Exploiting genomic information on purebred and crossbred pigs

ISSN: 1652-6880
ISBN (print version): 978-91-576-8438-7
ISBN (electronic version): 978-91-576-8439-4
ISBN: 678-94-6257-601-8

Thesis committee

Promotor
Prof. Dr M.A.M. Groenen
Personal chair at the Animal Breeding and Genomics Centre
Wageningen University

Thesis co-promotors
Prof. Dr D.J. de Koning
Professor in Animal Breeding at the Department of Animal Breeding and Genetics
Swedish University of Agricultural Sciences

Dr J.W.M. Bastiaansen
Researcher, Animal Breeding and Genomic Centre
Wageningen University

Other members (assessment committee)
Prof. Dr B. Kemp, Wageningen University
Prof. Dr F.A. van Eeuwijk, Wageningen University
Prof. Dr L. Rydhmer, Swedish University of Agricultural Sciences, Sweden
Prof. Dr P. Uimari, University of Helsinki, Finland

The research presented in this doctoral thesis was conducted under the joint auspices of the Swedish University of Agricultural Sciences and the Graduate School of Wageningen Institute of Animal Sciences of Wageningen University and is part of the Erasmus Mundus Joint Doctorate Program “EGS-ABG”.

Exploiting genomic information on purebred and crossbred pigs

André Marubayashi Hidalgo

ACTA UNIVERSITATIS AGRICULTURAE SUECIAEDOCTORAL THESIS No 2015:120

Thesis

submitted in fulfillment of the requirements for the degree of doctor from
Swedish University of Agricultural Sciences
by the authority of the Board of the Faculty of Veterinary Medicine and Animal
Science and from
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 of Wageningen University and
the Board of the Faculty of Veterinary Medicine and Animal Science at
the Swedish University of Agricultural Sciences
to be defended in public
on Wednesday December 9, 2015
at 4 p.m. in the Aula of Wageningen University

Hidalgo, A.M.
Exploiting genomic information on purebred and crossbred pigs

Joint PhD thesis, Swedish University of Agricultural Sciences, Uppsala, Sweden and
Wageningen University, Wageningen, the Netherlands (2015)
With references, with summary in English

Abstract

Hidalgo, A.M. (2015). Exploiting genomic information on purebred and crossbred pigs. Joint PhD thesis, between Swedish University of Agricultural Sciences, Sweden and Wageningen University, the Netherlands

The use of genomic information has become increasingly important in a breeding program. In a pig breeding program, where the final goal is an increased crossbred (CB) performance, the use of genomic information needs to be thoroughly evaluated as it may require a different strategy of what is applied in purebred (PB) breeding programs. In this thesis, I explore the use of genomic information for the genetic improvement of PB and CB pigs. I first focus on the identification of genomic regions affecting traits that are important to breeders. I identified two quantitative trait loci (QTL) regions for gestation length, one for Dutch Landrace on *Sus scrofa* chromosome (SSC) 2 and the other one for Large White on SSC5. I also fine-mapped and narrowed down the region of a previously detected QTL for androstenone level SSC6 from 3.75 Mbp to 1.94 Mbp. A tag-SNP of this fine-mapped region was further investigated and no unfavorable pleiotropic effects were found; indicating that using the studied marker for selection would not unfavorably affect the other studied traits. After that, the focus was changed to the application of genomic selection in pigs. Within-population predictions showed high accuracies, whereas across-population prediction had accuracies close to zero. Using combinations among Dutch Landrace and Large White populations plus their cross showed that multi-population prediction was not better than within-population. The exception was when the CB pigs were predicted with records from both parental populations added to the CB training data. When using PB pigs to train CB ones, the predictive ability found indicates that selection in the PB pigs results in response in the CB ones. When assessing the source of information used to estimate the breeding values used as response variable, I showed that a more accurate prediction of CB genetic merit was found when training on PB data with breeding values estimated using CB performance than training on PB data with breeding values estimated using PB performance. I also studied the accuracy of using CB pigs in the training population to select PB for CB performance. Predictive ability when using CB phenotypes for training was observed, however, the accuracy was lower than using PB phenotypes in the training population. Lastly, I evaluate the inclusion of dominance in the model when using a CB training population. Results showed that accounting for dominance effects can be slightly beneficial for genomic prediction compared with a model that accounts only for additive effects.

Table of contents

9 1 – General introduction

21 2 – Genome-wide association study reveals regions associated with gestation length in two pig populations

31 3 – 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

63 4 – Asian low-androstenone haplotype on pig chromosome 6 does not unfavorably affect production and reproduction traits

73 5 – Accuracy of predicted genomic breeding values in purebred and crossbred pigs

97 6 – Accuracy of genomic prediction using deregressed breeding values estimated from purebred and crossbred offspring phenotypes in pigs

117 7 – Accuracy of genomic prediction of purebreds for crossbred performance in pigs

133 8 – Evaluation of genomic prediction of purebreds for crossbred performance in pigs accounting for dominance effects

161 9 – General discussion

181 Summary

185 Acknowledgements

191 Curriculum vitae

197 Training and education

201 Colophon

1

General introduction

1.1 Introduction

Animal breeding aims to select the best animals to be the parents of the next generation. A large variety of techniques, strategies and methods have been developed to achieve this goal. In recent years, genotyping technology has improved considerably and high-throughput genomic information became available. Efficient use of this information, hence, is crucial for the competitiveness of a breeding company. In this work, therefore, I will explore the use of genomic information for the genetic improvement of purebred and crossbred pigs. In this general introduction, I will first concentrate on the identification of genomic regions that affect traits that are important to breeders. After that, I will focus on the application of genomic selection, and later on crossbreeding with emphasis on heterosis and dominance. These topics are relevant in the application of genomic information in the present breeding situation.

1.2 QTL mapping

Most traits of economic importance in livestock production are quantitative, i.e., are affected by many loci to various degrees. The genes affecting a quantitative trait, so-called "quantitative trait loci" (QTL), are difficult to identify, yet they are relevant for breeding purposes. Currently, 13,030 QTL for 663 traits have been described for pig (Animal QTLdb, http://www.animalgenome.org/QTLdb).

Genetic markers can be divided in three groups: 1) direct markers: loci that code for the causative mutation, 2) LD markers: loci are in population-wide linkage disequilibrium with the causative mutation, 3) LE markers: loci are in population-wide linkage equilibrium with the causative mutation in outbred populations (Dekkers 2004). Direct markers are the most difficult to detect because proving causality is extremely hard. The LD markers can be detected using candidate genes (Rothschild and Soller 1997), fine-mapping (Andersson 2001; Georges 2007) or genome-wide association studies (GWAS); LD markers are located close to the causative mutation so that linkage disequilibrium between marker and QTL exists. The LE markers within linkage distance of a QTL can be identified by using breed crosses or analysis of large half-sib families within the breed.

The first study that detected a QTL in pigs, identified a region affecting fat deposition on chromosome 4 (Andersson *et al.* 1994). This study, along with other contemporaneous studies, performed linkage mapping in an F2 design using microsatellite markers spread across the genome. The F2 were, in general, obtained from crosses between a European-descent commercial breed and either a European Wild Boar or Asian breed, such as Meishan (e.g. Knott *et al.*, 1998; De Koning *et al.*, 1999). Many QTL regions were detected using this methodology (reviewed by Rothschild *et al.* (2007)), however the confidence interval of these QTL were usually very large which hampered the use of this information in a breeding program. On top of the large confidence intervals, most of these QTL were detected in experimental populations using crosses, therefore the identified QTL could hardly be used directly for selection within breeds as they differ in frequency across breeds (Dekkers 2004). In practice, QTL analysis in crossed populations has been superseded by GWAS analyses within purebred populations, which will be described later.

The fine-mapping approach aims to find the causative mutation or at least refine the mapping resolution of a previously detected QTL region, which should lead to narrowing down this QTL region. The major factors affecting the mapping resolution are: 1) marker density, 2) crossover density, 3) accuracy of inferring the QTL genotype, and 4) molecular architecture of the QTL (Georges 2007). Provided that there are enough markers, then to increase the mapping resolution, there is the need to increase the number of recombinations. This increase can be achieved by breeding additional generations or increasing the population size (Darvasi and Soller 1995). The fine-mapping approach has been successful in detecting the causal mutation only for a small number of QTL, for example *FAT1* (Berg *et al.* 2006) and the insulin-like growth factor 2 gene (*IGF2*) (Van Laere *et al.* 2003).

Besides the linkage approach used for QTL mapping, other approaches were developed and applied in pig breeding, such as the candidate gene approach. The candidate gene approach involves 1) selecting the candidate gene based on its known biological function, 2) amplifying the gene, 3) finding polymorphic regions, 4) large scale genotyping of the polymorphic region, 5) phenotyping and genotyping a target population, 6) performing an association between phenotype and genotype, and finally 7) assessing the detected associations (Rothschild and Soller 1997). The candidate gene approach was successful in detecting few QTL, for example the porcine melanocortin-4 receptor (*MC4R*) gene (Kim *et al.* 2000). This approach discovered LD markers, which allows selection across animals of the same population, therefore is relevant for breeding (Dekkers 2004).

The pig genome sequence was published in 2012 by the Swine Genome Sequencing Consortium (Groenen *et al.* 2012). In the meantime, the identification of high numbers of single nucleotide polymorphisms (SNP) and the development of methodologies to simultaneously genotype large numbers of SNP, enabled the design of a SNP chip for pigs with approximately 60,000 markers (Ramos *et al.* 2009). The higher marker density across the genome allowed performance of genome-wide association mapping, for the identification of QTL. GWAS evaluates whether variations in the genome (e.g. SNP) are associated with variation in a given trait. The assumption underlying a GWAS is that significant associations occur because the SNP is in linkage disequilibrium (LD) with a causative mutation affecting the trait. The first study performing a GWAS in pigs identified a cluster of markers associated with androstenone level on chromosome 6 (Duijvesteijn *et al.* 2010).

To make use of markers linked to QTL in breeding, Fernando and Grossman (1989) developed a methodology that incorporated markers associated with quantitative traits into the conventional mixed models genetic evaluation. This method was applied by breeding companies as a complementary tool to the pedigree-based genetic evaluation (Ibáñez-Escriche *et al.* 2014). Before incorporating new markers in the genetic evaluation, it is recommended to assess the pleiotropic effects of that marker on other production and reproduction traits. This check is important to avoid unfavorable effects due to pleiotropy and/or due to genetic hitchhiking. Such unfavorable effects are examined by testing the association between the marker and the other traits.

So far, only a handful of causative mutations has been discovered and for the majority of QTL regions the causal variation has not been identified. The general finding from GWAS for quantitative traits, in livestock species, is that the majority of the economically important traits are controlled by many genes with small effects. Therefore, given the polygenic nature of most traits in livestock and the availability of a large number of genetic markers across the genome, genomic selection became the method of choice for application in animal breeding.

1.3 Genomic selection

Genomic selection (GS) entails using markers across the genome to estimate breeding values (Meuwissen *et al.* 2001). The assumption underlying genomic selection is that the effects of QTL will be captured by markers due to LD. In GS, individuals with both phenotypes and genotypes compose the so-called training population. Information on the training population is used to estimate direct genomic values (DGV) of selection candidates that are genotyped but do not have phenotypes. Selection based on DGV can be performed in these selection candidates. The DGV is an estimate, based on the animal's genomic information, of the value that an animal transfers to its progeny. To calculate the DGV, marker effects can be estimated by regressing the phenotypes on the marker genotypes in the training population. Afterwards, the genotypes of each selection candidate are multiplied by the marker effect and summed, resulting in the DGV. Various methods have been developed for the application of GS. These methods are generally based on mixed models, simple linear regression or shrinkage-based approaches. A detailed overview and evolution of these methods is described by Garrick *et al.* (2014).

In animal breeding, the selection of the best animals to be the parents of the next generation is performed typically to achieve a response to selection. The response to selection (*R*) is determined by the intensity of selection (*i*), the accuracy of prediction (*r*), the genetic standard deviation (σ_a) and the generation interval (*L*):

$$R = \frac{i * r * \sigma_a}{L}$$

Studies on genomic predictions have shown a solid increase in accuracy over pedigree-based predictions (BLUP). The degree of increase varies across traits, lines and species (e.g. Hayes *et al.*, 2009; Tussel *et al.*, 2013). In addition to the increase in accuracy, GS allows selection at a younger age of the selection candidates because the genotype that will be used for prediction can be obtained right after birth. Therefore, there is no need to spend a long time waiting for the expression and recording of the animals own phenotype, e.g. daily gain, or the phenotype of their offspring, e.g. milk production. This leads to a reduction in the generation interval, which is a larger benefit in some species (e.g. cattle) than in others (e.g. broilers). The potential for changing the intensity of selection with GS exists but it depends on the number of genotyped individuals; the more genotyped animals the higher the intensity and therefore a greater expected response to selection. Genomic selection, therefore, can affect response to selection through these three factors, *i*, *r* and *L*.

Genomic selection was first applied in dairy cattle (VanRaden *et al.* 2009), where the aim is to improve the performance of purebred animals. In pigs, two major pig breeding companies (PIC, Topigs Norsvin) began GS implementation in purebred lines in 2012-13. The delay in implementing GS in pigs, compared to cattle, can be attributed to: 1) the later release of the commercial SNP chip (Jan. 2008 for cattle vs Aug. 2009 for pigs), 2) the high genotyping cost compared to the value of an animal,

3) the different structure of the business (open nucleus vs. closed nucleus), 4) the need to distinguish from competitors in the market, 5) the uncertainty whether GS of purebreds results in gains in the crossbreds. The latter (crossbred production) plays an important role in pig production, and the crossbred breeding goals in pigs is probably a main difference between dairy cattle breeding and pig breeding. Implementation of GS in pigs for the crossbred breeding goals, hence, may require different strategies which are not yet fully developed. Besides the different strategies that need to be assessed, the accuracy of methods that are currently implemented for cattle may be reduced when the aim is to improve crossbred performance. Many factors affect this lower accuracy, such as the low number of genotyped crossbred individuals, genetic correlation between purebred and crossbred performance being different from 1, and the lower relationship between the purebred and crossbred individuals. Assessing accuracy of genomic prediction for the performance of purebred and crossbred animals, therefore, is a research field in development and of great interest for pig and poultry breeding companies.

1.4 Crossbreeding

Crossbreeding is the process of mating individuals from different breeds or lines to produce a crossbred offspring. It is standard practice in the modern pig production set-up, and as indicated in the preceding section, is a relevant difference compared to, for instance, dairy cattle breeding. Crossbreeding is applied to capitalize on breed complementarity and heterosis, and to protect the genetic progress in the pure lines.

Focusing on the importance of heterosis for crossbreeding, three types can be distinguished: individual, maternal and paternal (Clutter 2010). It is the individual heterosis that benefits the crossbred progeny and is a result of its own hybrid state and the primary aim for improving production traits. The maternal heterosis benefits the crossbred progeny and is a result of the hybrid state of its dam. Maternal heterosis is highly relevant for reproduction traits, e.g. mothering ability, because it benefits the offspring especially in the period that the offspring is dependent on its dam. Maternal heterosis is therefore a major reason for the extensive use of two-generation crossbreeding schemes in pig production (Bidanel 2010). The paternal heterosis benefits the crossbred progeny and is a result of the hybrid state of its sire. The benefit of paternal heterosis is limited, not having the same relevance as the maternal heterosis in crossbreeding. In general, heterosis is found across traits and species and varies roughly from 0% to 30%, including negative values as well (Bondoc *et al.* 2001; Bidanel 2010).

Dominance is labelled to be one of the main causes of heterosis (Falconer and Mackay 1996; Charlesworth and Willis 2009). This is because the hybrid superiority is attributed to the advantage of the heterozygotes over the mean of the two homozygotes. Studies in pigs and cattle have found that there is dominance variance for different traits in purebred populations (Su *et al.* 2012; Nishio and Satoh 2014; Sun *et al.* 2014). In addition, these studies have also reported that using a model that accounts for dominance resulted in either higher or similar accuracy for prediction of breeding values than using a model that only fits additive effects. Prediction of crossbred performance, accounting for dominance, has not been reported. Accounting for dominance in prediction of crossbreds is expected to result in a considerable increase of accuracy compared to purebred results because more dominance is envisaged in crossbred than purebred populations (Nishio and Satoh 2014). Therefore, using a model that accounts for dominance when crossbred individuals are used in the prediction might be important.

1.5 This thesis

The objective of my research is to exploit genomic information in purebred and crossbred pigs to generate knowledge and results that could be used to improve genetic progress. The thesis can be divided in two parts: 1) in this part the aim is to discover and investigate genomic regions that affect gestation length and boar taint, including an assessment of pleiotropic effects of the identified marker; 2) in this part the potential of genomic selection in pig breeding is investigated by determining the accuracy of genomic prediction using different training and validation populations, selected from multiple purebred lines and their crossbred offspring, and different models.

The first part of this thesis comprises Chapters 2-4 and concentrates on finding important genomic regions and test for possible application of these results in pig breeding. In **Chapter 2**, a GWAS is described with the aim to detect SNP and also to identify candidate genes that are associated with gestation length. Gestation length is an important trait in pig breeding due to its relation with maturity of the piglet at birth. Detecting significant SNP with effects on gestation length is therefore desired. In **Chapter 3**, the region of a previously detected QTL is fine-mapped, aiming at the identification of SNP that affect androstenone levels. This fine-mapped region is evaluated in **Chapter 4** for possible pleiotropic effects on production and reproduction traits in pigs. The combined results of Chapters 3 and 4 allow an informed decision on the usage of these markers in a breeding program.

The second part comprises Chapters 5-8 and focuses on strategies to implement GS in pig breeding when crossbreeding schemes are accounted for. In **Chapter 5**, the accuracy of genomic breeding values from within-, multi- and across-population predictions in pigs is evaluated, including the accuracy of using purebred training data to predict performance of crossbred pigs. This last analysis will indicate how well crossbred performance will respond to the current practice of selecting within purebred populations. For this chapter, the response variable used for training was the deregressed breeding value (DEBV) from a routine genetic evaluation, which contains a mix of purebred and crossbred animals. To separately assess the value of phenotypic information from purebred and crossbred pigs I investigated the source of information used to estimate the DEBV: should it be based on purebred or crossbred performance? Therefore, in **Chapter 6**, while the training and validation populations were the same as in Chapter 5, the training was performed twice with different phenotypes as input: first using DEBV based on purebred offspring, and second using DEBV based on crossbred offspring. The DEBV from crossbred offspring is expected to lead to better predictions of purebred animals for crossbred offspring performance. Later, more genotyped crossbred animals became available and a training population could be constructed that consisted of genotyped crossbred animals. Hence, in **Chapter 7** we compare the accuracy of prediction from using either only crossbred or only purebred animals as training population when predicting purebred animals for crossbred performance. Finally, as indicated above, the performance of crossbreds typically shows heterosis, and dominance is expected to strongly contribute to this heterosis. Therefore in **Chapter 8**, the performance of the dominance model is empirically compared to the additive model for prediction of purebreds for crossbred performance based on a training with data from crossbred pigs.

Lastly, **Chapter 9** is where the two parts, mapping and prediction, come together. I discuss the relevance of my findings, how breeders can benefit from the combination of genomic selection with the information of important genomic regions identified in GWAS. Also, I discuss the impact that high-density SNP chips and sequence data can have in GWAS studies. In addition, I expatiate on strategies for applying genomic selection, especially when crossbreeding information is used. To finalize, I give concluding remarks by summarizing the new insights from this thesis.

References

Andersson L (2001) Genetic dissection of phenotypic diversity in farm animals. Nat Rev Genet 2:130–138.

Andersson L, Haley CS, Ellegren H, et al (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science 263:1771–1774.

Animal QTLdb. http://www.animalgenome.org/QTLdb. Accessed 21 Aug 2015

Berg F, Stern S, Andersson K, et al (2006) Refined localization of the FAT1 quantitative trait locus on pig chromosome 4 by marker-assisted backcrossing. BMC Genet 7:17.

Bidanel J (2010) Biology and genetics of reproduction. In: Rothschild MF (ed) The genetics of the pig. CABI, pp 218–233

Bondoc OL, Santiago CAT, Tec JDP (2001) Least-square analysis of published heterosis estimates in farm animals. Philipp J Vet Anim Sci 27:12–26.

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nat Rev Genet 10:783–96.

Clutter AC (2010) Genetics of Performance Traits. In: Rothschild MF (ed) The genetics of the pig. CABI, pp 325–348

Darvasi A, Soller M (1995) Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141:1199–1207.

De Koning DJ, Janss LL, Rattink AP, et al (1999) Detection of quantitative trait loci for backfat thickness and intramuscular fat content in pigs (*Sus scrofa*). Genetics 152:1679–1690.

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

Duijvesteijn N, Knol EF, Merks JWM, et al (2010) A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC Genet 11:42.

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, 4th edn. Longman, Harlow, England

Fernando RL, Grossman M (1989) Marker assisted selection unbiased prediction using best linear. Genet Sel Evol 21:467–477.

Garrick D, Dekkers J, Fernando R (2014) The evolution of methodologies for genomic prediction. Livest Sci 166:10–18.

Georges M (2007) Mapping, fine mapping, and molecular dissection of quantitative trait Loci in domestic animals. Annu Rev Genomics Hum Genet 8:131–62.

Groenen MAM, Archibald AL, Uenishi H, et al (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393–8.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–43.

Ibáñez-Escriche N, Forni S, Noguera JL, Varona L (2014) Genomic information in pig breeding: Science meets industry needs. Livest Sci 166:94–100.

Kim KS, Larsen N, Short T, et al (2000) A missense variant of the porcine melanocortin-4 receptor (MC4R) gene is associated with fatness, growth, and feed intake traits. Mamm Genome 11:131–135.

Knott SA, Marklund L, Haley CS, et al (1998) Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. Genetics 149:1069–80.

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

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

Ramos AM, Crooijmans RPMA, Affara NA, et al (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.

Rothschild MF, Hu Z, Jiang Z (2007) Advances in QTL mapping in pigs. Int J Biol Sci 3:192–7.

Rothschild MF, Soller M (1997) Candidate gene analysis to detect genes controlling traits of economic importance in domestic livestock. Probe (Lond) 8:13–20.

Su G, Christensen OF, Ostersen T, et al (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 PM, Cole JB, O’Connell JR (2014) Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PLoS One 9:e103934.

Tusell L, Pérez-Rodriguez P, Forni S, et al (2013) Genome-enabled methods for predicting litter size in pigs : a comparison. Animal 7:1739–1749.

Van Laere A-S, Nguyen M, Braunschweig M, et al (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425:832–836.

VanRaden PM, Van Tassell CP, Wiggans GR, et al (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16–24.

2

Genome-wide association study reveals regions associated with gestation length in two pig populations

A.M. Hidalgo[1,2], M.S. Lopes[1,3], B. Harlizius[3], J.W.M. Bastiaansen[1]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6708WD, the Netherlands; [2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, 750 07, Sweden; [3] Topigs Norsvin, Beuningen 6640AA, the Netherlands

Animal Genetics (Accepted)

Abstract

Reproduction traits, such as gestation length (GLE), play an important role in dam line breeding in pigs. The objective of our study was to identify single nucleotide polymorphisms (SNP) that are associated with GLE in two pig populations. Genotypes and deregressed breeding values were available for 2,081 Dutch Landrace-based (DL) and 2,301 Large White-based (LW) pigs. We identified two QTL regions for GLE, one in each line. For DL, three associated SNP were detected in one QTL region spanning 0.52 Mbp on *Sus scrofa* chromosome (SSC) 2. For LW, four associated SNP were detected in one region of 0.14 Mbp on SSC5. The region on SSC2 contains the heparin-binding EGF-like growth factor (*HBEGF*) gene which promotes embryo implantation and has been described to be involved in embryo survival throughout gestation.The associated SNP can be used for marker-assisted selection in the studied populations, and further studies of *HBEGF* gene are warranted to investigate its role in GLE.

