points that I learned with each one. With Johan, I learned that two meetings per year are enough when the guidance and aims are clear (it is all about about being a good leader). With Egbert, I learned how to place my results in a more practical way, while, with Henk, I learned how to exploit my results in a more scientific way. With John, I learned how a small change in a sentence can make a statement much clearer and attractive at the same time. I still can improve quite a lot my writing skills. However, thanks especially to John, my writing skills have already improved a lot in the last years.

SUPPORTERS: in my team, the supporters are additional players. My supporters are my friends and players of my former team (Universidade Federal de Viçosa). In Viçosa, I made friends that I will keep for life. The "República Fim do Mundo" became a family that always supported me to go ahead and win. Special thanks to Paulo, Pedro, Lucas, João, Mateus, Adelmo, DuCarmo, Luís and Saiuri. From UFV, I would like to thank the people that were there to help me out when I gave my first steps into the animal breeding world: Carlos, Katiene, Ana Paula, Miguel, Crisoca, Priscila, Nicola, Lucas, Renata and many others. Simone, Paulo Sávio and Fabyano: thanks to all support and opportunities that you gave me. Special thanks also to Sr. Zé Geraldo, the manager of the pig-breeding farm. What I learned with you Sr. Zé, I will take for life. Thanks for all your advice and support. From the Dutch side, I would like to thank my Dutch families: Jos and Ans, and the family de Haardt. These two families helped to transform a shy Brazilian guy in a better person (that now talks a bit too much). I will always be grateful to you all.

STADIUM: Finally, I would like to say thanks to the place that made everything possible. I want to thank the Netherlands for being the stadium, the stage, my home during all these years. I will always have a green and yellow heart, proud of being Brazilian. However, in the last years, my heart became a little bit orange as well. Bedank Nederland en nederlanders!

Colophon

Colophon

The research presented in this thesis was financially supported by Topigs Norsvin (Vught, the Netherlands) and was part of the Breed4Food Consortium, a public-private partnership in the domain of animal breeding and genomics.

The cover of this thesis was designed by Marcos Soares Lopes.

This thesis was printed by GVO drukkers & vormgevers B.V. | Ponsen & Looijen, Ede, the Netherlands.