Key words: *HBEGF* gene, length of pregnancy, quantitative trait loci, reproduction trait

Reproduction traits, such as gestation length (GLE), play an important role in dam line breeding in pigs (Hanenberg *et al.* 2001; Onteru *et al.* 2012). GLE is defined as the interval (days) between insemination and farrowing. The last days of gestation are crucial for the maturation of the piglet at birth, therefore a gestation that is not shorter than the average (~114 days) will result in better development of the piglet at birth and lower postpartum mortality (Rydhmer *et al.* 2008). GLE has positive genetic and phenotypic correlation with mothering ability (Hanenberg *et al.* 2001) and a longer gestation is also favourably linked with birth weight and piglet growth rate (Rydhmer *et al.* 2008).

Previous studies using microsatellite markers and divergent crosses have identified quantitative trait loci (QTL) for GLE (Wilkie *et al.* 1999; Chen *et al.* 2010). These QTL, however, are based on linkage maps with few genetic markers resulting in large confidence intervals. With the development of dense panels of single nucleotide polymorphisms (SNP) we have now the opportunity to narrow down the confidence interval and to identify novel associated variants via genome-wide association studies (GWAS). Apart from Onteru *et al.* (2012), who reported many associations spread across the pig genome, there is a lack of GWAS for GLE. The objective of our study was to identify markers that are associated with GLE in two commercial pig populations and then identify candidate genes that lie within the associated QTL region.

Genotypes were available on animals from two dam lines: 2,081 Dutch Landrace-based (DL) and 2,301 Large White-based (LW). All animals were genotyped using the Illumina PorcineSNP60 BeadChip (Ramos *et al.* 2009). We excluded SNP with call rate <0.95, minor allele frequency <0.01, strong deviation from Hardy-Weinberg equilibrium (χ^2>600), GenCall<0.15, unmapped SNP and SNP located on sex chromosomes, according to the Sscrofa10.2 assembly of the reference genome (Groenen *et al.*, 2012). After quality control, the remaining missing genotypes were imputed using BEAGLE 3.3.2 (Browning & Browning 2007). This quality control was performed for each population separately, leaving 40,776 SNP for DL and 42,244 SNP for LW of the initial 64,232 SNP. None of the animals had more than 5% missing genotypes.

Phenotypes consisted of repeated observations of GLE for the DL and LW populations (Table S2.1). GLE was normally distributed and its mean differed by 0.85 days between populations. The phenotypes were used to estimate the breeding values (EBV) in a single-trait analysis using a repeatability model in ASReml 3.0

(Gilmour *et al.* 2009). The model used to estimate the EBV included the fixed effects of genetic line, parity number, total number of piglets born, multiple inseminations performed (yes or no) and herd-year-season. Random effects included were service sire, a permanent environmental effect and an additive genetic effect. Deregressed EBV (DEBV) of the genotyped animals were used as response variable in this study. EBV were deregressed using the methodology proposed by Garrick *et al.* (2009).

A single-SNP GWAS was performed in ASReml 3.0 (Gilmour *et al.* 2009) using the following model:

$$\mathbf{y} = \mu + \mathbf{b}_1\mathbf{SNP} + \mathbf{Zg} + \mathbf{e},$$

where **y** is the vector of DEBV of the genotyped animals, μ is the overall mean, $\mathbf{b}_1$ is the vector of regression coefficients of each SNP, **SNP** is the incidence vector for $\mathbf{b}_1$ with genotypic information (0, 1 and 2), **Z** is the incidence matrix for **g**, **g** is the vector of random additive genetic effects, assumed to be $\sim N(\mathbf{0}, \mathbf{G}\sigma_a^2)$, where **G** is the genomic relationship matrix, and e is the residual error, assumed to be $\sim N(\mathbf{0}, \mathbf{D}\sigma_e^2)$, where **D** is a diagonal matrix calculated as $\mathbf{I}*w_{i,}$ where **I** is an identity matrix and w_i is the weight of the i^{th} DEBV based on its reliability. The **G** matrix was built as $\mathbf{G} = \mathbf{MM}'/2\sum p_i q_i$, where **M** is a matrix of centered genotypes and p_i and q_i are the allelic frequencies of the i^{th} marker based on observed genotypes (VanRaden 2008).

The genomic inflation factor (lambda) for the distribution of *p*-values from the GWAS was estimated using the R package GenABEL (Aulchenko *et al.* 2007). A genome-wide false discovery rate (FDR) was applied to correct for multiple testing using the R package qvalue (Dabney & Storey 2015). A FDR threshold (FDR ≤ 0.05) was applied. We used the Haploview software (Barrett *et al.* 2005) to compute the LD (linkage disequilibrium) between significant markers.

The QTL variance (σ^2_{QTL}) was estimated according to Falconer & Mackay (1996). The proportion of the DEBV variance explained by the QTL region was estimated by:

$$\left(\sigma^2_{\mathrm{QTL}}/\sigma^2_{\mathrm{DEBV}}\right)*100,$$

where σ^2_{DEBV} is the total variance of the DEBV estimated based on the GWAS model without a SNP effect.

We searched for genes that were located nearby the significant markers using Ensembl database (www.ensembl.org). All genes within ± 1 Mbp from the QTL region (associated SNP peak) were selected for further investigation.

The lambda values of 1.03 for DL and 1.06 (Fig. S2.1) for LW suggest that there is no population stratification in our data indicating no increased risk for type I errors. Given our stringent false discovery rate threshold the reported associations are likely to be true associations.

For DL, three significant SNP were detected in one QTL region spanning 0.52 Mbp (Fig. 2.1A and Table 2.1). This QTL region was located on *Sus scrofa* chromosome (SSC) 2 between 147.8 Mbp and 148.4 Mbp. The percentage of DEBV variance explained by the QTL region was 1.12%. For LW, four significant SNP were identified in one QTL region spanning 0.14 Mbp (Fig. 2.1B and Table 2.1). This QTL region was located on SSC5 between 2.9 Mbp and 3.1 Mbp. The percentage of DEBV variance explained by this QTL region was 0.77%. The minor allele frequency (MAF) of the significant SNP on SSC5 was low, 0.01, which made us cautious about these results. Associations may appear by chance due to an extreme observation in the low frequency genotype class. However, the distribution of animals per genotype class (0 AA, 65 AB, and 2,236 BB) gives us confidence in this association, as the only contrast is between AB and BB genotypes that each have a reasonable number of observations. There was high LD between markers of the same peak in each population (Fig. S2.2), indicating that all markers from each peak belong to one LD block and are likely capturing variance from the same QTL.

Table 2.1 Significant SNPs genome-wide associated with gestation length for the two populations under study.

SNP	Pop.	Chr[a]	Pos[b]	Location	Effect[c]	$-\log_{10}$(p-value)	q-value[d]	MAF[e]
ASGA0012523	DL	2	147.83	Intergenic	0.30	7.02	0.003	0.14
DBMA0000166	DL	2	148.26	Intronic	0.27	6.86	0.003	0.15
DIAS0004579	DL	2	148.35	5’ UTR	0.22	5.46	0.048	0.18
INRA0018114	LW	5	2.92	Intergenic	-0.65	6.57	0.004	0.01
ALGA0029956	LW	5	2.95	Intergenic	-0.65	6.57	0.004	0.01
MARC0027217	LW	5	2.98	Intergenic	-0.63	6.00	0.010	0.01
H3GA0015235	LW	5	3.06	Intronic	-0.65	6.57	0.004	0.01

[a] Chromosome

[b] Position on the chromosome (Mbp)

[c] Allele substitution effect

[d] FDR-based q-value

[e] Minor allele frequency

DL: Dutch Landrace, LW: Large White

The QTL on SSC2 for DL and on SSC5 for LW were not previously reported. Onteru *et al.* (2012) have reported significant SNP for GLE on SSC2 and on SSC5, however they are not located near the region identified in our study. Chen *et al.* (2010), using microsatellite markers, have also detected one QTL region on SSC2, however that peak also does not overlap with ours.

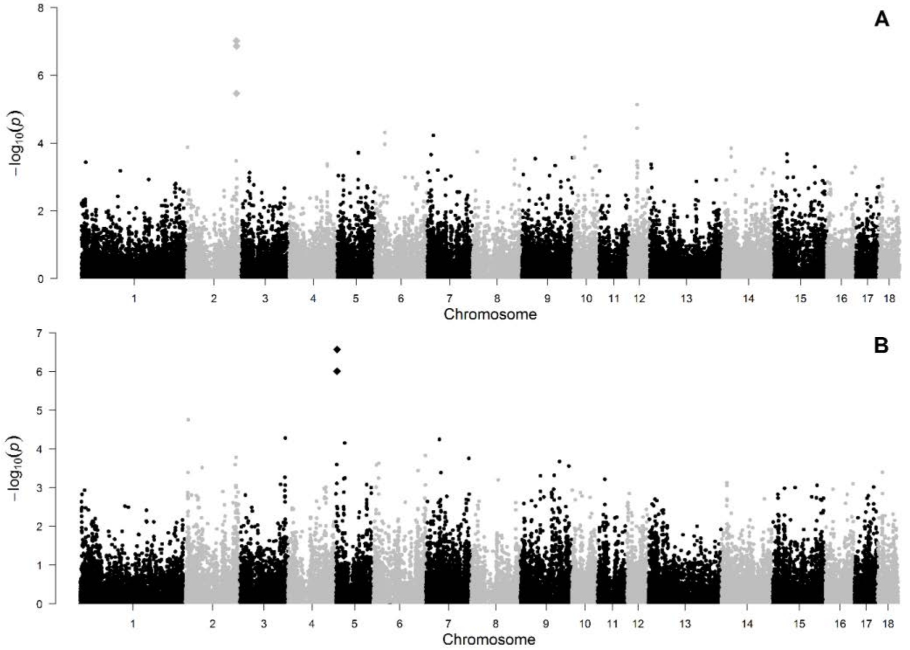

Figure 1 (A) Manhattan plot for gestation length for the Dutch Landrace population. (B) Manhattan plot for gestation length for the Large White population. The significant SNP (q-value ≤ 0.05) are shown as large diamonds.

The biology of parturition is not completely understood and differs considerably between species (Bezold *et al.* 2013). Candidate genes known from physiology or candidate gene studies in human and model organisms (Bezold *et al.* 2013) do not map to this region. Candidate genes were identified using the synteny tool of the Ensembl genome browser at http://www.ensembl.org/Sus_scrofa/Info/Index, build 10.2 of the pig reference genome sequence. For the region identified on SSC2, 73 genes with a gene name in the porcine annotation or homologous region in human were located within the QTL area ± 1 Mbp. The heparin-binding EGF-like growth factor (*HBEGF*) gene located between 148.04 Mbp and 148.06 Mbp has been suggested to promote embryo implantation in humans and mice (Leach *et al.*1999; Xie *et al.* 2007) and is expected to have an important role as survival factor throughout gestation (Jessmon *et al.* 2009). Xie *et al.* (2007) studying mice with a deletion of uterine *HBEGF* showed that it resulted in delayed implantation and compromised term pregnancy. For the region identified on SSC5, 25 genes were identified within the QTL region ± 1 Mbp. No gene, however, appeared to be a relevant candidate gene based on their currently known functions.

In summary, we have detected two QTL regions for GLE. Different populations had different QTL regions. SNP located in these QTL regions might be useful for marker-assisted selection in the studied populations. The *HBEGF* gene, suggested to promote embryo implantation, is the most compelling candidate gene for the QTL region on SSC2.

Acknowledgements

AMH currently benefits from a joint grant from the European Commission (within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”), Breed4Food and Topigs Norsvin.

References

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. Bioinformatics 23:1294–6.

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–5.

Bezold K, Karjalainen M, Hallman M, et al (2013) The genomics of preterm birth: from animal models to human studies. Genome Med 5:34.

Browning SR, Browning BL (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 81:1084–97.

Chen CY, Guo YM, Zhang ZY, et al (2010) A whole genome scan to detect quantitative trait loci for gestation length and sow maternal ability related traits in a White Duroc × Erhualian F2 resource population. Animal 4:861–6.

Dabney A, Storey JD (2015) qvalue: Q-value estimation for false discovery rate control.

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, 4th edn. Longman, Harlow, England

Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 41:55.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.

Groenen MAM, Archibald AL, Uenishi H, et al (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393–8.

Hanenberg EHAT, Knol EF, Merks JWM (2001) Estimates of genetic parameters for reproduction traits at different parities in Dutch Landrace pigs. Livest Prod Sci 69:179–186.

Jessmon P, Leach RE, Armant DR (2009) Diverse functions of *HBEGF* during pregnancy. Mol Reprod Dev 76:1116–1127.

Leach RE, Khalifa R, Ramirez ND, et al (1999) Multiple roles for heparin-binding epidermal growth factor-like growth factor are suggested by its cell-specific expression during the human endometrial cycle and early placentation. J Clin Endocrinol Metab 84:3355–3363.

Onteru SK, Fan B, Du Z-Q, et al (2012) A whole-genome association study for pig reproductive traits. Anim Genet 43:18–26.

Ramos AM, Crooijmans RPMA, Affara NA, et al (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.

Rydhmer L, Lundeheim N, Canario L (2008) Genetic correlations between gestation length, piglet survival and early growth. Livest Sci 115:287–293.

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–23.

Wilkie PJ, Paszek AA, Beattie CW, et al (1999) A genomic scan of porcine reproductive traits reveals possible quantitative trait loci (QTLs) for number of corpora lutea. Mamm Genome 10:573–578.

Xie H, Wang H, Tranguch S, et al (2007) Maternal heparin-binding-EGF deficiency limits pregnancy success in mice. Proc Natl Acad Sci U S A 104:18315–18320.

Supplementary material

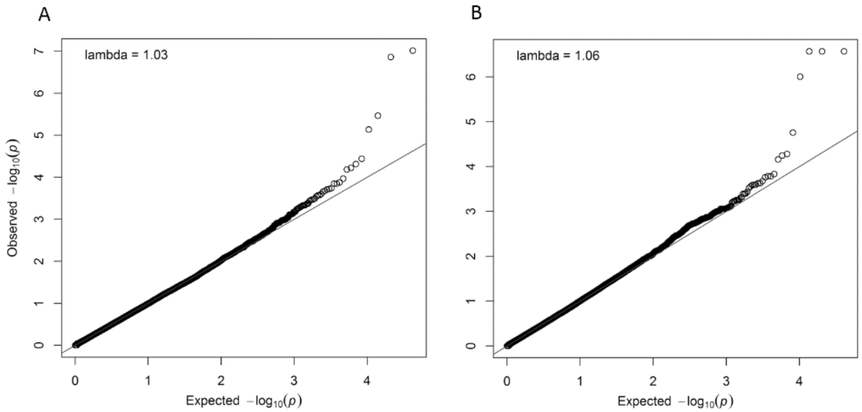

Figure S2.1 Quantile-quantile (Q-Q) plot with lambda values. (A) Q-Q plot for the Dutch Landrace population. (B) Q-Q plot for the Large White population.

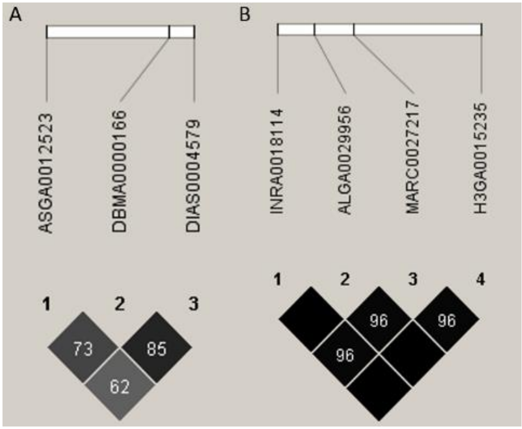

Figure S2.2 Linkage disequilibrium (LD) between significant markers. (A) LD plot (r^2) of the significant SNP in the Dutch Landrace line. (B) LD plot of the significant SNP in the Large White line.

Table S2.1 Descriptive statistics of gestation length for the two lines under study.

Line	Nr. Observations	Nr. Animals	Mean (SD[a])	Minimum	Maximum
Dutch Landrace	236,803	68,070	115.79 (1.53)	105	125
Large White	134,477	41,522	114.94 (1.53)	105	125

[a] Standard deviation

3

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

A.M. Hidalgo[1,2], J.W.M. Bastiaansen[1], B. Harlizius[3], H.J. Megens[1], O. Madsen[1], R.P.M.A. Crooijmans[1], M.A.M. Groenen[1]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6708WD, the Netherlands; [2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, 750 07, Sweden; [3] Topigs Norsvin, Beuningen 6640AA, the Netherlands

BMC Genetics (2014) 15:4

Abstract

Background

Androstenone is one of the major compounds responsible for boar taint, a pronounced urine-like odor produced when cooking boar meat. Several studies have identified quantitative trait loci (QTL) for androstenone level on *Sus scrofa* chromosome (SSC) 6. For one of the candidate genes in the region *SULT2A1*, a difference in expression levels in the testis has been shown at the protein and RNA level.

Results

Haplotypes were predicted for the QTL region and their effects were estimated showing that haplotype 1 was consistently related with a lower level, and haplotype 2 with a higher level of androstenone. A recombinant haplotype allowed us to narrow down the QTL region from 3.75 Mbp to 1.94 Mbp. An RNA-seq analysis of the liver and testis revealed six genes that were differentially expressed between homozygotes of haplotypes 1 and 2. Genomic sequences of these differentially expressed genes were checked for variations within potential regulatory regions. We identified one variant located within a CpG island that could affect expression of *SULT2A1* gene. An allele-specific expression analysis in the testis did not show differential expression between the alleles of *SULT2A1* located on the different haplotypes in heterozygous animals. However a synonymous mutation C166T (SSC6: 49,117,861 bp in Sscrofa 10.2; *C/T*) was identified within the exon 2 of *SULT2A1* for which the haplotype 2 only had the *C* allele which was higher expressed than the *T* allele, indicating haplotype-independent allelic-imbalanced expression between the two alleles. A phylogenetic analysis for the 1.94 Mbp region revealed that haplotype 1, associated with low androstenone level, originated from Asia.

Conclusions

Differential expression could be observed for six genes by RNA-seq analysis. No difference in the ratio of *C*:*T* expression of *SULT2A1* for the haplotypes was found by the allele-specific expression analysis, however, a difference in expression between the *C* over *T* allele was found for a variation within *SULT2A1*, showing that the difference in androstenone levels between the haplotypes is not caused by the SNP in exon 2.

Key words: Asian haplotype, boar taint, RNA-seq, SSC6, whole genome sequencing

3.1 Background

Androstenone (5α-androst-16-en-3-one) is a steroid hormone synthesized in the Leydig cells of the testis in a stepwise conversion involving 3β-hydroxysteroid dehydrogenase (HSD) and 5α-reductase enzymes (Dufort *et al.* 2001). In pigs, androstenone acts as a sex pheromone which attracts female pigs making them more receptive to mating (Dorries *et al.* 1997). Androstenone is degraded in the liver and salivary gland by 3α-HSD enzymes resulting in α-androstenol and by 3β-HSD enzymes resulting in β-androstenol (Dorries *et al.* 1997; Dufort *et al.* 2001; Sinclair *et al.* 2005). Sulfoconjugated androstenols are eliminated mainly in the urine and bile. Androstenone is one of the major compounds responsible for boar taint, a pronounced urine-like odor produced when cooking meat from intact male pigs, or boar meat (Bonneau 1982). As unconjugated androstenone and androstenol are the forms that most easily accumulate in adipose tissue and hereby lead to boar taint (Sinclair and Squires 2005), conjugation plays an important role in the prevention of boar taint. At high concentrations in the fat, androstenone influences consumer acceptability of pork (Bonneau and Chevillon 2012). In current breeding practice, castration of male piglets is used to prevent the boar taint. Castration, however, is undesirable not only for technical reasons, as castrated male pigs have fatter carcasses and reduced feed efficiency (Seideman *et al.* 1982), but also because of animal welfare concerns and future legislation restriction. Therefore, the development of an alternative to castration is needed.

The development of a medium-density 60K porcine single nucleotide polymorphism (SNP) chip (Ramos *et al.* 2009), has enabled genome-wide association studies (GWAS) to efficiently map regions throughout the genome affecting phenotypic traits such as the androstenone level. While GWAS can identify significant marker associations, the current SNP density on the Illumina PorcineSNP60 BeadChip often leads to clusters of markers covering a region that is still too large to allow accurate identification of the responsible genes or variants. Hence, there is still the need to reduce the size of these clusters if the aim is to find causative relations between gene(s) or variants that affect phenotypic traits like androstenone level in fat.

Several studies (Duijvesteijn *et al.* 2010; Grindflek *et al.* 2011a; Grindflek *et al.* 2011b; Gregersen *et al.* 2012) have identified quantitative trait loci (QTL) for androstenone level on *Sus scrofa* chromosome (SSC) 6. Duijvesteijn *et al.* (2010) performed a GWAS unveiling an 8 Mbp region on SSC6 associated with androstenone level in boars of a Duroc-based population. Similarly, Grindflek *et al.* (2011a) reported a QTL for Duroc animals within a 7.1 Mbp region that overlaps with the region found by Duijvesteijn *et al.* (2010). Within the QTL region on SSC6, Duijvesteijn *et al.* (2010) showed that there are haplotypes related to low and high average levels of androstenone in fat. Expression studies that compared boars with low and high androstenone levels (Moe *et al.* 2007; Grindflek *et al.* 2010; Leung *et al.* 2010) found differential expression of several genes including sulfotransferase family 2A dehydroepiandrosterone-preferring member 1 (*SULT2A1*). *SULT2A1* is located within the QTL region and is a strong candidate gene to have an effect on androstenone level. Since QTL regions are large, fine mapping studies need to be carried out to identify causative variants and to enable the use of these QTL in breeding programs.

The goal of this study was to narrow down the QTL region on SSC6 previously reported by Duijvesteijn *et al.* (2010), to identify and characterize genes and SNP variants that affect androstenone level in pigs, and to determine whether the effects of low- and high-androstenone haplotypes are caused by differential expression of *SULT2A1*.

3.2 Results

Androstenone level was obtained from 2,750 boars that belonged to six purebred populations and five crosses. The flowchart (Fig. 3.1) provides an overview of the study to clarify the steps that were taken in association mapping, whole genome sequencing, and functional analyses.

3.2.1 Region of interest and haplotypes

Markers associated with Androstenone were identified by Duijvesteijn *et al.* (2010) in the region from position 36,907,969 bp to 44,939,360 bp on SSC6 using the Sscrofa9 assembly of the reference genome. A target region of 2.8 Mbp, from position 36,907,969 bp to 39,697,649 bp, containing the peak associations, was defined by Duijvesteijn *et al.* (2010) and was used in our study. The present study used the Sscrofa10.2 assembly of the reference genome (Groenen *et al.* 2012), in which the 2.8 Mbp region in genome build Sscrofa9 corresponded to a 3.75 Mbp region on SSC6, from position 48,585,961 bp to 52,336,598 bp. This region contained two linkage disequilibrium (LD) blocks with a total of 46 markers on the Illumina PorcineSNP60 BeadChip that were polymorphic in our study (Table S3.1), 29 of which are significant for androstenone level, identical to the ones reported by Duijvesteijn *et al.* (2010). Prediction of the haplotypes for the 46 SNP across populations revealed ten haplotypes with a frequency above 1%.

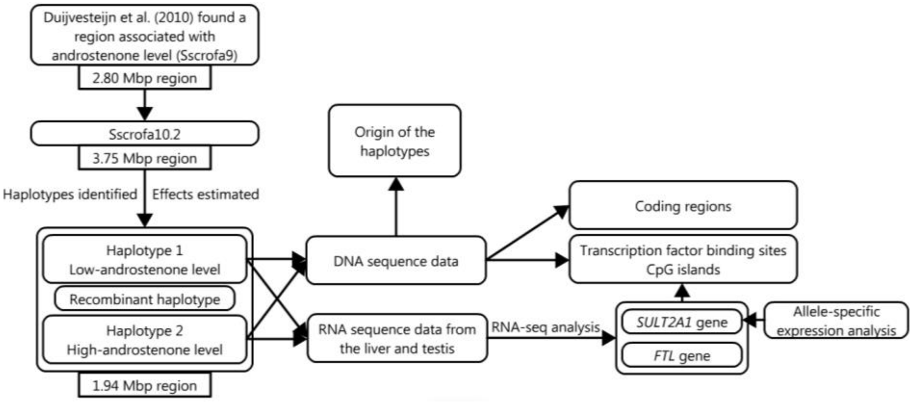
Duijvesteijn et al. (2010) found a region associated with androstenone level (Sscrofa9)
2.80 Mbp region
Sscrofa10.2
3.75 Mbp region
Haplotypes identified
Effects estimated
Haplotype 1 Low-androstenone level
Recombinant haplotype
Haplotype 2 High-androstenone level
1.94 Mbp region
Origin of the haplotypes
DNA sequence data
RNA sequence data from the liver and testis
RNA-seq analysis
Coding regions
Transcription factor binding sites CpG islands
SULT2A1 gene
FTL gene
Allele-specific expression analysis

Figure 3.1 Flowchart of the steps taken in the current study. We started the analysis with a region associated with androstenone level previously found on SSC6. The position of the markers were adjusted to the Sscrofa10.2 assembly of the reference genome. Afterwards, haplotypes for the region were identified and their effects were estimated. The region was narrowed down using information from a recombinant haplotype. RNA-seq analysis, within the narrowed region, was performed in the liver and testis. For candidate genes that were differentially expressed, we used whole genome sequence data to look for variations within regulatory regions and also to look for variations within coding regions for all genes within the narrowed region. Allele-specific expression analysis in the testis was performed for *SULT2A1* gene because a variation is located within a regulatory region. Finally, we analyzed genomic sequence data to assess the origin of the haplotypes.

3.2.2 Effects and association analysis of haplotypes

Effects on androstenone level were estimated for the ten haplotypes (Table S3.2). Haplotypes 1 and 2 were present in all populations; haplotype 1 was consistently related with a lower level, and haplotype 2 with a higher level of androstenone (Fig. 3.2).

A phylogenetic tree was constructed using MEGA 5 (Tamura *et al.* 2011) based on similarities among the 46 SNP of the haplotypes (Fig. 3.3A). Haplotypes were arranged in two groups, with haplotypes 1, 3, 7, 8 and 10 forming one group, and haplotypes 2, 4, 5, 6 and 9 another group. To determine whether the relation of haplotypes 1 and 2 to androstenone level followed the phylogenetic division, we analyzed the association between haplotypes and androstenone using Treescan (Posada *et al.* 2005) which "cuts" the phylogenetic tree at different branches and tests whether the groups created by the cut are statistically different in their effect on the phenotype. Dividing the tree between haplotypes 1, 3, 8, 10, and 7, 5, 9, 6, 2, 4; or between haplotypes 1, 3, 8, 10, 7, and 5, 9, 6, 2, 4 resulted in statistically significant differences (P<0.0001) and explained the largest proportion of phenotypic variation (0.049 and 0.051, respectively).

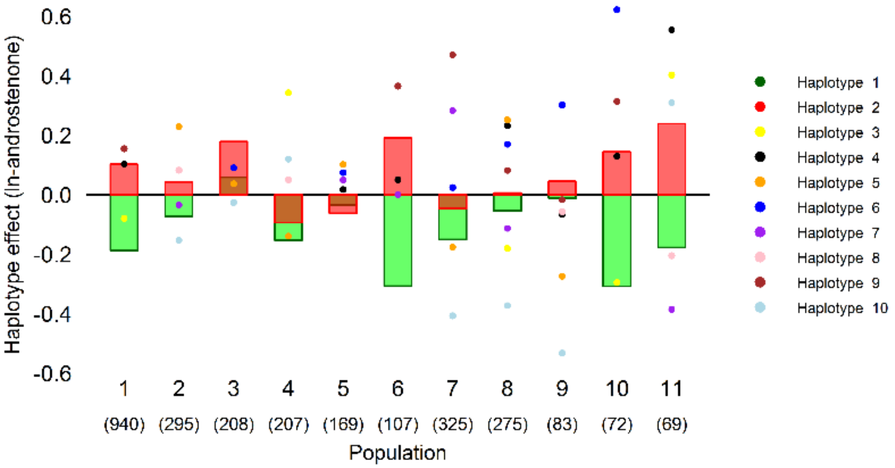

Figure 3.2 Haplotype effects across all populations. Effects estimated for the haplotypes 1 and 2 (bars) and eight other haplotypes (dots) across 11 populations (number of animals).

Haplotype 7 is a recombinant haplotype: the region from SNP 1 to 13 is similar to haplotype 2 (high androstenone level), whereas the region from SNP 14 to 46 is similar to haplotype 1 (low androstenone level) (Fig. 3.3B). Therefore, with a posterior analysis, the effects of haplotypes 1, 2, and 7 were estimated using only populations in which haplotype 7 is segregating (N = 1240). The effect of haplotype 7 (effect = 0.05) is significantly different from haplotype 1 (effect = -0.16), whereas it is not significantly different from haplotype 2 (effect = 0.00). The effect of haplotype 7 could therefore be grouped together with haplotype 2, indicating that the region from haplotype 7 that is similar to haplotype 2 causes the effect. This allowed us to narrow down the associated region from 3.75 Mbp (48,585,961 bp - 52,336,598 bp) to 1.94 Mbp (48,317,509 bp – 50,259,057 bp). When testing the remaining haplotypes 3, 4, 5, 6, 8, 9, and 10 in the same way, their effects were all congruent with the expectation, except for the rare haplotypes 8 and 10 which were not significantly different from either haplotype 1 or 2 (Table S3.3).

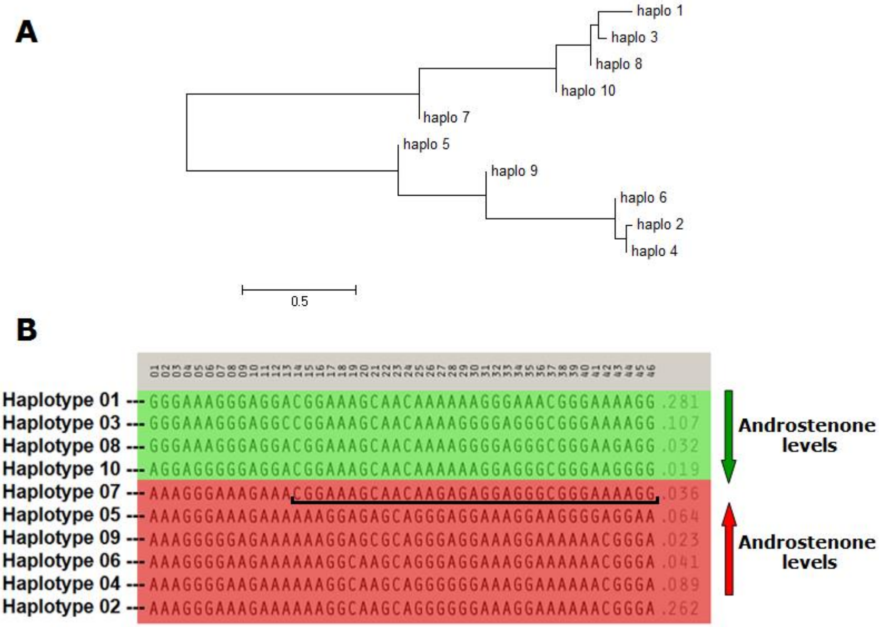
A
haplo 1
haplo 3
haplo 8
haplo 10
haplo 7
haplo 5
haplo 9
haplo 6
haplo 2
haplo 4
0.5
B
01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
Haplotype 01 --- GGGAAAGGGAGGACGGAAAGCAACAAAAAAGGGAAACGGGAAAAGG .281
Haplotype 03 --- GGGAAAGGGAGGCCGGAAAGCAACAAAAGGGGAGGGCGGGAAAAGG .107
Haplotype 08 --- GGGAAAGGGAGGACGGAAAGCAACAAAAGGGGAGGGCGGGAAGAGG .032
Haplotype 10 --- AGGAGGGGGAGGACGGAAAGCAACAAAAAAGGAGGGCGGGAAGGGG .019
Haplotype 07 --- AAAGGGAAAGAAACGGAAAGCAACAAGAGAGGAGGGCGGGAAAAGG .036
Haplotype 05 --- AAAGGGAAAGAAAAAAGGAGAGCAGGGAGGAAAGGAAGGGGAGGAA .064
Haplotype 09 --- AAAGGGGGAGAAAAAAAGGAGCGCAGGGAGGAAAGGAAAAAACGGA .023
Haplotype 06 --- AAAGGGGAAGAAAAAAAGGCAAGCAGGGAGGAAAGGAAAAAACGGA .041
Haplotype 04 --- AAAGGGGAAGAAAAAAAGGCAAGCAGGGGGGAAAGGAAAAAACGGA .089
Haplotype 02 --- AAAGGGAAAGAAAAAAAGGCAAGCAGGGGGGAAAGGAAAAAACGGA .262
Androstenone levels
Androstenone levels

Figure 3.3 Phylogeny of the 10 haplotypes. (**A**) Phylogenetic tree for the 10 most frequent haplotypes and (**B**) the 10 most frequent haplotypes across all populations ordered according to the phylogenetic tree and colored according to their effect on androstenone level. The underlined region indicates similarity between haplotype 7 and low-androstenone haplotypes.

3.2.3 RNA-seq analysis

To determine whether genes within the narrowed region were differentially expressed between haplotypes 1 and 2, an RNA-seq analysis was performed in the liver and testis. A "haplotype-1 pool" was made from four animals homozygous for haplotype 1, and a "haplotype-2 pool" from four animals homozygous for haplotype 2. After the rearrangement of the map in the new reference genome assembly build10.2 it was found that the haplotype-2 pool contained six copies of haplotype 2 and two copies that were haplotype-2-like (rare haplotypes that differed from haplotype 2 in three positions).

A total of 79 genes were located within the narrowed region. Among these, three genes (4%) were found to be differentially expressed in the liver: sperm acrosome membrane-associated protein 4 (*SPACA4*), synaptogyrin 4 (*SYNGR4*), tubby-like protein 2 (*TULP2*); and three genes (4%) in the testis: ferritin light polypeptide (*FTL*), glioma tumor suppressor candidate region gene 1 (*GLTSCR1*), and sulfotransferase family 2A dehydroepiandrosterone-preferring member 1 (*SULT2A1*) (Table 3.1).

3.2.4 Functional analysis

Whole-genome sequencing data were used to investigate the complete set of SNP variants in the narrowed region. From the 55 animals for which the whole-genome sequencing data were available (see M&M for details), 10 animals were homozygous for either haplotype 1 or 2 according to the 13 SNP located within the 1.94 Mbp interval. An animal was considered to be homozygous when it met two criteria: (1) average heterozygosity for the narrowed region was very low (Fig. S3.1); (2) the genotypes of the 13 SNP overlapping with the Illumina PorcineSNP60 BeadChip in the region were identical to the sequencing data.

Table 3.1 Results for genes differentially expressed in the liver and testis between pools of animals homozygous for low- and high-androstenone haplotypes.

Gene	Location	Haplo-1 Pool	Haplo-2 Pool	Log2 FC*	P value
		Liver			
SYNGR4	49625413-49627375	0.71	7.81	3.47	3.2 x10^{-4}
SPACA4	49711175-49719812	2.00	11.76	2.56	2.21 x10^{-3}
TULP2	50142849-50151782	0.67	6.45	3.27	6.72 x10^{-6}
		Testis			
GLTSCR1	48928610-48946950	4.39	1.88	-1.22	1.34 x10^{-6}
SULT2A1	49108566-49119941	40.06	104.83	1.39	1.41 x10^{-7}
FTL	50078375-50097694	782.79	1618.46	1.05	1.95 x10^{-5}

*Fold changes (FC) are calculated relative to low-androstenone haplotypes, hence indicates the times of up-regulation in the high-androstenone group compared to the low-androstenone haplotype.

3.2.4.1 Coding regions

From the genome sequencing data, 1,897 single nucleotide differences were found between haplotypes 1 and 2 in the 1.94 Mbp interval. To detect functional genetic variants between the haplotypes, differences were annotated using ANNOVAR (Wang *et al.* 2010). Of the 1,897 SNP, 75 (3.95%) were located in exonic regions, with 17 of them being non-synonymous and 58 being synonymous variations (Table S3.4). Within the five genes previously suggested as candidates for effects on androstenone by Duijvesteijn *et al.* (2010) and Grindflek *et al.* (2010), we found three synonymous and one non-synonymous variation. Gene *SULT2A1* contained one synonymous variation (C/T) within exon 2 at position 166 (SSC6: 49,117,861 bp in Sscfrofa10.2); Hydroxysteroid (17-beta) dehydrogenase 14 (*HSDB17B14*) contained a non-synonymous variation (T/G) within exon 4 at position 217 (SSC6: 49,889,443 bp in Sscfrofa10.2); Lutropin subunit beta (*LHB*) contained a synonymous variation (C/T) within exon 1 at position 147 (SSC6: 50,064,434 bp in Sscfrofa10.2); *FTL* contained a synonymous variation (C/G) within exon 4 at position 474 (SSC6:

50,096,119 bp in Sscrofa10.2); and for sulfotransferase family cytosolic, 2B member 1 (*SULT2B1*) no nucleotide variation was found in any of the exons.

The impact of the non-synonymous variations was assessed using PolyPhen2 (Adzhubei *et al.* 2010), which showed that the T/G variation in *HSDB17B14* was unlikely to affect the function of the protein. Regarding other genes within the 1.94 Mbp interval that have non-synonymous variations but are not considered as a candidate gene, a variation within exon 1 of fucosyltransferase 1 (*FUT1*) is probably damaging the functionality of the protein (PSIC score: 2.092). Probably-damaging status indicates that the variation is, with high confidence, expected to affect protein function.

3.2.4.2 Regulatory regions

Because *SULT2A1* and *FTL* were differentially expressed between pools of haplotype 1 and haplotype 2 animals and are functional candidate genes, their up and downstream sequences (±2,000 bp) were examined for the presence of potential transcription factor binding sites (TFBS) that were conserved across three species (*Sus scrofa*, *Bos taurus*, and *Homo sapiens*). None of the fixed differences between haplotypes 1 and 2 were located within predicted TFBS. In addition to the absence of single nucleotide differences between haplotypes 1 and 2 within the TFBS, we checked the overlap of the TFBS with copy number variations (CNVs) identified by Paudel *et al.* (2013). No CNVs were identified that could affect TFBS near *SULT2A1* and *FTL* genes.

CpG islands were predicted for the *SULT2A1* and *FTL* genes including their up and downstream sequence (±2000 bp). Two CpG islands of at least 200 bp, 50% of GC content, and 60% of average observed-to-expected ratio of *C* plus *G* were detected for each of the genes. One of the four CpG islands (49,110,687 bp - 49,110,889 bp) which were predicted for *SULT2A1* contained a SNP (*C/G*, 49,110,873 bp). This CpG island had 18 CpGs. None of the identified CpG islands contained CNVs.

3.2.5 Validation of differential *SULT2A1* expression in the testis

Of the genes that were differentially expressed between the low- and high-androstenone pools, *SULT2A1,* based on its function*,* is a particularly strong positional candidate gene. To validate the difference in expression in the testis found between the haplotype-1 and haplotype-2 pools, we made use of a synonymous SNP (*C/T*; SSC6: 49,117,861 bp in Sscfrofa10.2 – see details below on detection of this SNP) within exon 2 of the *SULT2A1* gene described by Sinclair *et al.* (2006). The *C/T* variation was not in complete LD with the low- and high-androstenone haplotypes. The *T* allele had a frequency of 0.34 and was only found on the haplotype 1, the *C* allele was found on both low- and high-androstenone haplotypes (2, 3, and 4). It allowed for a comparison of not only the ratio of *C*:*T* expression between low- and high-androstenone haplotypes but also between the *SULT2A1* alleles. The QTL effects of the *C* and *T* alleles were investigated and both the *T* and *C* alleles that were located on low-androstenone haplotypes were significantly different from *C* alleles that were located on high-androstenone haplotypes while within the low androstenone haplotypes the *C* and *T* alleles were not different (Table S3.5). To certify that the allele-specific expression analysis was sensitive enough to detect the expression differences found in the RNA-seq data, genomic DNA from two animals, homozygous for the *C* or *T* allele, were mixed in seven ratios and in three concentrations. The result of this analysis showed high sensitivity to distinguish between the expected difference in expression levels and a strong linear relation between the observed and expected ratios (R^2 in three concentrations: 2.5ng = 0.93, 10ng = 0.72, 40ng = 0.97) (Fig. S3.2).

Heterozygous animals (C/T) were one of two different androstenone diplotypes (low/low or high/low) but there was no difference in the ratio of *C*:*T* expression between the low/low and high/low diplotype. A difference in expression was however observed between the *C* and *T* alleles of *SULT2A1* with the *C* allele always being higher expressed than the *T* allele (ratio *C:T* = 1.5:1, s.d. = 0.13, Fig. 3.4). The mean ratio of 1.5:1 was calculated based on the 67 heterozygous animals studied in the allele-specific expression analysis. Genotyping of this SNP on the animals from the pools used in the RNA-seq analysis showed that the haplotype-1 pool contained 4*C* and 4*T* alleles whereas the haplotype-2 pool contained only *C* alleles. Thus, the observed difference in expression between haplotype-1 and haplotype-2 pools in *SULT2A1* expression could be related to differences in expression of the *C/T* alleles in exon 2.

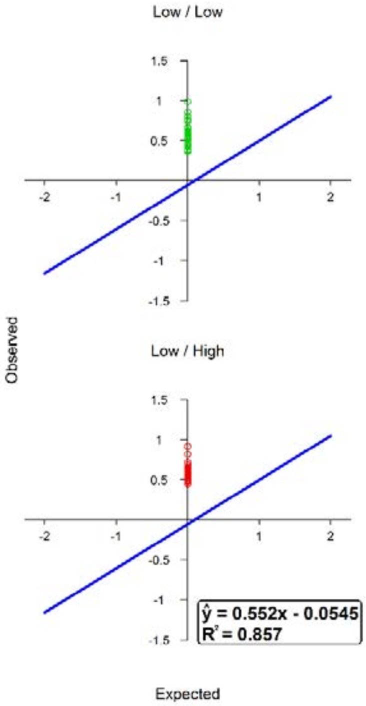

Figure 3.4 Allele-specific expression analysis. Ratios of *C*:*T* cDNA expression levels in the testis from heterozygous animals for exonic variation in *SULT2A1*. Heterozygous animals' diplotypes are indicated in the titles (low/low, low/high). Standard curve fitted to all control samples (blue line) and its regression equation and coefficient of determination are shown.

3.2.6 Origin of the haplotypes

A phylogenetic analysis was applied to investigate the origin of the haplotypes by extracting the 1.94 Mbp region from sequencing data from the 55 sequenced animals (Fig. 3.5) (Bosse *et al.* 2012).

The clustering of animals revealed by the phylogenetic tree computed using sequencing data was concordant with the tree based on the haplotypes from this region computed from the Illumina PorcineSNP60 BeadChip data. Haplotypes were grouped into three clusters: animals homozygous for haplotype 1 or haplotypes 1-like, animals homozygous for haplotype 2 or haplotypes 2-like, and animals heterozygous for haplotypes 1 and 2. Among the Asian animals only haplotype 1 or haplotypes 1-like were found, whereas in European wild boars only haplotype 2 or haplotypes 2-like were found. On the other hand, commercial European breeds are located within all three groups, showing that those animals carry all haplotypes.

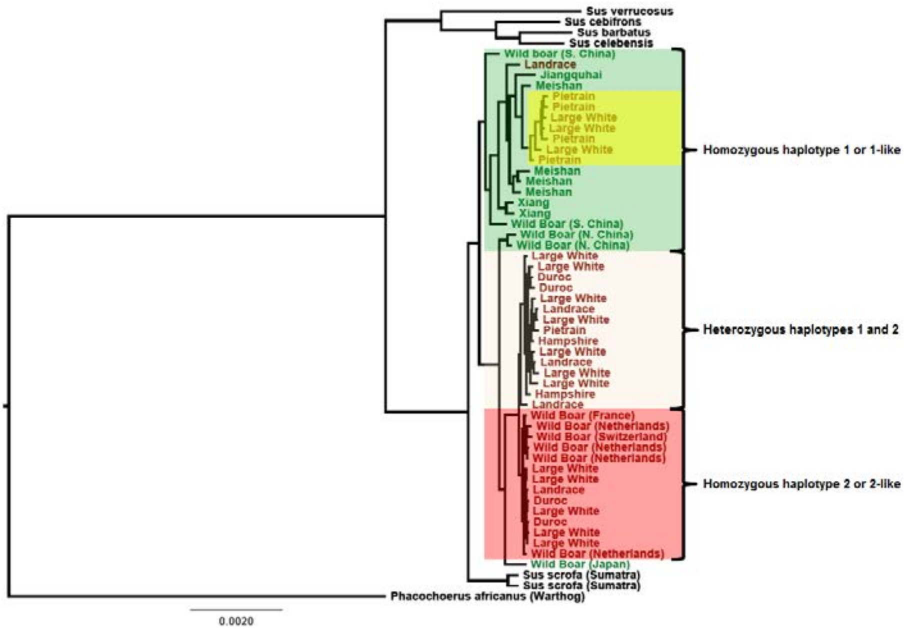

Figure 3.5 Phylogenetic tree for the haplotypes within the narrowed region in whole genome sequenced animals. Asian animals (green) are homozygous for haplotype 1 or 1-like, whereas European animals (red) have both haplotypes. European cluster of animals within the Asian animals group (yellow) shows that haplotype 1 (low androstenone) originated from Asian breeds.

3.3 Discussion

The SSC6 region associated with androstenone level was reduced from 3.75 Mbp to 1.94 Mbp and the association of haplotypes in the region with androstenone was replicated in independent populations. Haplotype 1 reduces the androstenone level across populations and can be potentially implemented in marker-assisted selection by pig breeding companies. Selection for haplotype 1 would speed up the genetic response for lower androstenone level, which would reduce the incidence of boar taint, countering the effects of international policies regarding castration of piglets. The association of *SULT2A1* expression in the testis with the level of androstenone (Moe *et al.* 2007; Grindflek *et al.* 2010; Leung *et al.* 2010) was confirmed by sequence analysis of RNA pools. Validation of differential expression showed that a SNP located within exon 2 of *SULT2A1* presented higher expression of the *C* over the *T* allele, confirming the result from the RNA-seq analysis and suggesting allelic-imbalanced expression of the two alleles. This difference in the ratio of *C*:*T* is however not associated with the haplotypes. A thorough search for functional SNP variation was carried out and resulted in a limited number of non-synonymous variants, despite the very high density of genes in the region.

3.3.1 Region of interest and haplotypes

The number of SNP within the region of interest is higher in our study compared to Duijvesteijn *et al.* (2010), due to the improved assembly of the reference genome Sscrofa10.2. Non-associated SNP that were previously located within the associated region were moved elsewhere on the genome; simultaneously, additional non-associated SNP were now included in the associated region.

Predicted haplotypes varied in number and frequency among the 11 populations. A larger number of haplotypes were found in those populations that represent crosses of purebred populations (populations 7 to 11). This was expected as crosses are made between divergent purebred populations that have different frequencies of haplotypes. Crosses will therefore combine haplotypes present in the purebred populations.

3.3.2 Effects and association analysis of haplotypes

Across all populations, haplotype 1 was consistently related with lower levels, and haplotype 2 with higher levels of androstenone. The haplotype tree showed two very distinct groups of haplotypes. When this tree was used to detect associations between (groups of) haplotypes and phenotypes, the estimated effects from the regression analyses were in good agreement with the evolutionary history of the haplotypes. Haplotypes similar in sequence to haplotype 1 also have similar effects, decreasing androstenone, and haplotypes similar to haplotype 2 have effects that increase androstenone.

After confirming that in general, haplotypes similar to 1 are associated with low and haplotypes similar to 2 are associated with high androstenone level, a posterior analysis using these two haplotypes together with the recombinant haplotype 7, placed haplotype 7 in the high-androstenone group. This placement was important because the haplotype 7 sequence is a recombination between haplotypes 1 and 2. From this result it was possible to deduce that the region from SNP 1 to 13 harbors the genetic variation responsible for the QTL for androstenone level in boars. Because it is unknown where the recombination took place the region was defined including the flanking intervals, 3’ up to SNP 14, and 5' up to the next SNP outside the LD block (SSC6: 48,317,509 bp – 50,259,057 bp, between genes *SAE1* and

SLC17A7). The assignment of haplotype 7 allowed us to narrow down the associated region from 3.75 Mbp to 1.94 Mbp.

This region is very gene-dense and contains several candidate genes for androstenone-level QTL (Duijvesteijn *et al.* 2010; Grindflek *et al.* 2010): *SULT2A1*, *SULT2B1*, *HSD17B14*, *LHB*, and *FTL*. The region is only ~0.3 cM long and has a low recombination rate (Tortereau *et al.* 2012) (Table S3.6). This is consistent with the low number of haplotypes identified within this region, even when using multiple populations. Across all 11 populations the same small set of haplotypes was found with consistently replicated effects of the haplotypes on androstenone, making the results very robust and useful for breeding programs selecting animals with reduced androstenone level.

3.3.3 RNA-seq analysis

From the six genes that were differentially expressed in the liver and testis, *SULT2A1* is an obvious candidate gene as it is involved in the metabolism of steroids. This gene is a sulfotransferase enzyme which sulfoconjugates α-androstenone. Increased expression of *SULT2A1* in the testis was found in the pool of animals with high-androstenone haplotype 2 (Table 3.1).

The higher level of *SULT2A1* in the testis was associated with higher androstenone level in fat tissue. This was unexpected based on the predictions by Sinclair and Squires (2005) that animals with low ability to sulfoconjugate 5α-androstenone in the testis would have higher accumulation of this hormone in fat tissue. Nevertheless, three other studies on different breeds (Duroc, Norwegian Landrace, and Yorkshire) (Moe *et al.* 2007; Grindflek *et al.* 2010; Leung *et al.* 2010) are in accordance with our results, showing up-regulation of *SULT2A1* in the testis of high-androstenone animals. Androstenone is known to be sulfoconjugated in the testis (Sinclair and Squires 2005), presumably to facilitate excretion and subsequent transport as androstenonesulfate in the blood. As suggested by Moe *et al.* (2007), high androstenone levels might induce an increase in *SULT2A1* expression in the testis. Recent results suggest, however, that *SULT2A1* might not be involved in the sulfoconjugation of androstenone and that another sulfotransferase is involved in this step, or that it is involved only in combination with enolase (Desnoyer 2011). Moe *et al.* (2008) also studied gene expression in the liver and found many genes to be differentially expressed but not *SULT2A1*, similar to our observation for the liver. Another candidate gene that was differentially expressed in the testis is *FTL*. The *FTL* gene codes for the ferritin light chain, an iron storage protein involved in numerous essential cellular functions. Although the function of *FTL* in the synthesis of androstenone has not been investigated (Leung *et al.* 2010), it was suggested by Moe *et al.* (2007) that *FTL* may influence androstenone level by interaction with *CYB5A* that may affect the *CYB5/CYP450* electron transfer. As the role of *FTL* affecting androstenone has not been investigated in more detail and in our study we did not find any variants that could explain a difference in expression, it remains unclear whether it has a direct effect on androstenone level. It was, therefore, not considered to be a strong candidate gene. Our expression data for *FTL* is consistent with the findings of three other studies (Moe *et al.* 2007; Grindflek *et al.* 2010; Leung *et al.* 2010), where *FTL* was up-regulated in Duroc, Norwegian Landrace, and Yorkshire boars with high androstenone levels.

3.3.4 Functional analysis using DNA sequence data

3.3.4.1 Coding regions

The only gene within the 1.94 Mbp region for which a non-synonymous variation was identified between haplotypes 1 and 2 that might have an impact on protein function was *FUT1*. *FUT1* has been identified as a candidate gene controlling the adhesion of enterotoxigenic *Escherichia coli* (ETEC) F18 to the F18 receptor (Bao *et al.* 2011). However, *FUT1* is not known to have an influence on androstenone level, and based on the functions of the protein encoded by this gene, it is unlikely that it affects androstenone level.

3.3.4.2 Regulatory regions

We studied the regulatory regions of *SULT2A1* and *FTL* because they were the two candidate genes that were differentially expressed according to the RNA-seq analysis. We checked potential TFBS and CpG islands, and only one variation (C/G, 49,110,873 bp) was found within a CpG island (49,110,687 bp - 49,110,889 bp) predicted for *SULT2A1*.

CpG islands are known to play a role in regulating gene expression where, in general, higher methylation levels are related to repression of gene expression (Du *et al.* 2012). This one variation found within the CpG island could explain the difference in expression of *SULT2A1* caused by the haplotype, however, this difference in expression between haplotypes identified by RNA-seq could not be validated subsequently, making it very unlikely that this variation plays a role in gene regulation.

3.3.5 Validation of differential *SULT2A1* expression

Allele-specific expression analysis was a follow-up step to the RNA-seq experiment to test the association of the haplotypes with difference in the ratio of *C*:*T* expression of *SULT2A1* within heterozygous animals. The quantitative difference in the relative expression found for RNA-seq (2.5:1) and allele-specific expression analysis (1.5:1) may simply be due to random error in the estimate from RNA-seq analysis which was based on only two pooled samples. Other reasons include systematic or technical differences that affect the amplification in the RNA-seq assay. There may be other biological mechanisms that trigger a higher expression of *SULT2A1* allele *C* that cannot be captured by allele-specific expression analysis. Unraveling such a mechanism can however not be achieved using our data. Surprisingly, in the allele-specific expression analysis we did not observe differential expression between heterozygous animals (C/T) with either low/low or high/low androstenone diplotypes (Fig. 3.4). We concluded that the difference in *SULT2A1* expression was not regulated by the haplotypes surrounding the *SULT2A1* gene. Instead, an increase in expression of allele *C* over allele *T* in *SULT2A1* was observed, indicating haplotype independent allelic-imbalanced expression between these two alleles. One option for the cause of this allelic-imbalanced expression is a potential regulatory SNP-variant in LD with the *SULT2A1* SNP that affects expression. Other options are transcriptional regulation of the two alleles, like in an enhancer element, that resides outside the investigated region, or differences in RNA decay between the two alleles. It is known that the RNA folding structures play a role in the degree of RNA decay. Prediction of the fold structure indicated considerable difference in structure around the two alleles (Madsen, O., unpublished observation) making RNA decay a possible participant in the observed allelic-imbalanced expression.

3.3.6 Origin of the haplotypes

Since the entire region between 48.3 Mbp and 50.2 Mbp on SSC6 has a very low recombination rate (Tortereau *et al.* 2012), the integrity of the haplotypes found in this study has been retained across different populations. Because of this retained integrity, a phylogenetic analysis could be applied to construct a phylogenetic tree of this region from sequencing data from the 55 sequenced animals (Fig. 3.5) (Bosse *et al.* 2012).

This tree revealed that haplotype 1 of the 1.94 Mbp region, associated with low androstenone level, originated from Asia. It is likely, therefore, that haplotype 1 was introgressed into European breeds during the 18th and 19th centuries, generating hybrid European breeds (Giuffra *et al.* 2000). Introgression of favorable Asian haplotypes has been observed for other traits as well. A well-known example is an *IGF2* haplotype conferring increased muscle mass and leaner pigs (Andersson and Georges 2004). This haplotype is currently in high frequency in several commercial pig populations, but originated from Asian pigs. There is currently only a handful of gene variants described from European pigs that originate from the late 18th- early 19th century introgression of Asian breeding stock (e.g. Wilkinson *et al.* 2013).

The likely relatively recent (i.e., around 200 years ago or less) introgression of the Asian haplotypes into the European pigs, combined with the very low recombination rate in the genomic region, further explains the paucity of recombinant haplotypes, and difficulty in fine-mapping even across breeds.

Pigs with Asian origin haplotypes were associated with low-androstenone level, whereas European-origin haplotypes were associated with high androstenone level. This is consistent with Lee *et al.* (2005) who found that Large White alleles have an additive effect on androstenone level for a QTL found on SSC6 at 91 cM, between SW782 (49,996,734 bp - 49,996,825 bp) and SW1823 (79,653,393 bp - 79,653,597 bp), in an F2 Large White x Meishan population.

Taking into account that haplotypes of European breeds originated from Asian breeds and that Asian breeds have high genetic diversity (Megens *et al.* 2008), further studies are needed either to identify additional haplotypes that are recombinant between European and Asian animals or to fine-map the region further in Asian pigs since LD will be much lower than in European pigs.

3.4 Conclusions

In summary, the androstenone QTL region previously identified on SSC6 (Duijvesteijn *et al.* 2010) was narrowed down from 3.75 Mbp to 1.94 Mbp. Differential expression was observed for six genes by RNA-seq analysis. No difference in the ratio of *C*:*T* expression of *SULT2A1* for the haplotypes was found by the allele-specific expression analysis, however, a difference in expression between the *C* over *T* allele was found for a variation within *SULT2A1*, showing that the difference in androstenone levels between the haplotypes is not caused by the SNP in exon 2. Nonetheless, a difference in ln-androstenone level across populations in case of fixation of the Asian-origin haplotype 1 would yield a change, on average, of -0.19 ln-androstenone (ranging from -0.57 to +0.08). Use of tag-SNP from the haplotype-1 group will be valuable in animal breeding programs to select animals with lower androstenone levels.

3.5 Methods

This study was conducted according to regulations of Dutch law on protection of animals.

3.5.1 Phenotypes, animals, and genotypes

Phenotypes for androstenone level were obtained from 2,750 boars slaughtered at a mean hot carcass weight of 91.33 kg (SD = 9.21 kg). Androstenone level was measured in fat samples; details of measurements and fat extraction are described in earlier studies (Duijvesteijn *et al.* 2010; Ampuero Kragten *et al.* 2011). Androstenone level was log-transformed (ln-androstenone) because it was not normally distributed. Boars belonged to six purebred populations (population 1 to 6, Duroc-based, Yorkshire-based, Dutch Landrace, Pietrain, Finish Landrace, and Large White) and five terminal crosses based on populations 1-6 (population 7 to 11). Number of animals per population ranged from 940 for a Duroc-based population to 69 for one of the crosses (Table 3.2).

Table 3.2 Number of animals and means (standard deviation) for ln-androstenone and androstenone (μg/g) per population

Population	Number of animals	Ln-androstenone (SD)	Androstenone μg/g (SD)
1	940	0.24 (0.89)	1.84 (1.62)
2	295	-0.31 (0.83)	1.02 (0.91)
3	208	-0.04 (0.83)	1.33 (1.21)
4	207	-1.29 (0.99)	0.45 (0.49)
5	169	0.25 (0.94)	1.88 (1.64)
6	107	0.17 (1.15)	2.14 (2.46)
7	325	-0.61 (0.87)	0.82 (0.96)
8	275	-0.12 (0.88)	1.31 (1.29)
9	83	0.03 (0.82)	1.43 (1.25)
10	72	0.34 (0.94)	2.18 (2.37)
11	69	0.19 (0.83)	1.67 (1.45)

Genotyping was performed using the Illumina PorcineSNP60 Beadchip (San Diego, CA, USA) (Ramos *et al.* 2009). Quality control involved removing SNP with low quality score (GenCall score <0.7), and those with a minor allele frequency lower than 0.01 (Duijvesteijn *et al.* 2010). A total of 3,025 SNP located on SSC6 remained and 46 were used in the analyses.

3.5.2 Linkage Disequilibrium (LD) analysis

Significant SNP (N=29) previously identified (Duijvesteijn *et al.* 2010) were rearranged according to the Sscrofa10.2 reference genome and LD blocks were defined based on the criteria of Gabriel *et al.* (2002).

3.5.3 Haplotype diversity

Haplotypes with frequencies greater than 1% across all populations were identified using Haploview 4.2 (Barrett *et al.* 2005). A phylogenetic tree for haplotypes based on the similarities among the 46 SNP from the haplotypes was constructed using the neighbor-joining method as implemented in MEGA 5 (Tamura *et al.* 2011).

3.5.4 Association analysis

Phasing and imputation of sporadic missing data were performed using BEAGLE (Browning and Browning 2007).

Linear regression was used to estimate effect of haplotypes for each population using ASReml v3.0 (Gilmour *et al.* 2009). The following model was used

$$y_i = b_1 haplo + a_i + e_i$$

where y_i is the ln-androstenone of the ith animal, b_1 is the regression coefficient on the haplotype, a_i is the random additive genetic effect of the ith animal, e_i is the random residual effect.

For the posterior analysis using three haplotypes, only populations that had the third haplotype that was being compared to haplotypes 1 and 2 were included in the analysis. In this analysis, the model was corrected for population.

To test whether groups of haplotypes that are on different branches of the phylogenetic tree have a statistically significant different effect on the androstenone level, we used Treescan which "cuts" the haplotype tree at different branch points to identify functional grouping of haplotypes.

3.5.5 DNA sequence data

The sequencing procedure used for the 55 sequenced individuals in the present study was described in Bosse *et al.* (2012). Briefly, Illumina-formatted (v. 1.3-1.7) fastq files, with sequence reads of 100 bp (Illumina HiSeq2000), were subject to quality trimming prior to sequence alignment. A minimum average quality score of

13 (i.e. average error probability equal to 0.05) in a 3 bp window was used as cut-off, with 3-prime sequences being discarded if the criterion was not met. Only sequences where both mates were at least 45 bp in length were retained.

Sequences were aligned against the Sscrofa10.2 reference genome using Mosaik align v.1.1.0017 (http://bioinformatics.bc.edu/marthlab/Mosaik). Alignment was performed using a hash-size of 15, with a maximum of 10 matches retained, and 7% maximum mismatch score, for all pig populations and outgroup species. Alignment files were then sorted using the “mosaiksort” function, which entails removing ambiguously mapped reads that are either orphaned or fall outside a computed insert-size distribution.

Variant allele-calling was performed per individual using the “pileup” function in SAMtools v.1.12a (Li *et al.* 2009), and variations were initially filtered to have minimum quality of 50 for indels, and 20 for SNP. In addition, all variants showing higher than 3x average read-density, estimated from the number of raw sequence reads, were also discarded to remove false-positive variant-calling originating from off-site mapping as much as possible. This procedure yielded high-quality variants for 55 pigs, wild boars and outgroup species (European Nucleotide Archive (ENA) under project number ERP001813).

3.5.6 Phylogenetic analysis

Sequence assemblies for the region on SSC6 between Sscrofa10.2 reference genome base 48,317,509 and 50,259,057 were extracted per individual according to their genomic coordinates from the BAM files generated from individually-sequenced pigs using SAMtools v.1.12a. Phylogenetic analysis was done using RAxML (Stamatakis *et al.* 2005), using 10 iterations and implementing a GTR-Γ model of sequence evolution, and with an African warthog as outgroup.

3.5.7 Functional analysis using DNA sequence data

The two haplotypes present across all populations were numbered haplotype 1 and 2 and re-sequencing data were used to identify putative functional differences between them.

To obtain genotype calls for all polymorphic sites identified across the 55 individuals for which whole-genome sequence data were available, every individual was examined for the genotype call for each of the sites found to be polymorphic in the region of interest, including the species-specific differences. Filtering was based on sequence depth (genotype retained if depth ranged between four reads, and twice the average genome-wide depth), where for this procedure the average sequence depth was based directly on the actual sequence depth measured for each individual separately. Further filtering of these sequence-derived genotypes was performed on SNP and consensus quality (for homozygotes, either a SNP or consensus quality > 20 was applied, and for heterozygotes, both SNP and consensus qualities > 20 were applied).

With the SNP that contributed different alleles to haplotype 1 and 2, we performed an analysis to annotate functional genetic variants detected between the haplotypes using ANNOVAR. For SNP that resulted in an amino acid substitution we used PolyPhen2 to predict the impact of this substitution on the structure and function of the protein. To identify mutations outside the exonic regions that have the potential to influence androstenone level, we used MULAN (Ovcharenko *et al.* 2005) to detect potential TFBS that were conserved across three species (*Sus scrofa*, *Bos taurus*, and *Homo sapiens*). SNP between haplotypes 1 and 2 were checked for being located within a TFBS. CpG islands were predicted using EMBOSS/CpGPlot with default settings to identify SNP between haplotypes 1 and 2 that could be located within a CpG island.

3.5.8 RNA sequencing data and gene expression analysis

Forty-eight animals of the Duroc-based population were slaughtered at a mean age of 173 days, and liver and testis tissue were collected and stored in RNALater (Qiagen Inc., Valencia, CA, USA). These samples were genotyped for the 29 SNP that were significant in Duijvesteijn *et al.* (2010). The liver and testis were collected for this analysis because androstenone is synthesized in the testis and metabolized in the liver, leading us believe that they are the most interesting tissues to use for the analysis of the effect of gene expression on androstenone levels. Four animals homozygous for the haplotype associated with high androstenone level were selected for RNA isolation in both the liver and testis, as well as four animals homozygous for the haplotype associated with a low androstenone level. From these samples, total RNA was extracted with the RNeasy mini kit (Qiagen Inc., Valencia, CA, USA) following manufacturer's instructions. RNA from low- and high-androstenone haplotypes were pooled, respectively, and stored at -80 ºC until being used. The Illumina mRNA-seq Sample Preparation Kit was used for sample preparation (~5 µg of total RNA) following manufacturer's instructions and used for 100 bp single-end cDNA sequencing on the Illumina HighSeq 2000 platform.

The sequence data obtained from the two RNA pools were clipped to remove adapter sequence and quality trimmed (Phred score > 20) with Trim galore v.0.2.2 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore). After cleaning the data, between 58 – 62 million sequence reads were available from each pool. To compare gene and transcript expression, we followed the protocol described in Trapnell *et al.* (2012). Reads were aligned against the Sscrofa10.2 reference genome with TopHat v.1.4.1 (Trapnell *et al.* 2011) using the -T option (all other options were default) in order to align the reads only against an annotated transcriptome. Estimation of differences in expression was done with Cufflinks v.1.3.0 (Trapnell *et al.* 2011) and visualized with the CummeRbund package (Goff and Trapnell 2012). Many imperfections exist in the current Sscrofa10.2 reference genome that may affect the details of the gene model, however we do not think that these imperfections compromise our conclusions. The position of *SULT2A1* is not in doubt because the BAC-by-BAC sequencing strategy, based on a rather good physical map, has been shown to result in overall well-assembled genome sequence at a larger scale (e.g. Tortereau *et al.* 2012). The most important inconsistencies are within BAC, because BAC were shotgun sequenced (using classical Sanger sequencing strategies) at low average depth of ~4-6x (Groenen *et al.* 2012).

Allele-specific expression analysis was performed for the exonic mutation within *SULT2A1* (*C/T* change within exon 2 at position 166; SSC6: 49,117,861 bp in Sscfrofa10.2), on testis tissue from 67 heterozygous animals. We prepared a Taqman PCR Reaction using the assaymix 40x for *SULT2A1* (AHN1RKQ, Applied Biosystems). Taqman PCR was performed on ABI 7500 RT-PCR system. Output values of cycle 40 (exponential phase) were used for both *C* and *T* signals. These values were corrected for background noise by subtracting the value for the respective signal of cycle 1. To certify that the allele-specific expression analysis was accurate, we quantified the two alleles of the SNP in genomic DNA mixes with known ratios: 4:1, 2.33:1, 1.5:1, 1:1, 1:1.5, 1:2.33, 1:4, and in three concentrations of genomic DNA: 2.5 ng, 10 ng, 40 ng. Ratios were a mix of genomic DNA from two homozygous animals for different alleles. For 10 ng and 40 ng dilutions, 13 heterozygous animals were included in the 1:1 class (Sun *et al.* 2010).

3.6 Acknowledgements

We thank Bert Dibbits, Animal Breeding and Genomics Centre - Wageningen University, for performing allele-specific expression analysis. The first author benefited from a joint grant from the European Commission and Topigs Norsvin within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG". Topigs Norsvin had a role in study design, data collection, and preparation of the manuscript.

References

Adzhubei IA, Schmidt S, Peshkin L, et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249.

Ampuero Kragten S, Verkuylen B, Dahlmans H, et al (2011) Inter-laboratory comparison of methods to measure androstenone in pork fat. Animal 5:1634–1642.

Andersson L, Georges M (2004) Domestic-animal genomics: deciphering the genetics of complex traits. Nat Rev Genet 5:202–212.

Bao W Bin, Ye L, Pan ZY, et al (2011) Beneficial genotype of swine FUT1 gene governing resistance to E. coli F18 is associated with important economic traits. J Genet 90:315–318.

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–5.

Bonneau M (1982) Compounds responsible for boar taint, with special emphasis on androstenone: a review. Livest Prod Sci 9:687–705.

Bonneau M, Chevillon P (2012) Acceptability of entire male pork with various levels of androstenone and skatole by consumers according to their sensitivity to androstenone. Meat Sci 90:330–337.

Bosse M, Megens HJ, Madsen O, et al (2012) Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. Plos Genet 8:e1003100.

Browning SR, Browning BL (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 81:1084–97.

Desnoyer JEVE (2011) The formation of androstenone conjugates in testis tissue of the mature boar. MsC thesis. Dept of Animal and Poultry Science, University of Guelph. http://hdl.handle.net/10214/3165.

Dorries KM, Adkins-Regan E, Halpem BP (1997) Sensitivity and behavioral responses to the perfomone androstenone are not mediated by the vomeronasal organ in domestic pigs. Brain Behav Evol 49:53–62.

Du X, Han L, Guo AY, Zhao Z (2012) Features of methylation and gene expression in the promoter-associated CpG islands using human methylome data. Int J Genomics 2012:1-8.

Dufort I, Soucy P, Lacoste L, Luu-The V (2001) Comparative biosynthetic pathway of androstenol and androgens. J Steroid Biochem Mol Biol 77:223–227.

Duijvesteijn N, Knol EF, Merks JWM, et al (2010) A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC Genet 11:42.

Gabriel SB, Schaffner SF, Nguyen H, et al (2002) The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.

Giuffra E, Kijas JM, Amarger V, et al (2000) The origin of the domestic pig: independent domestication and subsequent introgression. Genetics 154:1785–91.

Goff LA, Trapnell C, Kelley D (2012) CummeRbund : analysis, exploration, manipulation and visualization of Cufflinks high-throughput sequencing data. R package version 2.2.0.

Gregersen VR, Conley LN, Sørensen KK, et al (2012) Genome-wide association scan and phased haplotype construction for quantitative trait loci affecting boar taint in three pig breeds. BMC Genomics 13:22.

Grindflek E, Berget I, Moe M, et al (2010) Transcript profiling of candidate genes in testis of pigs exhibiting large differences in androstenone levels. BMC Genet 11:4.

Grindflek E, Lien S, Hamland H, et al (2011a) Large scale genome-wide association and LDLA mapping study identifies QTLs for boar taint and related sex steroids. BMC Genomics 12:362.

Grindflek E, Meuwissen THE, Aasmundstad T, et al (2011b) Revealing genetic relationships between compounds affecting boar taint and reproduction in pigs. J Anim Sci 89:680–92.

Groenen MAM, Archibald AL, Uenishi H, et al (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393–8.

Lee GJ, Archibald a L, Law a S, et al (2005) Detection of quantitative trait loci for androstenone, skatole and boar taint in a cross between Large White and Meishan pigs. Anim Genet 36:14–22.

Leung MCK, Bowley K-L, Squires EJ (2010) Examination of testicular gene expression patterns in Yorkshire pigs with high and low levels of boar taint. Anim Biotechnol 21:77–87.

Li H, Handsaker B, Wysoker A, et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079.

Megens HJ, Crooijmans RP, Cristobal MS, et al (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.

Moe M, Lien S, Bendixen C, et al (2008) Gene expression profiles in liver of pigs with extreme high and low levels of androstenone. BMC Vet Res 4:29.

Moe M, Meuwissen T, Lien S, et al (2007) Gene expression profiles in testis of pigs with extreme high and low levels of androstenone. BMC Genomics 8:405.

Ovcharenko I, Loots GG, Giardine BM, et al (2005) Mulan: multiple-sequence local alignment and visualization for studying function and evolution. Genome Res 15:184–194.

Paudel Y, Madsen O, Megens HJ, et al (2013) Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication. BMC Genomics 14:449.

Posada D, Maxwell TJ, Templeton AR (2005) TreeScan: A bioinformatic application to search for genotype/phenotype associations using haplotype trees. Bioinformatics 21:2130–2132.

Ramos AM, Crooijmans RPMA, Affara NA, et al (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.

Seideman SC, Cross HR, Oltjen RR, Schanbacher BD (1982) Utilization of the intact male for red meat production : a review. J Anim Sci 55:826–840.

Sinclair PA, Gilmore WJ, Lin Z, et al (2006) Molecular cloning and regulation of porcine *SULT2A1*: Relationship between *SULT2A1* expression and sulfoconjugation of androstenone. J Mol Endocrinol 36:301–311.

Sinclair PA, Hancock S, Gilmore WJ, Squires EJ (2005) Metabolism of the 16-androstene steroids in primary cultured porcine hepatocytes. J Steroid Biochem Mol Biol 96:79–87.

Sinclair PA, Squires EJ (2005) Testicular sulfoconjugation of the 16-androstene steroids by hydroxysteroid sulfotransferase: its effect on the concentrations of 5α-androstenone in plasma and fat of the mature domestic boar. J Anim Sci 83:358–365.

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.

Sun C, Southard C, Witonsky DB, et al (2010) Allelic Imbalance (AI) identifies novel tissue-specific cis-regulatory variation for human *UGT2B15*. Hum Mutat 31:99–107.

Tamura K, Peterson D, Peterson N, et al (2011) MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739.

Tortereau F, Servin B, Frantz L, et al (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC Genomics 13:586.

Trapnell C, Roberts A, Goff L, et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–78.

Trapnell C, Williams B A, Pertea G, et al (2011) Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat Biotechnol 28:511–515.

Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38:1–7.

Wilkinson S, Lu ZH, Megens HJ, et al (2013) Signatures of diversifying selection in European pg breeds. PLoS Genet 9:e1003453.

Supplementary material

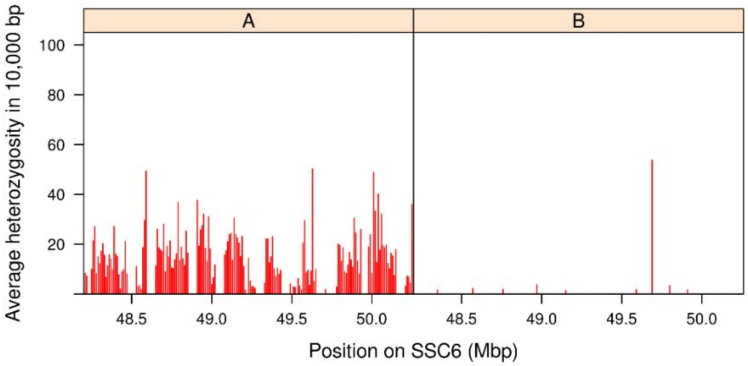

Figure S3.1 Average heterozygosity for the narrowed region. Average heterozygosity in 10,000 bp bins along the narrowed region for a heterozygous animal (A) and a homozygous animal (B).

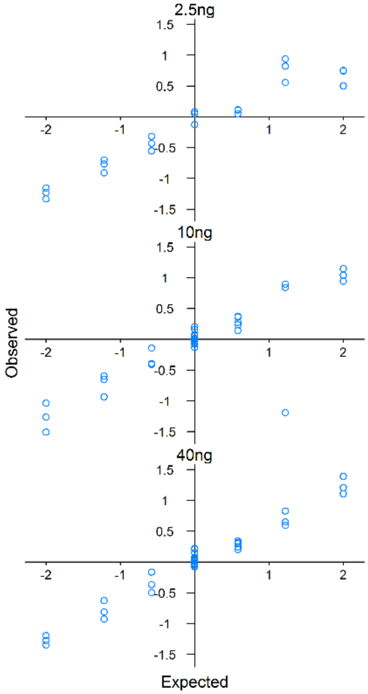

Figure S3.2. Panels with control allele-specific expression. Log-transformed (base 2) data relation between observed and expected ratios of mixed genomic DNA from two homozygous animals for different alleles in three concentrations: 2.5 ng, 10 ng, and 40 ng.

Table S3.1 Position of the 46 SNP in the QTL region on SSC6 of Sscrofa10.2.

SNP name	Position (bp)	SNP name	Position (bp)
H3GA0053864	48585961	ASGA0028216	50742441
ALGA0102689	48717238	ASGA0028211	50803585
ASGA0104037	48792292	M1GA0008539	50847065
ASGA0089838	49146524	ASGA0103898	50867656
ASGA0093393	49168322	MARC0098482	50888554
MARC0019764	49351202	H3GA0056609	50922233
MARC0015928	49538608	ALGA0122867	51104922
DIAS0000492	49802217	ALGA0116613	51139647
MARC0011519	49817264	ASGA0103416	51352837
H3GA0056470	50006716	ALGA0035323	51611976
DIAS0004447	50037571	ALGA0035324	51636474
DIAS0003231	50065951	ALGA0035318	51678926
ASGA0084861	50079246	MARC0021351	51692785
MARC0032442	50259057	ASGA0028223	51757391
DIAS0000822	50264414	ALGA0035326	51775907
H3GA0053555	50307537	ASGA0028228	51805308
DIAS0003830	50339827	ALGA0035330	51843873
MARC0049189	50364492	MARC0086794	52063034
MARC0044346	50478565	ALGA0115158	52085979
M1GA0008536	50495796	ASGA0097167	52127558
H3GA0017949	50532885	ALGA0112704	52226606
ASGA0028206	50556192	MARC0005462	52262806
M1GA0008527	50606084	MARC0049139	52336598

Table S3.2 Haplotype effects and total number of animals per pig population

Hap*	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5	Pop. 6	Pop. 7	Pop. 8	Pop. 9	Pop. 10	Pop. 11
1	-0.19	-0.07	0.06	-0.15	-0.03	-0.31	-0.15	-0.05	-0.01	-0.31	-0.18
2	0.10	0.04	0.18	-0.09	-0.06	0.19	-0.05	0.01	0.04	0.14	0.24
3	-0.08	NA	NA	0.34	NA	NA	NA	-0.18	NA	-0.30	0.40
4	0.10	NA	NA	NA	0.02	0.05	NA	0.23	-0.07	0.13	0.55
5	NA	0.23	0.04	-0.14	0.10	NA	-0.18	0.25	-0.27	NA	NA
6	NA	NA	0.09	NA	0.07	NA	0.02	0.17	0.30	0.62	NA
7	NA	-0.04	NA	NA	0.05	0.00	0.28	-0.11	NA	NA	-0.39
8	NA	0.08	NA	0.05	NA	NA	NA	NA	-0.06	NA	-0.21
9	0.15	NA	NA	NA	NA	0.36	0.47	0.08	-0.02	0.31	NA
10	NA	-0.15	-0.03	0.12	NA	NA	-0.41	-0.37	-0.53	NA	0.31
N	940	295	208	207	169	107	325	275	83	72	69

* Haplotype

Table S3.3 Haplotype effects estimated using only populations that have the third haplotype that is being compared to haplotypes 1 and 2

Haplotype	Estimate	Std. Error
	Haplotype 3	
Haplotype1	**-0.22**	**0.04**
Haplotype2	-0.01	0.04
Haplotype3	**-0.15**	**0.05**
	Haplotype 4	
Haplotype1	-0.14	0.04
Haplotype2	**0.12**	**0.04**
Haplotype4	**0.16**	**0.05**
	Haplotype 5	
Haplotype1	-0.15	0.04
Haplotype2	**-0.03**	**0.04**
Haplotype5	**-0.01**	**0.05**
	Haplotype 6	
Haplotype1	-0.11	0.05
Haplotype2	**-0.01**	**0.05**
Haplotype6	**0.10**	**0.06**
	Haplotype 7	
Haplotype1	-0.16	0.04
Haplotype2	**0.00**	**0.05**
Haplotype7	**0.05**	**0.07**
	Haplotype 8	
Haplotype1	-0.11	0.06
Haplotype2	0.03	0.07
Haplotype8	0.07	0.09
	Haplotype 9	
Haplotype1	-0.15	0.04
Haplotype2	**0.07**	**0.04**
Haplotype9	**0.26**	**0.08**
	Haplotype 10	
Haplotype1	-0.15	0.04
Haplotype2	-0.05	0.04
Haplotype10	-0.13	0.08

Table S3.4 List of genes and the status of the exonic variation

Gene	Exon	Status	Gene	Exon	Status
GPR77	Exon1	Synonymous	*MAMSTR*	Exon2	Synonymous
GPR77	Exon1	Synonymous	*IZUMO1*	Exon1	Non-synonymous
GPR77	Exon1	Synonymous	*FUT1*	Exon1	Non-synonymous
DHX34	Exon2	Synonymous	*HSD17B14*	Exon9	Synonymous
DHX34	Exon9	Synonymous	*HSD17B14*	Exon4	Non-synonymous
DHX34	Exon15	Synonymous	*PLEKHA4*	Exon17	Synonymous
DHX34	Exon16	Synonymous	*PLEKHA4*	Exon6	Synonymous
ZNF541	Exon1	Synonymous	*PPP1R15A*	Exon1	Synonymous
ZNF541	Exon3	Synonymous	*PPP1R15A*	Exon2	Non-synonymous
ZNF541	Exon3	Synonymous	*PPP1R15A*	Exon2	Synonymous
ZNF541	Exon5	Synonymous	*HRC*	Exon1	Non-synonymous
ZNF541	Exon15	Non-synonymous	*HRC*	Exon1	Non-synonymous
KPTN	Exon10	Synonymous	*PPFIA3*	Exon26	Synonymous
SLC8A2	Exon1	Synonymous	*PPFIA3*	Exon25	Non-synonymous
SLC8A2	Exon1	Synonymous	*PPFIA3*	Exon21	Synonymous
MEIS3	Exon1	Synonymous	*PPFIA3*	Exon15	Non-synonymous
MEIS3	Exon10	Non-synonymous	*PPFIA3*	Exon14	Non-synonymous
GLTSCR1	Exon9	Synonymous	*PPFIA3*	Exon13	Non-synonymous
GLTSCR1	Exon12	Synonymous	*PPFIA3*	Exon13	Non-synonymous
GLTSCR1	Exon13	Synonymous	*PPFIA3*	Exon3	Synonymous
EHD2	Exon2	Synonymous	*LIN7B*	Exon3	Synonymous
GLTSCR2	Exon1	Synonymous	*SNRNP70*	Exon5	Synonymous
GLTSCR2	Exon5	Synonymous	*SNRNP70*	Exon2	Synonymous
GLTSCR2	Exon8	Non-synonymous	*KCNA7*	Exon1	Synonymous
GLTSCR2	Exon10	Synonymous	*KCNA7*	Exon2	Synonymous
CRX	Exon3	Synonymous	*KCNA7*	Exon1	Synonymous
CRX	Exon3	Synonymous	*NTF4*	Exon1	Synonymous
SULT2A1	Exon2	Synonymous	*LHB*	Exon1	Synonymous
LIG1	Exon14	Synonymous	*RUVBL2*	Exon13	Synonymous
TMEM143	Exon5	Synonymous	*GYS1*	Exon5	Synonymous
LMTK3	Exon5	Synonymous	*FTL*	Exon4	Synonymous
LMTK3	Exon11	Synonymous	*NUCB1*	Exon5	Synonymous
CYTH2	Exon4	Synonymous	*TULP2*	Exon6	Non-synonymous
GRWD1	Exon6	Synonymous	*TULP2*	Exon8	Non-synonymous
GRIN2D	Exon11	Synonymous	*TULP2*	Exon10	Non-synonymous
GRIN2D	Exon11	Synonymous	*SLC17A7*	Exon10	Synonymous
GRIN2D	Exon4	Synonymous	*SLC17A7*	Exon7	Synonymous
KDELR1	Exon4	Synonymous			

Table S3.5 The QTL effects for the *C* and *T* allele of the high- and low-androstenone haplotypes

Allele (haplotype)	Estimate	Std. Error
C allele (high-androstenone haplotype)	0.00	0.00
***T* allele (low-androstenone haplotype)**	**-0.44**	**0.11**
***C* allele (low-androstenone haplotype)**	**-0.29**	**0.11**

Table S3.6 Genetic and physical map of the narrowed region (SSC6: 48,317,509 bp – 50,259,057 bp), showing its low-recombining nature

SNP	Order	cM	Position	SNP	Order	cM	Position
H3GA0053864	623	62.263	48585961	ASGA0028211	641	62.441	50803585
ASGA0104037	624	62.276	48792292	M1GA0008539	642	62.447	50847065
ALGA0102689	625	62.290	48717238	ASGA0103898	643	62.453	50867656
ASGA0089838	626	62.303	49146524	H3GA0056609	644	62.460	50922233
MARC0015928	627	62.317	49538608	ALGA0122867	645	62.466	51104922
MARC0011519	628	62.330	49817264	ALGA0116613	646	62.476	51139647
DIAS0000492	629	62.344	49802217	ASGA0103416	647	62.485	51352837
DIAS0004447	630	62.357	50037571	ALGA0035318	648	62.495	51678926
ASGA0084861	631	62.371	50079246	MARC0021351	649	62.504	51692785
DIAS0000822	632	62.378	50264414	ASGA0028223	650	62.511	51757391
MARC0032442	633	62.386	50259057	ASGA0028228	651	62.517	51805308
MARC0049189	634	62.393	50364492	ALGA0035330	652	62.524	51843873
H3GA0053555	635	62.401	50307537	MARC0086794	653	62.530	52063034
MARC0044346	636	62.408	50478565	ALGA0115158	654	62.537	52085979
M1GA0008536	637	62.415	50495796	ASGA0097167	655	62.543	52127558
H3GA0017949	638	62.421	50532885	MARC0049139	656	62.549	52336598
ASGA0028206	639	62.428	50556192	MARC0005462	657	62.554	52262806
M1GA0008527	640	62.434	50606084	ALGA0112704	658	62.559	52226606

4

Asian low-androstenone haplotype on pig chromosome 6 does not unfavorably affect production and reproduction traits

A.M. Hidalgo[1,2], J.W.M. Bastiaansen[1], B. Harlizius[3], E.F. Knol[3], M.S. Lopes[1,3], D.J. de Koning[2], M.A.M. Groenen[1]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6708WD, the Netherlands; [2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, 756 51, Sweden; [3] Topigs Norsvin, Beuningen 6640AA, the Netherlands

Animal Genetics (2014) 45(6):874-877

Abstract

European pigs that carry Asian haplotypes of a 1.94 Mbp region on pig chromosome 6 have lower levels of androstenone, one of the two main compounds causing boar taint. The objective of our study was to examine potential pleiotropic effects of the Asian low-androstenone haplotypes. A single-nucleotide polymorphism marker, rs81308021, distinguishes the Asian from European haplotypes and was used to investigate possible associations of androstenone with production and reproduction traits. Eight traits were available from three European commercial breeds. For the two sow lines studied, a favorable effect on number of teats was detected for the low-androstenone haplotype. In one of these sow lines, a favorable effect on number of spermatozoa per ejaculation was detected for the low-androstenone haplotype. No unfavorable pleiotropic effects were found, which suggests that selection for low-androstenone haplotypes within the 1.94 Mbp would not unfavorably affect the other eight relevant traits.

Key words: boar taint, genetic correlations, number of spermatozoa, number of teats, pleiotropic effects

Androstenone is one of the two main compounds causing boar taint, a pronounced urine-like odor produced when cooking meat from non-castrated male pigs (Bonneau 1982). A 1.94 Mbp (48,317,509 bp – 50,259,057 bp) region on pig chromosome 6 (SSC6) was found to affect androstenone levels. Two major haplotype groups (low and high androstenone), spanning this entire region, can be distinguished (Hidalgo *et al.* 2014). Other genome scans also reported this quantitative trait loci (QTL) region on SSC6 (Lee *et al.* 2005; Grindflek *et al.* 2011). Analysis of sequence data showed that the low-androstenone haplotypes originated from Asian breeds; these haplotypes are found at intermediate frequencies in several European commercial breeds. Asian-origin haplotypes were introgressed into European breeds during the 18th and 19th centuries when Asian animals were used to improve European breeds, generating hybrid European breeds (Giuffra *et al.* 2000). Because androstenone levels were not taken into account in typical breeding programs in the past, we hypothesize that low-androstenone haplotypes accumulated indirectly by selection for another correlated trait.

Androstenone is chiefly synthesized in the testis and is a product of a metabolic pathway that also produces other sex hormones such as estrogens and testosterone, known to be important factors affecting fertility. Studies have found unfavorable correlations between the androstenone level in fat and reproduction traits (Willeke *et al.* 1987; Sellier & Bonneau 1988; Tajet *et al.* 2006; Mathur *et al.* 2013), which would complicate selection for a reduced androstenone level in a breeding program. This finding was confirmed by Grindflek *et al*. (2011) who studied the genetic relationship of androstenone in fat with levels of sex hormones. They detected many QTL that affected both androstenone and other sex hormones and concluded that “Most of the QTLs for androstenone are affecting both androstenone and estrogens, making practical implementation in breeding challenging”. A QTL they discovered on SSC6 in a Duroc population overlaps with our QTL region. The QTL from their study did not affect other sex hormones. Such QTL for androstenone that are “not affecting any of the other sex hormones” were suggested as being “very interesting for selection purposes” by Grindflek *et al.* (2011). Production and reproduction traits were however not reported by Grindflek *et al*. (2011). Hence, the objective of our study was to examine potential pleiotropic effects on important pig production and reproduction traits from the Asian low-androstenone haplotypes on SSC6.

Phenotypes and genotypes were available from animals of three commercial pig lines: 1) Dutch Landrace; 2) Large White; 3) Pietrain (Table 4.1). The Dutch Landrace and the Large White are dam lines and the Pietrain is a sire line. The three lines studied were also included in the study of Hidalgo *et al*. (2014) where the Dutch Landrace corresponds to population 3, Large White corresponds to population 6, and Pietrain corresponds to population 4. Phenotypes were available for eight traits: growth rate (g/day), backfat thickness (mm), litter birth weight (kg), total number of piglets born, birth weight (kg), number of teats, sperm motility (% of motile cells) and number of spermatozoa per ejaculation (billions). Genotyping was performed using the Illumina PorcineSNP60 BeadChip (Ramos *et al.* 2009). Based on the haplotypes of our previous study Hidalgo *et al*. (2014), we selected single nucleotide polymorphism (SNP) marker rs81308021 to test the trait associations. This marker is a tag SNP that distinguishes between the low- and high-androstenone haplotype groups.

Genetic correlations between androstenone and the production traits, and androstenone and the reproduction traits were estimated in a bivariate analysis using ASReml v3.0 (Gilmour *et al.* 2009). The association of the polymorphism, representing the haplotypes, with the phenotype was fitted in a linear mixed model using ASReml v3.0 (Gilmour *et al.* 2009) as shown below:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{SNP} + \mathbf{e}$$

where **y** is a vector of individual trait observations; **X** is an incidence matrix for fixed effects contained in vector **b**; **Z** is an incidence matrix connecting genetic values contained in vector **a** to phenotypes in **y**; **e** is a vector of random errors associated with *y*. **SNP** was coded as 0, 1 or 2, for the number of copies of the low-androstenone haplotype. The additive genetic effect of the individual was derived from a polygenic evaluation including up to ten generations of pedigree. Analyses were performed separately within each line. More details about the models used for each trait are described in Table S4.1 of the supplementary material.

In total, we tested 21 associations between phenotypes and the marker. From these associations, two were considered to be significant ($P \leq 0.05$), which is slightly more than what would be expected by chance alone (Table 4.1).

For the dam lines, a tendency for ($P = 0.06$ for Dutch Landrace) and a significant ($P = 0.01$ for Large White) favorable effect for number of teats (NTEAT) were detected for the low-androstenone haplotype. For Large White, a significant ($P = 0.03$) favorable effect of the low-androstenone haplotype on number of spermatozoa per ejaculation (NSPERM) was detected even though the number of observations was not very large (n = 200). No effects were identified on production and other reproduction traits, indicating that this QTL region does not unfavorably affect other breeding-goal traits. Because the low-androstenone haplotype shows a favorable effect on NTEAT and NSPERM, it suggests that selection for animals with a low-androstenone haplotype may lead to a greater NTEAT for Dutch Landrace and both NTEAT and NSPERM for Large White.

Table 4.1 SNP association and effect for traits under study in three commercial lines.

Trait	N[1]	Freq. FA[2]	SNP effect[3]	SE[4]	p-value
	♀ Dutch Landrace				
Growth rate (g/day)	1,265	0.19	5.27	3.61	0.15
Backfat thickness (mm)	1,265	0.19	-0.03	0.12	0.77
Litter birth weight (kg)	735	0.20	-0.09	0.15	0.56
Total number of piglets born	736	0.20	-0.08	0.15	0.57
Birth weight (kg)	1,364	0.19	0.02	0.02	0.27
Nr. of teats	1,430	0.19	0.11	0.06	0.06
Sperm motility (% of motile cells)	313	0.18	1.18	0.36	0.89
NSPERM (billions)	314	0.18	-1.33	1.57	0.40
	♀ Large White				
Growth rate (g/day)	1,279	0.46	2.05	2.97	0.49
Backfat thickness (mm)	1,277	0.46	0.08	0.08	0.33
Litter birth weight (kg)	901	0.47	-0.06	0.10	0.52
Total number of piglets born	922	0.47	-0.09	0.10	0.38
Birth weight (kg)	1,076	0.44	-0.01	0.01	0.50
Nr. of teats	1,280	0.46	0.11	0.05	0.01*
Sperm motility (% of motile cells)	200	0.41	-0.36	0.33	0.29
NSPERM (billions)	200	0.41	3.58	1.59	0.03*
	♂ Pietrain				
Growth rate (g/day)	864	0.85	0.24	3.02	0.93
Backfat thickness (mm)	859	0.85	-0.06	0.07	0.38
Total number of piglets born	194	0.86	0.06	0.22	0.78
Sperm motility (% of motile cells)	145	0.82	-0.19	0.59	0.75
NSPERM (billions)	159	0.83	3.00	2.36	0.21

[1] Number of animals genotyped for the SNP
[2] Frequency of the favorable allele (low-androstenone)
[3] Effect of each SNP measured in the same unit as the analyzed trait
[4] Standard error for the SNP effect
* Significant association at the 5% level between SNP and trait
NSPERM - number of spermatozoa per ejaculation

The favorable effect of androstenone on NTEAT concurs with a QTL found for NTEAT in a Yorkshire x Meishan F2 population (Zhang *et al.* 2007). The QTL was found on SSC6 at 95 cM, corresponding to the 52.7 - 85.8 Mbp region on Sscrofa10.2 (Hu *et al.*

2005), which does not overlap, but is near the androstenone QTL identified by Hidalgo *et al.* (2014) (2.5 Mbp downstream). The Meishan alleles were reported to increase the NTEAT, in accordance with the Asian origin of the haplotypes increasing NTEAT in the current study. These results indicate that the continued presence of the Asian haplotypes in European breeds could be explained by a favorable effect of the Asian haplotypes on NTEAT for the dam lines, which are often selected for NTEAT.

Uzu & Bonneau (1980) and Strathe *et al*. (2013) did not find a significant relation between androstenone level in fat and NSPERM. The biological explanation for a favorable effect of androstenone on NSPERM is still uncertain. We speculate that in the course of time since introgression of the Asian haplotype, a low level of androstenone may have led to an increase in NSPERM because of a compensatory system: animals with lower levels of androstenone would have fewer matings, but when there is a pleiotropic effect that results in more NSPERM, a higher proportion of these matings would result in a pregnancy. The product of the effect on number of matings and on number of successful matings would lead to a similar or even higher fitness for low-androstenone animals. On the other hand, this relation can also be considered as a trade-off in NSPERM when levels of androstenone are increased.

Within the 1.94 Mbp region, we looked for candidate genes that may affect NSPERM. The zinc finger protein 541 gene (*ZNF541*) is expressed in testicular cells in mice and is involved in chromatin remodeling during spermatogenesis. *ZNF541* encodes a nuclear protein that has a potential role in chromatin remodeling and its expression is dependent on the developmental stage during spermatogenesis (Choi *et al*. 2008). In dairy cattle, Peñagaricano *et al*. (2012) performed a genome-wide study on sire conception rate and detected a significant SNP located 16 kb upstream of the *ZNF541* gene, corroborating the possible role of *ZNF541* in spermatogenesis and, therefore, in male fertility. The DEAH (Asp-Glu-Ala-His) box polypeptide 34 gene (*DHX34*) is a candidate tumor suppressor gene for gliomas (Abdelhaleem *et al.* 2003). It codes for a member of the DEAD box proteins, which are putative RNA helicases. Some members of this family have been reported to be involved in embryogenesis, spermatogenesis, and cellular growth and division. The observed pleiotropic effect between low-androstenone and NSPERM could therefore be caused by genetic hitchhiking of the low-androstenone mutation with an allele within *ZNF541* or *DHX34* affecting NSPERM.

The genetic correlations found between androstenone and production traits, and androstenone and reproduction traits (Table 4.2) varied in their direction. For male and female reproduction traits, unfavorable correlations were found for NSPERM for Large White and NTEAT in both dam lines. These correlations indicate that the significant associations found between the marker and these traits may have an influence in the overall genetic correlation, as those were the only ones different from 0. On the other hand, for production traits, a favorable correlation between androstenone and backfat thickness was found in all lines (also found by Duijvesteijn *et al*. (2012)) and an unfavorable correlation was found for average growth in the sire line. The lack of association between the marker and production traits is a positive point in light of the unfavorable genetic correlation found for average growth. Knowledge about the pleiotropic effects of a marker is important even when the genetic correlation is known; especially for traits that can only be measured after slaughter. This is important to assure that haplotypes selected early in life will not unfavorably affect other traits.

Table 4.2 Genetic correlations ± standard error between androstenone and production traits, and androstenone and reproduction traits.

Trait	Dutch Landrace	Large White	Pietrain
GR	0.07 ± 0.08	-0.08 ± 0.14	0.22 ± 0.10
BF	0.35 ± 0.07	0.39 ± 0.12	0.27 ± 0.09
LBW	-0.08 ± 0.10	0.06 ± 0.16	-
TNB	-0.01 ± 0.11	-0.10 ± 0.18	-0.06 ± 0.20
BW	-0.28 ± 0.15	-0.34 ± 0.22	-
NTEAT	-0.13 ± 0.06	-0.24 ± 0.12	-
MOTILITY	-0.37 ± 0.29	-0.37 ± 0.33	-0.11 ± 0.27
NSPERM	0.00 ± 0.19	-0.59 ± 0.24	-0.24 ± 0.20

GR - growth rate (g/day); BF - backfat thickness (mm); LBW - litter birth weight (kg); TNB - total number of piglets born; BW - birth weight (kg); NTEAT - number of teats; MOTILITY - sperm motility (% of motile cells); NSPERM - number of spermatozoa per ejaculation (billions)

In summary, selection for the Asian low-androstenone haplotypes within the 1.94 Mbp region would not unfavorably affect other breeding goal traits, even suggesting favorable results for NTEAT and NSPERM in some lines.

Acknowledgements

The first author currently benefits from a joint grant from the European Commission (within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG"), Breed4Food (a public-private partnership in the domain of animal breeding and genomics) and Topigs Norsvin. Topigs Norsvin had a role in study design, data collection, and preparation of the manuscript.

References

Abdelhaleem M, Maltais L, Wain H (2003) The human DDX and DHX gene families of putative RNA helicases. Genomics 81:618–622.

Bonneau M (1982) Compounds responsible for boar taint, with special emphasis on androstenone: a review. Livest Prod Sci 9:687–705.

Choi E, Han C, Park I, et al (2008) A novel germ cell-specific protein, SHIP1, forms a complex with chromatin remodeling activity during spermatogenesis. J Biol Chem 283:35283–94.

Duijvesteijn N, Knol EF, Bijma P (2012) Direct and associative effects for androstenone and genetic correlations with backfat and growth in entire male pigs 1. J Anim Sci 90:2465–2475.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.

Giuffra E, Kijas JM, Amarger V, et al (2000) The origin of the domestic pig: independent domestication and subsequent introgression. Genetics 154:1785–91.

Grindflek E, Lien S, Hamland H, et al (2011) Large scale genome-wide association and LDLA mapping study identifies QTLs for boar taint and related sex steroids. BMC Genomics 12:362.

Hidalgo AM, Bastiaansen JWM, Harlizius B, et al (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.

Hu Z-L, Dracheva S, Jang W, et al (2005) A QTL resource and comparison tool for pigs: PigQTLDB. Mamm Genome 15:792–800.

Lee GJ, Archibald a L, Law a S, et al (2005) Detection of quantitative trait loci for androstenone, skatole and boar taint in a cross between Large White and Meishan pigs. Anim Genet 36:14–22.

Mathur PK, ten Napel J, Crump RE, et al (2013) Genetic relationship between boar taint compounds, human nose scores, and reproduction traits in pigs. J Anim Sci 91:4080–9.

Peñagaricano F, Weigel K a, Khatib H (2012) Genome-wide association study identifies candidate markers for bull fertility in Holstein dairy cattle. Anim Genet 43:65–71.

Ramos AM, Crooijmans RPMA, Affara NA, et al (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.

Sellier P, Bonneau M (1988) Genetic relationships between fat androstenone level in males and development of male and female genital tract in pigs. J Anim Breed Genet 105:11–20.

Strathe AB, Velander IH, Mark T, et al (2013) Genetic parameters for male fertility and its relationship to skatole and androstenone in Danish Landrace boars. J Anim Sci 91:4659–68.

Tajet H, Andresen Ø, Meuwissen THE (2006) Estimation of genetic parameters of boar taint; skatole and androstenone and their correlations with sexual maturation. Acta Vet Scand 48:S9.

Uzu G, Bonneau M (1980) Relations entre la production spermatique et la teneur en androsténone dans les graisses du jeune verrat. Ann Zootech 29:23–30.

Willeke H, Claus R, Müller E, et al (1987) Selection for high and low level of 5α-androst-16-en-3-one in boars. J Anim Breed Genet 104:64–73.

Zhang J, Xiong Y, Zuo B, et al (2007) Detection of quantitative trait loci associated with several internal organ traits and teat number trait in a pig population. J Genet Genomics 34:307–14.

Supplementary material

Table S4.1 Models used for association study

Models
GR = u + sex + HYM + SNP + *animal* + *litter*
BF = u + sex + HYM + weight + method + SNP + *animal* + *litter*
LBW = u + HYM + parity + tnb + SNP + *dam* + *pe*
TNB = u + HYM + parity + SNP + *dam* + *pe*
BW = u + sex + HYM + tnb + parity + SNP + *animal* + *litter*
NTEAT = u + sex + HYM + SNP + *animal* + *litter*
MOTILITY = u + HYM + station + lab + interval + age + conc + SNP + *animal* + *pe*
NSPERM = u + HYM + station + lab + interval + age + SNP + *animal* + *pe*

Random effects are in italics

GR - growth rate (g/day); BF - backfat thickness (mm); LBW - litter birth weight (kg); TNB - total number of piglets born; BW - birth weight (kg); NTEAT - number of teats; MOTILITY - sperm motility (% of motile cells); NSPERM - number of spermatozoa per ejaculation (billions)

u - mean; sex - sex of the animal; HYM - contemporary group defined by farm of birth, year, and month; SNP - genotype of the SNP marker; weight - weight of the animal when measurement was taken; method - method of backfat measurement; parity - number of parities of the dam; station - artificial insemination station; lab - laboratory where sperm was examined; interval - interval of semen collection; age - age of animal when semen was collected; conc - sperm concentration; animal - additive genetic effect of the animal; litter - litter in which the animal was born; pe - permanent environment effect

5

Accuracy of predicted genomic breeding values in purebred and crossbred pigs

A.M. Hidalgo[1,2], J.W.M. Bastiaansen[1], M.S. Lopes[1,3], B. Harlizius[3], M.A.M. Groenen[1], D.J. de Koning[2]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6708WD, the Netherlands; [2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, 750 07, Sweden; [3] Topigs Norsvin, Beuningen 6640AA, the Netherlands

G3: Genes | Genomes | Genetics (2015) 5:1575-1583

Abstract

Genomic selection has been widely implemented in dairy cattle breeding where the aim is to improve performance of purebred animals. In pigs, however, the final product is a crossbred animal. This may affect the efficiency of methods that are currently implemented for dairy cattle. Therefore, the objective of this study was to determine the accuracy of predicted breeding values in crossbred pigs using purebred genomic and phenotypic data. A second objective was to compare the predictive ability of SNP when training is done in either single or multiple populations for four traits: age at first insemination (AFI), total number of piglets born (TNB), litter birth weight (LBW) and litter variation (LVR). We performed marker- and pedigree-based predictions. Within-population predictions for the four traits ranged from 0.21 to 0.72. Multi-population prediction yielded accuracies ranging from 0.18 to 0.67. Predictions across purebred populations as well as predicting genetic merit of crossbreds from their purebred parental lines for AFI performed poorly (not significantly different from zero). In contrast, accuracies of across-population predictions and accuracies of purebred to crossbred predictions for LBW and LVR ranged from 0.08 to 0.31 and 0.11 to 0.31, respectively. Accuracy for TNB was zero for across-population prediction, whereas for purebred to crossbred prediction it ranged from 0.08 to 0.22. In general, marker-based outperformed pedigree-based prediction across populations and traits. However in some cases pedigree-based prediction performed similarly or outperformed marker-based prediction. There was predictive ability when purebred population(s) were used to predict crossbred genetic merit using an additive model in the populations studied. AFI was the only exception, indicating that predictive ability depends largely on the genetic correlation between PB and CB performance, which was 0.31 for AFI. Multi-population prediction was no better than within-population prediction for the purebred validation set. Accuracy of prediction was very trait dependent.

Key words: across-population, genomic selection, multi-population, reproduction traits, within-population

5.1 Introduction

Genomic selection has been widely implemented in dairy cattle breeding where the aim is to improve performance of purebred animals (Berry *et al*. 2009; VanRaden *et al*. 2009; Hayes *et al*. 2009b). In pigs and poultry, however, the final product is a crossbred animal. This may affect the efficiency of methods that are currently implemented for dairy cattle. In pig breeding, multiple sire and dam lines are used, with a minimum of two lines (typically for crossbred sows) and often additional sire lines to produce a three- or four-way cross finisher pig (Lutaaya *et al*. 2001; Merks and De Vries 2002).

Selection based on genomic estimated breeding values (GEBV) for purebreds (PB) using phenotypes on crossbreds (CB) is expected to increase the response to selection observed in CB compared to the situation in which only PB phenotypes are used. This increased response is expected when the genetic correlation between the PB and CB trait is less than 1, especially when the genetic correlation is 0.7 or less (Dekkers 2007). Genetic correlations between PB and CB performance vary and can be considerably less than 1 (Lutaaya *et al.* 2001; Zumbach *et al.* 2007; Cecchinato *et al.* 2010). Adding CB individuals to the training data is very expensive because, besides genotyping, it also requires additional identification and individual recording of target traits. Breeding companies are not inclined to make these investments unless there is evidence that predictions yield greater gains and higher accuracies. Simulation studies have shown that the response to selection is greater when PB animals are selected based on CB performance and that accuracy of prediction is high (Dekkers 2007; Ibánez-Escriche *et al.* 2009; Kinghorn *et al.* 2010; Toosi *et al.* 2010; Zeng *et al.* 2013). There is, however, a lack of studies using real data. The number of genotyped crossbreds is not yet large enough to test the superiority of training on CB for PB selection. A first step towards finding the optimal genomic selection scenario for pigs is to determine predictive ability (accuracy), in real data, of GEBV for CB pigs based on PB genomic and phenotypic data. This will show how CB performance responds to the current practice of selection on GEBV in PB pigs.

Recently, accuracies of within-population genomic prediction in pigs have been reported (Cleveland *et al.* 2010; Forni *et al.* 2011; Christensen *et al.* 2012; Tusell *et al.* 2013; Badke *et al.* 2014). These studies have shown that all traits had more than zero predictive ability within population in a variety of pig breeds using different methods. It has also been shown that using genomic information generally increased the accuracy of prediction compared to using only pedigree information (Forni *et al.*

2011; Christensen *et al.* 2012; Tusell *et al.* 2013). Using multi-population training might be a way fo increase the accuracy of prediction further. This is especially relevant to enable genomic selection for small populations when a closely related breed, or the same breed from another country, is added to the training set (Lund *et al.* 2014). An unresolved question is how to obtain accurate predictions from multi-population datasets. The effectiveness of a multi-population genomic evaluation depends on many factors, e.g., differences in allele frequency and consistency of linkage disequilibrium (LD) between quantitative trait loci (QTL) and single nucleotide polymorphism (SNP), which could reduce the accuracy of prediction (Wientjes *et al.* 2013) whereas the larger reference population would potentially improve the accuracy.

The objective of our study was to determine predictive ability (accuracy) in CB pigs using real PB genomic and phenotypic data. The outcome is a first step towards determining the optimal genomic selection scenario to select PB for CB performance. As in cattle, studying accuracy of prediction for multi-population datasets is important for species in which population size imposes upper limits to the training population size. Therefore, a second objective was to compare the predictive ability of SNP when training is done in either single or multiple populations in pigs.

5.2 Material & Methods

5.2.1 Data

Genotypes were available from sows with own-performance information of three pig populations born from 2005 through 2012: 1,070 Dutch Landrace-based (DL) sows from 19 farms, 1,389 Large White-based (LW) sows from 14 farms and 287 individuals from an F1 cross between these two commercial lines (DL sire/LW dams) originating from three farms. The genotyped CB animals had no specific family structure and the majority of them were not offspring of the genotyped PB animals, i.e., a number of generations separated PB and CB. The 287 CB animals were offspring from 76 sires and 170 dams. Four female reproduction traits were analyzed: age at first insemination (AFI), total number of piglets born (TNB), litter birth weight (LBW) and litter variation (LVR). AFI consisted of the age at the second estrus, which was the time that the first insemination was performed. TNB was the sum of all piglets born alive and stillborn. LBW was the sum of individual birth weights of all piglets born in the same litter. Finally, LVR consisted of the standard deviation (SD) of individual birth weight of the piglets from the same litter.

The PB and CB sows that were selected for genotyping have phenotypic records from multiple parities on multiple traits and have a large genetic contribution to future descendants. All PB sows were breeding animals from nucleus farms, whereas the CB sows belonged to farms where combined crossbred and pure line selection (CCPS) is applied. There was no strong selection for first parity performance in the genotyped sows, reducing any possible bias in TNB and LBW due to culling after first parity.

Deregressed estimated breeding values (DEBV) were used as response variable for each trait undergoing study. The estimated breeding values (EBV) were deregressed for each trait separately using the methodology proposed by Garrick *et al.* (2009). DEBV, instead of EBV, were used to compute the GEBV accuracy because this removes the influence of the parents EBV and rescales the EBV according to its accuracy, i.e. the DEBV of the animals reflect their genetic merit. Ostersen *et al.* (2011) have shown that using DEBV rather than EBV for genomic prediction yields higher GEBV accuracies. The number of animals and records used to estimate the EBV are in Table 5.1. The EBV of each animal was obtained from the routine genetic evaluation by Topigs Norsvin using MiXBLUP (Mulder *et al.* 2012) in a multi-trait model (including all measured reproduction traits). The genetic evaluation was done across lines with phenotypes from the different populations treated as the same trait. A fixed line effect was included in the model for estimating EBV. In multi-population prediction scenarios, this line effect was added back to the random additive genetic effect after estimating the EBV, and subsequently, the line effect was again included in the genomic prediction model. Adding back the line effect allows the differences of the level of EBV between-population to be maintained in the data. Therefore, in the genomic-prediction step the mean differences between populations are still present and this allows SNP effects (that differ in allele frequencies between lines) to explain these differences between lines.

The model for obtaining the EBV for AFI included genetic line and herd-year-season as fixed effects and an additive genetic effect (animal) as random effect. For TNB, the fixed effects were genetic line, parity, interval between weaning and pregnancy (days), whether more than one insemination procedures were performed (yes or no) and herd-year-season. 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). EBV for LBW were obtained with a model that included genetic line, parity number, TNB and herd-year-season as fixed effects and a permanent effect and an additive genetic effect (animal) as random effects. The model used for LVR was similar to the one used for LBW, except that TNB was removed. The reliabilities per animal, needed for deregression, were extracted from the genetic evaluation based on the methodology of Tier and Meyer (Tier and Meyer 2004). The heritabilities (h^2) used for deregression were estimated via restricted maximum likelihood (REML) using a pedigree-based relationship matrix and were also obtained from the routine genetic evaluation. The h^2 of the traits were 0.30 for AFI, 0.11 for TNB, 0.38 for LBW and 0.14 for LVR. The genomic h^2 of the DEBV were estimated via REML using ASREML 3.0 (Gilmour *et al.* 2009).

Table 5.1 Number of phenotypes on crossbred and purebreds that were used to estimate the breeding values.

Trait	No.	DL	LW	F1	Total
AFI	Records	304853	203933	190828	699614
	Animals	304853	203933	190828	699614
TNB	Records	1483099	910349	864551	3257999
	Animals	344583	223088	211117	778788
LBW	Records	158546	152722	7051	318319
	Animals	46221	43403	2093	91717
LVR	Records	158167	146500	7037	311704
	Animals	46124	42350	2083	90557

AFI - age at first insemination, TNB - total number of piglets born, LBW - litter birth weight, LVR - litter variation

Sows were genotyped using the Illumina PorcineSNP60 BeadChip (Ramos *et al.* 2009). SNP with GenCall <0.15, unmapped SNP and SNP located on either the X or Y chromosome, according to the Sscrofa10.2 assembly of the reference genome (Groenen *et al.* 2012), were excluded. Quality control was performed in all populations simultaneously, which involved excluding SNP with call rate <0.95, minor allele frequency <0.01 and strong deviations of Hardy-Weinberg equilibrium (χ^2>600). After quality control, 42,139 SNP remained out of the initial 64,232 SNP. Individuals with missing genotype frequency >0.05 were also removed. Missing genotypes of the remaining animals were imputed using BEAGLE 3.3.2 (Browning and Browning 2007).

5.2.2 Statistical analyses

GEBV were computed based on the genomic best linear unbiased prediction method (GBLUP). GBLUP uses a genomic relationship matrix (**G**) instead of the numerator relationship matrix (**A**). The **G** matrix contains genomic kinship indicating relatedness between animals and was used for prediction in all scenarios with the model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e},$$

where **y** is the vector of DEBV, μ is the overall mean, **g** is the vector of random-additive genetic effects assumed to be ~N(**0**, $\mathbf{G}\sigma_a^2$), **Z** is a design matrix allocating **g** to **y**, and **e** is a residual with heterogeneous variance due to differences in reliabilities of the DEBV (Garrick *et al.* 2009). In predictions where the training set contained more than one population, the fixed line effect present in the model for estimating EBV was also included in the GBLUP model as a fixed effect.

The **G** matrix for within-population prediction was built according to VanRaden (2008), which was computed as

$$\mathbf{G} = \mathbf{ZZ}' / 2\sum p_i q_i \,,$$

where **Z** is a matrix of centered genotypes and p_i and q_i are the allelic frequencies of the *ith* marker based on observed genotypes. In predictions where the training set contained more than one population, the **G** matrix was built according to Chen *et al.* (2013), accounting for differences in allele frequencies between populations.
We used ASREML 3.0 (Gilmour *et al.* 2009) to predict the GEBV with the **G** matrix entered as a user-defined matrix. Animals assigned to the prediction set had their DEBV removed before predicting GEBV.

All scenarios were also analyzed using the **A** matrix, which contains the average additive genetic relationships of the animals based on the pedigree (PED-BLUP). The model for these analyses was similar to the GBLUP one, however the **g** vector of random-additive genetic effect was assumed to be ~N(**0**, $\mathbf{A}\sigma_a^2$).

Genetic correlations between PB and CB performance were estimated for the four traits. We used records for DL, LW and F1 animals born from 2005 through 2012 (Supporting Information, Table S5.1). Genetic correlations were estimated in bivariate analyses using REML in ASREML 3.0 (Gilmour *et al.* 2009). The effects of bivariate models were the same as those used to obtain the EBV (see above), however to estimate genetic correlations PB and CB performance were treated as different traits (Falconer 1952), which in matrix notation is:

$$\begin{bmatrix}\mathbf{y_1}\\ \mathbf{y_2}\end{bmatrix} = \begin{bmatrix}\mathbf{Z_1} & 0\\ 0 & \mathbf{Z_2}\end{bmatrix}\begin{bmatrix}\mathbf{g_1}\\ \mathbf{g_2}\end{bmatrix} + \begin{bmatrix}\mathbf{e_1}\\ \mathbf{e_2}\end{bmatrix}$$

where $\mathbf{y}_i$ is the vector of observations with *i* being 1 for purebred and 2 for crossbred data, $\mathbf{Z}_i$ is the incidence matrix for $\mathbf{g}_i$, which is a vector of random additive genetic effects. The additive genetic variance is expressed as:

$$\mathrm{var}\begin{bmatrix}\mathbf{g_1}\\ \mathbf{g_2}\end{bmatrix} = \mathbf{G_0} \otimes \mathbf{A}$$

where **A** is the numerator relationship matrix and $\mathbf{G_0}$ is a 2x2 covariance matrix with the purebred and crossbred variances in the diagonals, and the covariances in the off-diagonals.

5.2.3 Scenarios and accuracy of prediction

Seventeen scenarios were investigated that can be divided into four groups according to composition of the training and validation data sets as follows:

- Scenarios 1-3: Training and validation data were subsets from the same population, DL, LW and F1 respectively, i.e., prediction was within-population. These scenarios determine how well the within-population prediction performs for the different traits.
- Scenarios 4-7: Same as scenarios 1-3 but the remaining PB population(s) was/were added to the training data, i.e., prediction was multi-population. These scenarios determine whether adding data from a different PB population to the training data would increase the accuracy compared to the within-population prediction.
- Scenarios 8-11: One PB population was used for training to predict the other PB population. F1 data were not used in these scenarios, i.e., prediction was across breeds. These scenarios determine how well across-population predictions would perform.
- Scenarios 12-17: PB populations were used for training and CB animals were used for validation. These scenarios determine how well CB genetic merit can be predicted from PB data alone, and whether inclusion of more than one parental PB population increases the accuracy.

The accuracy of prediction was estimated as the correlation between the GEBV/EBV and the DEBV of the validation set animals for GBLUP/PED-BLUP. Prediction bias was calculated by regressing the validation variables (DEBV) on the prediction variables (GEBV/EBV). Accuracies were the average of a 20 random training-validation populations in scenarios 1-7, 9, 11, 13, 15 and 17. For scenarios 1-7, we randomly set aside part of the genotyped animals (N=50) and used those in a later step to determine the accuracy of prediction. These 50 were not included in the training for those scenarios. In scenarios 9, 11, 13, 15 and 17 not all the available animals were used for training. Subsets of the training populations were sampled such that the same number of animals was used from each population per trait under study. Any differences in accuracies would then be due to the different populations used, and not to differences in the number of animals. Scenarios 8, 10, 12, 14 and 16 only had one estimate of accuracy because all the animals were used in the training population to maximize prediction accuracy of animals in another population.

5.3 Results

Estimates of genomic h^2 of the DEBV across traits and populations ranged from 0.04 to 0.58 (Table 5.2). Estimates of pedigree-based h^2 of the DEBV across traits and populations ranged from 0.03 to 0.78 (Table S5.2). The genomic and pedigree-based heritabilities were similar in general. Genetic correlations between PB and CB performance for the four traits under study ranged from 0.31 for AFI to 0.90 for LBW (Table 5.3).

Table 5.2 Estimated genomic heritability (h^2) of the deregressed estimated breeding values across traits and populations under study.

	Heritability (SE)		
Trait	DL	LW	F1
AFI	0.18 (0.04)	0.07 (0.02)	0.64 (0.12)
TNB	0.04 (0.01)	0.05 (0.01)	0.12 (0.05)
LBW	0.58 (0.05)	0.57 (0.04)	0.43 (0.12)
LVR	0.21 (0.03)	0.11 (0.02)	0.17 (0.07)

DL - Dutch Landrace, LW - Large White, F1 - cross between DL and LW, SE - standard error, AFI - age at first insemination, TNB - total number of piglets born, LBW - litter birth weight, LVR - litter variation

Table 5.3 Genetic correlations between purebred and crossbred performance for the four traits under study

Trait	Genetic correlation (SE)
AFI	0.31 (0.02)
TNB	0.88 (0.01)
LBW	0.90 (0.05)
LVR	0.88 (0.06)

SE – standard error, AFI- age at first insemination, TNB- total number of piglets born, LBW- litter birth weight, LVR- litter variation

Accuracies for within-population predictions for scenarios 1-3 ranged from 0.22 to 0.72 for GBLUP and from 0.21 to 0.64 for PED-BLUP across the four traits and different training sets, indicating a modest to good predictive ability (Table 5.4). The regression coefficient of the GEBV/EBV on the DEBV for scenarios 1-3 ranged from 1.03 to 1.70 for GBLUP and from 0.90 to 2.21 for PED-BLUP.

For multi-population prediction of PB populations (scenarios 4 and 5) the accuracies ranged from 0.18 to 0.67, whereas for multi-population prediction (two PB + one CB) of the CB population (scenarios 6-7) the accuracies ranged from 0.17 to 0.45 for GBLUP and from 0.32 to 0.42 for PED-BLUP. When predicting PB (scenarios 4 and 5, Table 5.5), the addition of the other PB population resulted in lower accuracies for all four traits, in comparison to within-population prediction for GBLUP. When predicting CB (scenarios 6 and 7; Table 5.5) the addition of PB populations resulted in lower accuracies for AFI and TNB but higher accuracies for LBW and LVR. The regression coefficient of the GEBV/EBV on the DEBV for scenarios 4 and 5 ranged from 0.86 to 1.18 for GBLUP, whereas for scenarios 6 and 7 ranged from 0.80 to 3.11 for GBLUP and from 0.97 to 5.00 for PED-BLUP. Accuracies and regression coefficients of the EBV on the DEBV were not computed for PED-BLUP for scenarios 4 and 5 because the other PB population to be added is not related according to the pedigree.

GEBV accuracy of across-breed prediction, i.e., predicting genetic merit of one PB from a different PB population, performed poorly for AFI and TNB (Table 5.6), accuracies were not significantly different from zero (P>0.05). Accuracies for LBW and LVR ranged from 0.13 to 0.26 across the different training sets for GBLUP. The regression coefficient of the GEBV on the DEBV for AFI and TNB ranged from -0.71 to 1.37, whereas for LBW and LVR ranged from 0.70 to 1.40. Accuracies and regression coefficients of the EBV on the DEBV were not computed for PED-BLUP because the two PB populations are not related according to the pedigree.

Accuracy of prediction in scenarios 12-17 that predicted genetic merit of CB using PB parental populations as training data performed poorly for AFI (Table 5.7), accuracies were not significantly different from zero for both GBLUP and PED-BLUP (P>0.05). For the other three traits, TNB, LBW and LVR, however, predictive ability was observed. Accuracies ranged from 0.11 to 0.31 for GBLUP and from 0.08 to 0.22 for PED-BLUP. The regression coefficient of the GEBV/EBV on the DEBV for AFI ranged from -1.14 to -0.15 for GBLUP and from 0.15 to 0.95 for PED-BLUP, whereas for TNB, LBW and LVR it ranged from 0.48 to 3.82 for GBLUP and from 0.53 to 7.76 for PED-BLUP.

Table 5.4 GEBV accuracies from within-population prediction using GBLUP and PED-BLUP (scenarios 1-3)

			N training			N validation			Accuracy† (SD)		Slope*	
Trait	Scenario	r^2	DL	LW	F1	DL	LW	F1	GBLUP	PED-BLUP	GBLUP	PED-BLUP
	1	0.45	1017	-	-	50	-	-	0.26 (0.16)	0.25 (0.15)	1.21	1.13
GLE	2	0.45	-	1339	-	-	50	-	0.22 (0.09)	0.21 (0.19)	1.25	1.08
	3	0.33	-	-	237	-	-	50	0.37 (0.09)	0.30 (0.12)	1.04	0.90
	1	0.45	1016	-	-	50	-	-	0.26 (0.12)	0.25 (0.12)	1.70	1.90
TNB	2	0.49	-	1333	-	-	50	-	0.24 (0.15)	0.25 (0.15)	1.24	1.50
	3	0.40	-	-	231	-	-	50	0.40 (0.11)	0.35 (0.14)	1.52	2.21
	1	0.78	1020	-	-	50	-	-	0.64 (0.09)	0.58 (0.06)	1.08	1.06
LBW	2	0.80	-	1335	-	-	50	-	0.72 (0.06)	0.64 (0.07)	1.03	1.05
	3	0.77	-	-	236	-	-	50	0.40 (0.11)	0.39 (0.13)	1.10	1.27
	1	0.50	1019	-	-	50	-	-	0.50 (0.11)	0.40 (0.10)	1.04	1.03
LVR	2	0.53	-	1335	-	-	50	-	0.46 (0.09)	0.39 (0.15)	1.05	1.17
	3	0.49	-	-	235	-	-	50	0.34 (0.09)	0.33 (0.11)	1.03	1.19

SD - standard deviation, DL - Dutch Landrace, LW - Large White, F1 - cross between DL and LW

AFI - age at first insemination, TNB - total number of piglets born, LBW - litter birth weight, LVR - litter variation

† - Estimate obtained by 20-random training-validation populations

* - Regression coefficient of the GEBV/EBV on the DEBV

r^2 - Mean reliability of deregressed estimated breeding values from the training population

Table 5.5 GEBV accuracies from multi-population prediction using GBLUP and PED-BLUP (scenarios 4-7)

			N training			N validation			Accuracy[†] (SD)		Slope[*]	
Trait	Scenario	r^2	DL	LW	F1	DL	LW	F1	GBLUP	PED-BLUP	GBLUP	PED-BLUP
AFI	4	0.45	1017	1389	-	50	-	-	0.20 (0.13)	-	1.18	-
	5	0.45	1067	1339	-	-	50	-	0.18 (0.13)	-	1.11	-
	6	0.41	1067	1389	237	-	-	50	0.17 (0.09)	0.32 (0.12)	1.91	1.86
	7	0.41	237	237	237	-	-	50	0.27 (0.11)	0.35 (0.11)	3.11	2.17
TNB	4	0.47	1016	1383	-	50	-	-	0.21 (0.15)	-	1.12	-
	5	0.47	1066	1333	-	-	50	-	0.23 (0.17)	-	1.11	-
	6	0.44	1066	1383	231	-	-	50	0.31 (0.18)	0.34 (0.11)	1.69	3.06
	7	0.44	231	231	231	-	-	50	0.37 (0.14)	0.33 (0.12)	3.02	5.00
LBW	4	0.79	1020	1385	-	50	-	-	0.51 (0.13)	-	0.86	-
	5	0.79	1070	1335	-	-	50	-	0.67 (0.07)	-	1.09	-
	6	0.78	1070	1385	236	-	-	50	0.45 (0.11)	0.37 (0.09)	0.80	0.97
	7	0.78	236	236	236	-	-	50	0.41 (0.15)	0.37 (0.11)	1.03	1.10
LVR	4	0.52	1019	1385	-	50	-	-	0.38 (0.12)	-	0.99	-
	5	0.52	1069	1335	-	-	50	-	0.41 (0.12)	-	1.11	-
	6	0.51	1069	1385	235	-	-	50	0.44 (0.10)	0.40 (0.14)	1.22	1.59
	7	0.51	235	235	235	-	-	50	0.38 (0.12)	0.42 (0.08)	1.33	1.88

SD - standard deviation, DL - Dutch Landrace, LW - Large White, F1 - cross between DL and LW

AFI - age at first insemination, TNB - total number of piglets born, LBW - litter birth weight, LVR - litter variation

[†] - Estimate obtained by 20-random training-validation populations

[*] - Regression coefficient of the GEBV/EBV on the DEBV

r^2 - Mean reliability of deregressed estimated breeding values from the training population

Table 5.6 GEBV accuracies from across-population prediction using GBLUP (scenarios 8-11)

			N training			N validation			Accuracy (SD)	Slope*
Trait	Scenario	r^2	DL	LW	F1	DL	LW	F1	GBLUP	GBLUP
AFI	8	0.45	1067	-	-	-	1389	-	-0.05	-0.57
	9	0.45	711	-	-	-	1389	-	-0.04 (0.01)†	-0.71
	10	0.45	-	1389	-	1067	-	-	-0.02	-0.27
	11	0.45	-	711	-	1067	-	-	-0.02 (0.03)†	-0.43
TNB	8	0.45	1066	-	-	-	1383	-	0.05	1.01
	9	0.45	693	-	-	-	1383	-	0.04 (0.01)†	1.37
	10	0.49	-	1383	-	1066	-	-	0.03	0.56
	11	0.49	-	693	-	1066	-	-	0.00 (0.02)†	0.00
LBW	8	0.78	1070	-	-	-	1385	-	0.26	0.83
	9	0.78	708	-	-	-	1385	-	0.23 (0.04)†	0.83
	10	0.80	-	1385	-	1070	-	-	0.22	0.73
	11	0.80	-	708	-	1070	-	-	0.16 (0.03)†	0.65
LVR	8	0.50	1069	-	-	-	1385	-	0.17	0.70
	9	0.50	705	-	-	-	1385	-	0.15 (0.03)†	0.75
	10	0.53	-	1385	-	1069	-	-	0.20	1.40
	11	0.53	-	705	-	1069	-	-	0.13 (0.04)†	1.22

SD - standard deviation, DL - Dutch Landrace, LW - Large White, F1 - cross between DL and LW

AFI - age at first insemination, TNB - total number of piglets born, LBW - litter birth weight, LVR - litter variation

† - Estimate obtained by 20-random training-validation populations

* - Regression coefficient of the GEBV on the DEBV

r^2 - Mean reliability of deregressed estimated breeding values from the training population

Table 5.7 GEBV accuracies from prediction of crossbred genetic merit from purebred training data using GBLUP and PED-BLUP (scenarios 12-17)

			N training			N validation			Accuracy (SD)		Slope*	
Trait	Scenario	r^2	DL	LW	F1	DL	LW	F1	GBLUP	PED-BLUP	GBLUP	PED-BLUP
AFI	12	0.45	1067	1389	-	-	-	287	-0.07	0.09	-0.75	0.81
	13	0.45	356	356	-	-	-	287	-0.03 (0.06)†	0.04 (0.06)†	-0.40	0.53
	14	0.45	1067	-	-	-	-	287	-0.02	0.07	-0.15	0.95
	15	0.45	711	-	-	-	-	287	-0.02 (0.05)†	0.05 (0.04)†	-0.24	0.67
	16	0.45	-	1389	-	-	-	287	-0.07	0.06	-1.14	0.73
	17	0.45	-	711	-	-	-	287	-0.06 (0.05)†	0.01 (0.05)†	-0.95	0.15
TNB	12	0.47	1066	1383	-	-	-	281	0.20	0.21	1.51	3.02
	13	0.47	347	347	-	-	-	281	0.18 (0.09)†	0.22 (0.05)†	3.82	7.76
	14	0.45	1066	-	-	-	-	281	0.18	0.19	2.29	3.62
	15	0.45	693	-	-	-	-	281	0.18 (0.04)†	0.19 (0.04)†	3.15	4.71
	16	0.49	-	1383	-	-	-	281	0.13	0.10	1.17	2.23
	17	0.49	-	693	-	-	-	281	0.11 (0.04)†	0.08 (0.04)†	1.72	3.06
LBW	12	0.79	1070	1385	-	-	-	286	0.31	0.14	0.62	0.54
	13	0.79	354	354	-	-	-	286	0.18 (0.05)†	0.11 (0.05)†	0.52	0.60
	14	0.78	1070	-	-	-	-	286	0.26	0.10	0.65	0.53
	15	0.78	708	-	-	-	-	286	0.22 (0.04)†	0.14 (0.04)†	0.64	0.73
	16	0.80	-	1385	-	-	-	286	0.22	0.11	0.55	0.63
	17	0.80	-	708	-	-	-	286	0.17 (0.03)†	0.08 (0.05)†	0.48	0.59
LVR	12	0.52	1069	1385	-	-	-	285	0.27	0.15	0.90	0.91
	13	0.52	353	353	-	-	-	285	0.21 (0.08)†	0.13 (0.07)†	1.16	1.44
	14	0.50	1069	-	-	-	-	285	0.31	0.11	1.18	0.84
	15	0.50	705	-	-	-	-	285	0.28 (0.04)†	0.14 (0.05)†	1.24	1.25
	16	0.53	-	1385	-	-	-	285	0.15	0.11	0.74	1.33
	17	0.53	-	705	-	-	-	285	0.11 (0.04)†	0.12 (0.05)†	0.75	2.43

SD - standard deviation, DL - Dutch Landrace, LW - Large White, F1 - cross between DL and LW

AFI - age at first insemination, TNB - total number of piglets born, LBW - litter birth weight, LVR - litter variation

† - Estimate obtained by 20-random training-validation populations

* - Regression coefficient of the GEBV/EBV on the DEBV

r^2 - Mean reliability of deregressed estimated breeding values from the training population

5.4 Discussion

Accuracies of genomically-predicted breeding values in CB and PB pigs were estimated for four female reproduction traits in 17 scenarios to optimize the use of genomic data for crossbred animals. We have used DEBV as response variable with a moderate to high mean reliability (ranging from 0.33 to 0.80) for the different traits and populations. The SD of the accuracies in scenarios in which we had replicates of training-validation populations varied according to the type of prediction (within-, multi-, across-, or PB to CB). Within- and multi-population predictions showed higher SD because the relationship between training and validation in each replicate could substantially vary due to different degrees of relationship within a population. For across- and PB to CB predictions, the relationship between training and validation populations was naturally lower, therefore in each replicate there was less variation.

5.4.1 Within-population prediction

LBW and LVR showed generally higher accuracies than AFI and TNB. This difference between traits may occur due to the lower reliability of the DEBV for AFI and TNB, which lowers the accuracy when number of observations is preset. Another possibility is that there are non-additive genetic effects (e.g., dominance, epistasis) affecting AFI and TNB more, whereas LBW and LVR may be regulated mainly by an additive action of the genes. Therefore, the importance of non-additive effects needs to be further investigated. Even with the low number of genotyped CB pigs, all traits showed predictive ability within the CB. Therefore, a greater number of genotyped CB should increase these accuracies. In general, GBLUP outperformed PED-BLUP across populations and traits, which is mainly a result of a better estimation of relationship among individuals by the markers. Similar results have also been reported in other studies using pigs (Forni *et al.* 2011; Tusell *et al.* 2013). The regression coefficients of the GEBV/EBV on the DEBV for both GBLUP and PED-BLUP were, in general, close to 1, indicating that the predictions were not severely biased, except for TNB, where some of them deviated considerably from 1.

The level of accuracy found here is concordant with those found in other studies on pigs (Cleveland *et al.* 2010; Forni *et al.* 2011; Christensen *et al.* 2012; Tusell *et al.* 2013; Badke *et al.* 2014). In these studies, as well as in ours, many traits and breeds were studied and within-population prediction always showed to have predictive ability. One of the studies (Tusell *et al.* 2013) also studied TNB for two PB populations and their F1 cross and also found that prediction within the F1 cross has greater accuracy than within-PB prediction. They argued that this might be caused by the structure and effective sample size of the populations under study. Accuracies found by Christensen *et al.* (2012) were not statistically different between single-step BLUP (SS-BLUP) and GBLUP, but both were higher than pedigree-based prediction and GBLUP was shown to be more biased. The advantage of using SS-BLUP was an increase of accuracy for non-genotyped animals. Because our aim was to predict genotyped animals, we studied accuracies of prediction using GBLUP.

5.4.2 Multi-population prediction

Adding data from a different PB population to the training data (scenarios 4 and 5) decreased the accuracy of prediction compared with within-population predictions (scenarios 1-3) for GBLUP. Adding data from the two PB populations to the CB training data (scenario 6 and 7) had different results depending on the trait. LBW and LVR that had high genetic correlation between PB and CB performance had an increase in accuracy, whereas for AFI that had a low genetic correlation there was a decrease in accuracy. TNB had a high genetic correlation, however the accuracy also decreased, which was unexpected.

If traits are genetically very different (low genetic correlation between PB and CB), then adding more animals with the other trait to the training is not expected to increase the accuracy. When the trait is the same (high genetic correlation), however, including more animals with the other trait (PB vs. CB) is expected to increase the accuracy. Besides having a high genetic correlation between the traits, the additional animals also need to have some (genomic) relationship to the validation animals. In addition to a low genetic correlation between PB and CB performance, the degradation of accuracy might result from differences in non-additive effects.

For PED-BLUP, adding the two parental PB populations in the training also had different results depending on the trait. AFI and LBW had an increase in accuracy, whereas TNB and LVR had a slight decrease in accuracy. The regression coefficient of the GEBV/EBV on the DEBV estimated to investigate bias for scenarios 4-7 was, in general, close to 1, indicating that the predictions did not suffer from a large bias, except for AFI, and TNB in scenarios 6 and 7. For these traits, whenever the PB parental populations were used as training and CB as validation set, the regression coefficient of the GEBV/EBV on the DEBV indicated that the estimates were severely biased.

A review regarding multi-population prediction in cattle (Lund *et al.* 2014) has shown that combining populations, in general, increases the accuracy of prediction when the breeds are the same but from different countries, to a lesser degree when the breeds are closely related, and has little or no benefit when the breeds are distantly related. Another study (Hayes *et al.* 2009a) has reported slightly higher accuracies when using multi-population prediction compared to within-population prediction in dairy cattle. Chen *et al.* (2013) used Angus and Charolais steers to determine the accuracy of prediction with GBLUP for within-population and multi-population prediction. In their study, accuracies did not always increase, suggesting that noise was being added to the predictions. The maximum increment in accuracy that they obtained was of 0.05, whereas a decrement of 0.07 was also obtained, which is within the same range as the differences observed in the current study. These studies showed that adding another PB population to the training data in cattle did not necessarily increase the accuracy of prediction, similar to our current results in pigs.

De Roos *et al.* (2009) using simulated data, also showed that increasing the size of the training data by adding animals from a different population does not always increase the accuracy. An increase in accuracy over within-population was only found when the populations were closely related, when marker density was high, or when the size of initial within-population training data set was small. In our case, the number of markers was reasonable and in some scenarios the size of the within-population training data set was small, but still we did not have a great increase in accuracy of prediction. This suggests that the marker density might not be sufficient to have similar LD level between QTL and marker in the different populations that are mixed. The genetic distance between the populations was probably an important factor that limited the benefit of adding training data from other populations.

5.4.3 Across-population prediction

Some predictive ability was observed when predicting across populations for LBW and LVR, whereas for AFI and TNB all the accuracies were null. Increasing the size of the training population slightly improved the accuracies of prediction, on average by 0.05. Greater accuracies were found when DL predicted LW genetic merit, rather than the other way around (scenario 9 vs. scenario 11). The regression coefficients of the GEBV/EBV on the DEBV for scenarios 8-11 were, in general, close to 1 for LBW and LVR, indicating that the predictions did not suffer from much bias. For AFI and TNB, however, regression values greatly deviated from 1, sometimes with negative values, which we attribute to the very low accuracies we found.

In a study by Harris *et al.* (2008), the prediction across Holstein-Friesian and Jersey cattle breeds was also investigated. Predictions were not accurate, ranging from -0.1 to 0.3 for 25 traits. In another study, Hayes *et al.* (2009a) predicted the GEBV of Jersey animals using a Holstein population as training data and vice-versa, resulting in accuracies ranging from -0.06 to 0.23 for five traits. Both studies report results that were very similar to ours that ranged from -0.05 to 0.26.

The simulation study by De Roos *et al.* (2009) indicated that across-population prediction was substantially less accurate than within-population or multiple-population prediction. These lower accuracies were due to differences in marker–QTL LD phase between the populations. A marker may be in LD with QTL in a given population, but is not necessarily in LD with those QTL in the other population, resulting in poor predictions for the other population. These simulation results suggested that for our analyses a higher marker density would be required. Results of Veroneze *et al.* (2014) show that with the same 60K porcine SNP panel, the density of SNP is high enough to obtain reasonable levels of LD. This would predict that our SNP panel should be able to capture marker effects across breeds.

5.4.4 Using purebred training data to predict crossbred genetic merit

Using only the PB population(s) to predict the CB genetic merit with GBLUP has some predictive ability for TNB, LBW and LVR, whereas all the accuracies for AFI were null. Increasing the size of the training data by adding another PB population increased the accuracy for TNB and LBW, whereas for AFI and LVR it did not. However, when we increased the size of the training population by adding more animals of the same PB population, the accuracies usually increased. The accuracy of prediction for predicting CB animals based on PB animals appears to depend largely on the genetic correlation between PB and CB performance. As our results demonstrate, the greater the genetic correlation the higher the chances of having any or more predictive ability. AFI, for which the genetic correlation between PB and CB was very poor, had a zero accuracy of prediction showing that selection on PB is expected to have no effect on CB genetic merit.

For PED-BLUP, the accuracies were in general lower than for GBLUP especially for LBW and LVR. Adding the second PB population in the training slightly increased the accuracy of prediction.

The greater accuracies found for TNB, LBW and LVR when training with DL rather than LW population can be explained by the slightly greater relationship between DL and F1 populations than between LW and F1. This higher relationship is specific for the animals included in this study. The F1 animals that were genotyped are more closely related to the DL animals that were genotyped than the LW animals that were genotyped.

To test the impact of relationship between training and validation populations on the accuracy we split the training data into the 50% of animals that are MOST related to the validation set and the 50% that are LEAST related to the validation set (Supporting Information, Tables S5.3-S5.4). For AFI, TNB and LBW, using the 50% MOST related animals resulted in greater accuracies, whereas for LVR it did not. This indicates that if CB animals have closer genomic relationships to the PB animals used as training, higher accuracies for scenarios 12-17 could generally be expected.

In cattle, Harris *et al.* (2008) used PB populations (Holstein-Friesian and Jersey) to predict the genomic breeding values of a cross between these two breeds. They used data from 4,500 sires genotyped for approximately 44K SNP. Their results show that using the two breeds as training population increased the accuracy by 5 - 10% compared to using only one of the breeds to predict the cross. The actual level of accuracy was not reported in their study. Our results were similar to theirs for TNB, LBW and LVR, where the genetic correlation between PB and CB performance is close to 1, but not for AFI.

Results indicate that using a PB population to predict CB genetic merit can generate reasonable predictions. This, however, is not consistent for all traits. Although these results do not reflect the actual practice of genomic selection in pig breeding, they do provide an estimate of the accuracy of genomic prediction between CB and PB populations using real data. The results make a strong case for the genotyping and recording of CB animals, at least for a subset of traits where genetic correlations are away from 1.

The low genetic correlation between PB and CB performance for AFI was also found in another study (Nagyné-Kiszlinger *et al.* 2013). They have reported values of 0.28 and 0.39 for Hungarian Large White and Hungarian Landrace with their reciprocal cross. Possible reasons for this low genetic correlation were reported: 1) genes which affect the trait might be different among populations; 2) this trait is affected by non-additive effects or environmental factors due to different management of PB and CB animals (Nagyné-Kiszlinger *et al.* 2013). One clear environmental factor that probably reduces the genetic correlation of AFI between PB and CB is the use of batch farrowing systems in the production environment of CB sows. Suppression of estrus is used to synchronize the heat of the CB gilts which impacts the measurement of the trait and lead to these low correlations.

Standardized EBV were used, therefore, a bias would possibly be introduced during deregression due to different reliabilities between breeds (Garrick *et al.* 2009). Additional sources for potential bias affecting the SNP effect estimates are the differences in the population mean of the breeds. The differences in the mean between populations was remedied by reintroducing the line effect after deregression. To test the impact of deregression on bias we investigated all 17 scenarios for the trait AFI by analyzing phenotypes, which are not standardized, instead of DEBV. The correlation between the accuracies obtained by the two different approaches was 0.92, with a mean regression coefficient of the GEBV on the phenotype of 0.70. This correlation shows that using the phenotypes has good agreement with the accuracies calculated using DEBV; therefore any bias due to the process of standardization and deregression is expected to be limited.

The reasonable accuracy for PB predicting CB genetic merit shows that in a current typical breeding program, selection in the PB does result in a phenotypic response in CB. AFI was an exception in our study, as the genetic correlation between PB and CB performance was very low.

Further studies to compare the accuracy of genomic selection of PB for CB performance are needed. Other genomic models including breed-specific effects of SNP alleles or dominance (Ibánez-Escriche *et al.* 2009; Zeng *et al.* 2013) were described and were found to outperform an additive model only in specific cases, e.g., with high dominance levels or when the number of SNP is small relative to the size of the training population. Using these more complex models or a multiple-trait model (Christensen *et al.* 2014) to real data will be needed.

In conclusion, there was predictive ability for purebred population(s) predicting crossbred genetic merit using an additive model in the populations studied when PB and CB traits have high genetic correlation. For practical implementation, estimation of genomic breeding values of PB animals for CB performance needs to be further studied with models that take into account the crossbred nature of training data. Multi-population prediction was no better than within-population prediction for PB populations. Accuracy of prediction showed to be very trait dependent, hence, for the utility of data from other breeds in the application of genomic selection, each trait needs to be studied separately and no generalizations should be made. Finally, real-data accuracies were lower than what simulation studies have reported.

5.5 Acknowledgements

AMH currently benefits from a joint grant from the European Commission (within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG"), Breed4Food (a public-private partnership in the domain of animal breeding and genomics) and Topigs Norsvin. JWMB was supported by PPP Breed4Food. Topigs Norsvin had a role in study design, data collection, and preparation of the manuscript.

References

Badke YM, Bates RO, Ernst CW, et al (2014) Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3 Genes, Genomes, Genet 4:623–31.

Berry DP, Kearney F, Harris BL (2009) Genomic selection in Ireland. Interbull Bull 29–34.

Browning SR, Browning BL (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 81:1084–97.

Cecchinato A, de los Campos G, Gianola D, et al (2010) The relevance of purebred information for predicting genetic merit of survival at birth of crossbred piglets. J Anim Sci 88:481–490.

Chen L, Schenkel F, Vinsky M, et al (2013) Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. J Anim Sci 91:4669–78.

Christensen OF, Madsen P, Nielsen B, et al (2012) Single-step methods for genomic evaluation in pigs. Animal 6:1565–1571.

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

Cleveland MA, Forni S, Garrick DJ, Deeb N (2010) Prediction of genomic breeding values in a commercial pig population. Proc. 9th WCGALP. p 266

De Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. Genetics 183:1545–53.

Dekkers JCM (2007) Marker-assisted selection for commercial crossbred performance. J Anim Sci 85:2104–14.

Falconer DS (1952) The problem of environment and selection. Am Nat 86:293–298.

Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43:1.

Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 41:55.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.

Groenen MAM, Archibald AL, Uenishi H, et al (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393–8.

Harris BL, Johnson DL, Spelman RJ (2008) Genomic selection in New Zealand and the implications for national genetic evaluation. Proc. Interbull Meet. Niagara Falls, p 325–330

Hayes BJ, Bowman PJ, Chamberlain AC, et al (2009a) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41:51.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009b) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–43.

Ibánez-Escriche N, Fernando RL, Toosi A, Dekkers JCM (2009) Genomic selection of purebreds for crossbred performance. Genet Sel Evol 41:12.

Kinghorn BP, Hickey JM, Werf JHJ Van Der (2010) Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals. Proc. 9th WCGALP. p 36

Lund MS, Su G, Janss L, et al (2014) Genomic evaluation of cattle in a multi-breed context. Livest Sci 166:101–110.

Lutaaya E, Misztal I, Mabry JW, et al (2001) Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. J Anim Sci 79:3002–7.

Merks JWM, De Vries AG (2002) New sources of information in pig breeding. Proc. 7th WCGALP. p (30) 3–10

Mulder HA, Lidauer M, Strandén I, et al (2012) MiXBLUP Manual.

Nagyné-Kiszlinger H, Farkas J, Kövér G, Nagy I (2013) Selection for reproduction traits in hungarian pig breeding in a two-way cross. Anim Sci Pap Reports 31:315–322.

Ostersen T, Christensen OF, Henryon M, et al (2011) Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genet Sel Evol 43:38.

Ramos AM, Crooijmans RPMA, Affara NA, et al (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.

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.

Toosi A, Fernando RL, Dekkers JCM (2010) Genomic selection in admixed and crossbred populations. J Anim Sci 88:32–46.

Tusell L, Pérez-Rodriguez P, Forni S, et al (2013) Genome-enabled methods for predicting litter size in pigs : a comparison. Animal 7:1739–1749.

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–23.

VanRaden PM, Van Tassell CP, Wiggans GR, et al (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16–24.

Veroneze R, Bastiaansen JWM, Knol EF, et al (2014) Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (*Sus scrofa*) populations. BMC Genet 15:126.

Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621–31.

Zeng J, Toosi A, Fernando RL, et al (2013) Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol 45:11.

Zumbach B, Misztal I, Tsuruta S, et al (2007) Genetic correlations between two strains of Durocs and crossbreds from differing production environments for slaughter traits. J Anim Sci 85:901–908.

6

Accuracy of genomic prediction using deregressed breeding values estimated from purebred and crossbred offspring phenotypes in pigs

A.M. Hidalgo[1,2], J.W.M. Bastiaansen[1], M.S. Lopes[1,3], R. Veroneze[4], M.A.M. Groenen[1], D.J. de Koning[2]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6708WD, the Netherlands; [2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, 750 07, Sweden; [3] Topigs Norsvin, Beuningen 6640AA, the Netherlands; [4] Departamento de Zootecnia, Universidade Federal de Viçosa, Viçosa, 36570-900, Brazil

Journal of Animal Science (2015) 93(7): 3313-3321

Abstract

Genomic selection is applied in dairy cattle breeding to improve the genetic progress of purebred (PB) animals, whereas in pigs and poultry the target is a crossbred (CB) animal for which a different strategy appears to be needed. The source of information used to estimate the breeding values, i.e. using phenotypes of CB or PB animals, may affect the accuracy of prediction. The objective of our study was to assess the direct genomic value (DGV) accuracy of CB and PB pigs using different sources of phenotypic information. Data used were from three populations: 2,078 Dutch Landrace-based, 2,301 Large White-based and 497 crossbreds from an F1 cross between the two lines. Two female reproduction traits were analyzed: gestation length (GLE) and total number of piglets born (TNB). Phenotypes used in the analyses originated from offspring of genotyped individuals. Phenotypes collected on CB and PB animals were analyzed as separate traits using a single-trait model. Breeding values were estimated separately for each trait in a pedigree BLUP analysis and subsequently deregressed. Deregressed EBV for each trait originated from different sources (CB or PB offspring) were used to study the accuracy of genomic prediction. Accuracy of prediction was computed as the correlation between DGV and the DEBV of the validation population. Accuracy of prediction within PB populations ranged from 0.43 to 0.62 across GLE and TNB. Accuracies to predict genetic merit of CB animals with one PB population in the training set ranged from 0.12 to 0.28, with the exception when using CB offspring phenotype of the Dutch Landrace which resulted in an accuracy estimate around 0 for both traits. Accuracies to predict genetic merit of CB animals with both parental PB populations in the training set ranged from 0.17 to 0.30. We conclude that prediction within population and trait had good predictive ability regardless of the trait being the PB or CB performance, whereas using PB population(s) to predict genetic merit of CB animals had zero to moderate predictive ability. We observed that the DGV accuracy of CB animals when training on PB data was greater or equal than training on CB data. However, when results are corrected for the different levels of reliabilities in the PB and CB training data, we showed that training on CB data does outperform using PB data for the prediction of CB genetic merit, indicating that more CB animals should be phenotyped to increase the reliability and, consequently, accuracy of DGV for CB genetic merit.

Key words: genomic selection, pig, prediction, reproduction traits, within-population

6.1 Introduction

Genomic selection (Meuwissen *et al.* 2001) is applied in dairy cattle breeding with the aim to increase genetic progress of purebred (PB) animals (Berry *et al.* 2009; Hayes *et al.* 2009b; VanRaden *et al.* 2009). In pigs and poultry, however, the target is a crossbred (CB) animal. In pig breeding, multiple lines are usually used, with a minimum of two lines (normally for crossbred sows) and often additional sire lines to produce a three- or four-way cross finisher pig (e.g. Lutaaya *et al.*, 2001).

Selection on direct genomic values (DGV) for PB based on CB training data is expected to increase the genetic gain observed in CB animals over the use of PB training data (Dekkers 2007). Simulation studies (Ibánez-Escriche *et al.* 2009; Toosi *et al.* 2010; Zeng *et al.* 2013) have shown that this method of selection can result in high genetic gains and accuracies of prediction. Empirical studies that use genomic data to quantify these accuracies, however, are lacking. Moghaddar *et al.* (2014) studied accuracy of prediction when only CB sheep were used in the training set for three traits. They found accuracies ranging from 0.05 to 0.41 for PB validation populations. Hidalgo *et al.* (2015) studied prediction of CB genetic merit from a PB training set in pigs and showed that the accuracies are trait dependent (ranging from 0 to 0.31).

In many studies (Hayes *et al.* 2009a; Ding *et al.* 2013; Badke *et al.* 2014), (deregressed) estimated breeding values are used as the response variable for genomic selection. In genomic prediction, the source of information, i.e. using only the phenotypes of PB or CB animals, may influence the DGV accuracy. Therefore, the objective of our study was to assess the DGV accuracy in CB and PB pigs using different sources of phenotypic information for the estimation of the breeding values that are deregressed and used as response variable in the training data.

6.2 Material & Methods

The flowchart (Fig. 6.1) provides an overview of the study to clarify the steps that were taken in the genomic predictions.

6.2.1 Data

Genotypes were available from animals of three pig populations born between 2005 and 2012. Populations consisted of 2,078 Dutch Landrace-based (DL), 2,301 Large White-based (LW) and 497 crossbred individuals from an F1 cross between these two commercial lines.

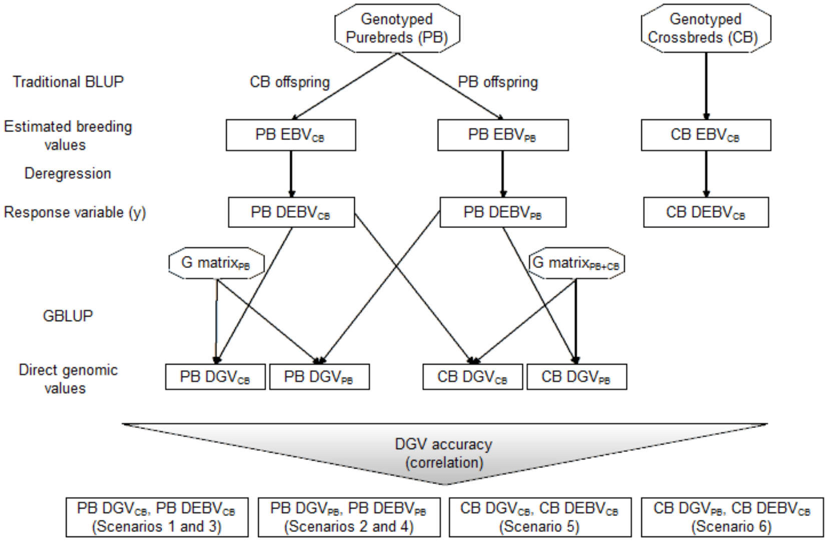

Figure 6.1 Flowchart of the steps taken in the current study. We used traditional BLUP to estimate the breeding values for purebred (PB) genotyped animals: one based on crossbred progeny and one based on the purebred progeny. These EBV based on crossbred and purebred offspring were deregressed. The DEBV were used as response variables (y) in the training data using a GBLUP model. The genomic relationship (G) matrix used in the GBLUP model varied according to the type of prediction (e.g. within-population, prediction PB to CB). DGV accuracies were computed as the correlation between DGV and DEBV from the different sources of information. The purebreds animals used in the study were from Dutch Landrace and Large White populations.

All animals were genotyped using the Illumina PorcineSNP60 BeadChip (Ramos *et al.* 2009). SNP with GenCall<0.15, unmapped SNP and SNP located on either the X or Y chromosome, according to the Sscrofa10.2 assembly of the reference genome (Groenen *et al.* 2012), were excluded. Quality control (QC) was performed within population for scenarios where the validation population was PB (scenarios 1-4), whereas QC was performed across all populations simultaneously for other scenarios (5-10). After within-population QC, 40,776 SNP for DL and 42,244 SNP for LW remained, whereas after across-populations QC, 43,027 SNP remained out of the initial 64,232 SNP. Quality control involved excluding SNP based on call rate (<0.95), strong deviations from Hardy-Weinberg equilibrium (χ^2>600) and minor allele frequency (<0.01). Animals with more than 5% of missing SNP genotypes were also removed. Missing genotypes of the remaining animals were imputed using BEAGLE 3.3.2 (Browning and Browning 2007).

We analyzed two female reproduction traits: gestation length (GLE) and total number of piglets born (TNB). GLE is the difference between insemination and farrowing dates in days, whereas TNB is the sum of all piglets born alive and stillborn. Phenotypes were available on the offspring of genotyped animals. Offspring included in the analyses were PB and CB animals that had at least one of their parents genotyped (Table 6.1). We divided the data into two sets of phenotypic records: 1) only PB offspring and 2) only CB offspring, which resulted in having four traits: GLE-PB, TNB-PB (PB traits) and GLE-CB, TNB-CB (CB traits). For the CB genotyped animals, phenotypes on their offspring were used. These data sets were used to estimate breeding values that were subsequently deregressed and used as response variable. Breeding values were estimated in a single-trait analysis using a repeatability model in ASReml 3.0 (Gilmour *et al.* 2009).

Table 6.1 Number of offspring with records used to estimate DEBV, and heritabilities for gestation length (GLE) and total number born (TNB)

Population	GLE	h^2_{GLE}	TNB	h^2_{TNB}
PB	109592	0.33	138472	0.13
CB	85875	0.31	85875	0.15

PB = Purebreds (Dutch Landrace + Large White)
CB = Crossbreds

The model used to obtain the estimated breeding values (EBV) for GLE included the fixed effects of genetic line, parity number, TNB, whether more than one insemination procedure was performed (yes or no) and heard-year-season, while the random effects consisted of service sire, a permanent environmental effect to account for the repeated observations of a single sow and an additive genetic effect (animal). The model for TNB was similar to the one used for GLE, except that the covariable TNB was replaced by interval weaning-pregnancy (days).

Deregressed EBV (DEBV) of the genotyped animals were used as response variable in this study. EBV were deregressed using the methodology proposed by Garrick *et al.* (2009). The heritabilities used for the deregression were estimated while estimating the breeding values. The reliabilities of the EBV, also required for the deregression procedure, were estimated using the following formula (Gilmour *et al.* 2009):

$$r^2 = 1 - \frac{s_i^2}{(1 + f_i)\sigma_a^2}$$

where s_i is the standard error reported for the EBV of the *i*th individual, f_i is the inbreeding coefficient, and σ_a^2 is the additive genetic variance. Further, reliabilities of the DEBV and the weighting factor (w) were also estimated following the methodology proposed by Garrick *et al.* (2009).

6.2.2 GBLUP and genomic relationship matrix

GBLUP uses a genomic relationship matrix (**G**) instead of the numerator relationship matrix (**A**). The **G** matrix contains the genomic relationship between animals and was used for prediction in all scenarios with the model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e}$$

where **y** is the vector of DEBV, μ is the overall mean, **g** is the vector of random additive genetic effects assumed to be ~N(0, $\mathbf{G}\sigma_a^2$), where **G** is the genomic relationship matrix and σ_a^2 is the additive genetic variance, **Z** is an incidence matrix, and **e** is the vector of random residual effects assumed to be ~N(0, $\mathbf{D}\sigma_e^2$), where **D** is a diagonal matrix calculated as $\mathbf{I}^*w_i$, where **I** is an identity matrix, w_i is the weight of the i^{th} DEBV based on its reliability and σ_e^2 is the residual variance. In multi-population prediction scenarios, the fixed line effect present in the model for estimating EBV was added back to the random additive-genetic effect and subsequently the line effect was again included in the genomic prediction model:

$$\mathbf{y}^* = \mathbf{1}\mu + \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$$

where **y*** is the vector of DEBV plus fixed line effect, **X** is an incidence matrix for the line effects, **b** is a vector containing the line effects, and the other factors are aforementioned.

The **G** matrix for within-population prediction was built according to VanRaden (2008), which was computed as

$$\mathbf{G} = \mathbf{ZZ}'/2\sum_{i=1}^{m} p_i q_i$$

where **Z** is a matrix of centered genotypes, $p_i = 1 - q_i$ is the allelic frequency for the i^{th} marker based on observed genotypes, and *m* is the number of makers. For multi-population prediction, the **G** matrix was built according to Chen *et al.* (2013), accounting for differences in allele frequencies between populations. Briefly, **X** is a matrix with genotype values coded as -1, 0 and 1 for the three SNP genotypes and with dimension *n* x *m* (number of animals x number of SNP). Matrix **X** includes all animals from both the training and validation sets. The matrix **X** was organized into two blocks: $[\mathbf{X}_1\ \mathbf{X}_2]'$where **X₁** represents the genotypes of line 1 and **X₂** the genotypes of line 2. **P** was a matrix of allele frequencies $[\mathbf{P}_1\ \mathbf{P}_2]'$corresponding to **X**, each row in **P₁** (or **P₂**) was a replicated row vector **p₁** (or **p₂**) with the frequency of allele *A* for SNP *k* in line 1 (or line 2). The matrix **Z** was computed to set mean values of the allele effects to 0: $[\mathbf{Z}_1\ \mathbf{Z}_2]'$=**X**-2**P**+**1** where **1** is a matrix of ones. Therefore, the two-population genomic relationship matrix was constructed as:

$$\mathbf{G} = \begin{bmatrix} \dfrac{\mathbf{Z}_1\mathbf{Z}_1'}{2\sum p_{1k}(1-p_{1k})} & \dfrac{\mathbf{Z}_1\mathbf{Z}_2'}{2\sum [p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})]^{1/2}} \\ \dfrac{\mathbf{Z}_2\mathbf{Z}_1'}{2\sum [p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})]^{1/2}} & \dfrac{\mathbf{Z}_2\mathbf{Z}_2'}{2\sum p_{2k}(1-p_{2k})} \end{bmatrix}$$

A 3-population genomic relationship matrix was constructed in a similar way as the 2-population one. The difference is that the **G** matrix was a 3x3 block matrix with the diagonal blocks accounting for the allele frequency of the populations and the off-diagonal blocks accounting for the combination of allele frequencies.

We used ASREML 3.0 (Gilmour *et al.* 2009) to predict the genomic breeding values with the **G** matrix inputted as a user defined matrix. Animals assigned to the prediction set had their DEBV removed before predicting DGV.

6.2.3 Scenarios and DGV accuracy

Ten scenarios were investigated. These can be divided into two groups according to the training and validation data sets:

- Within-population prediction (scenarios 1-4): training and validation data were subsets from the same PB population - DL or LW. The DEBV of the genotyped PB animals in the training and validation sets were estimated using the phenotypes of their offspring that were either CB or PB, i.e. prediction can be considered within-population. In scenario 1, for example, genotyped Dutch Landrace animals in the training and validation set had DEBV estimated using their CB offspring. Whereas, for scenario 2, genotyped Large White animals in the training and validation set had DEBV estimated using their PB offspring. These DEBV were used as response variable in the training set and used to compute the accuracy in the validation set. These scenarios determine how well the model fits the within-population prediction (Fig. 6.1).

- PB population(s), using different sources of phenotypic information, predicting CB genetic merit (scenarios 5-10): PB genotyped animals were used for training and CB genotyped animals were used for validation. DEBV used in the training by the PB genotyped animals were provided by their offspring phenotypes that were either PB or CB. These scenarios determine how well CB genetic merit can be predicted from DEBV of PB animals using different sources of phenotypic information and using single or both parental populations (Fig. 6.1 and 6.2).

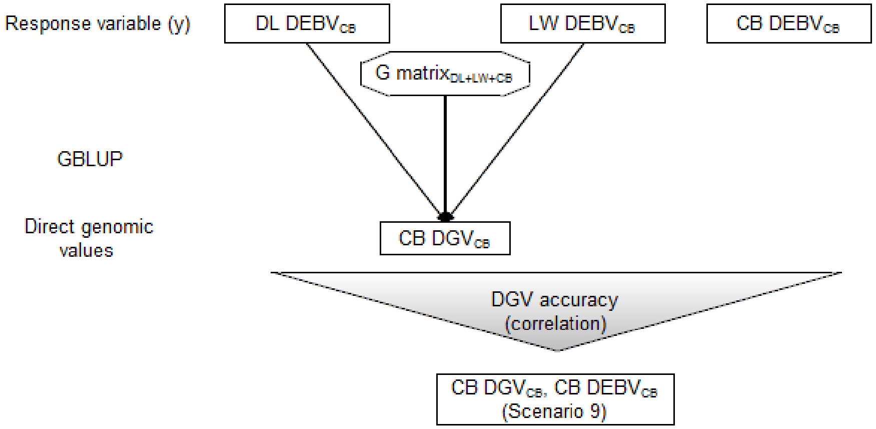

Figure 6.2 Flowchart of the steps taken for the prediction of crossbred (CB) genetic merit using both purebred (PB) parental populations in the training set. The deregressed estimated breeding values (DEBV) were computed according to Fig. 6.1. The DEBV of both parental populations were used as response variables (y) in the training data using a GBLUP model. The genomic relationship (G) matrix used in the GBLUP model included all genotyped animals from the three populations under study. The output of the GBLUP model is the DGV of the CB animals. These DGV were correlated to their DEBV and resulted in the DGV accuracies. The same steps were taken when the breeding values of both parental breeds were estimated based on their PB progeny.

The DGV accuracy was estimated as the correlation between the DGV and the DEBV of the validation set animals. Accuracies were the average of 20 random training-validation populations in scenarios 1-4. For these scenarios we randomly set aside part of the genotyped animals (N=100) and used those in a later step to determine the accuracy of prediction. These 100 animals were not included in the training for those scenarios. Scenarios 5-10 only had one estimate of accuracy to predict CB animals, and all PB animals were used in the training set. DEBV, instead of EBV, were used to compute the DGV accuracy because this removes the influence of the parents' EBV and rescales the EBV according to its accuracy, i.e., the DEBV of the animals reflect their genetic merit. It has been shown that using DEBV rather than EBV for genomic prediction yields higher DGV accuracies (Ostersen *et al.* 2011). Prediction bias was calculated regressing the validation variables (DEBV) on the prediction variables (DGV).

6.2.4 Comparison between DGV accuracies

Using different sources of information to estimate the DEBV resulted in DEBV with a range of reliabilities (Table 6.2). To correct for the difference in reliability between CB and PB estimation of EBV and enable a fair comparison, we computed expected accuracies for a situation where the mean reliability would be 0.5 (called "corrected accuracy" henceforth). We used the formula derived by Daetwyler *et al.* (2010):

$$\text{Exp.Accuracy}=\sqrt{\frac{\mathrm{N_p h^2}}{\mathrm{N_p h^2+M_e}}}$$

where *Np* is the number of individuals in the training set, h^2 is the "heritability" of the response variable (we replaced the heritability by the mean reliability (r^2) of the DEBV of a given trait of all genotyped animals), and M_e is the number of independent chromosome which accounts for the genomic structure of the population. Firstly, we inserted the observed accuracy on the left-hand side in place of the "Exp. Accuracy", and entered values for N_p and h^2 (r^2) obtained in the training set to estimate the M_e of the training set. With the estimated M_e, we could then calculate the corrected accuracy for our standardized situation of h^2 = r^2 = 0.5. The corrected accuracies of the different scenarios under standardized values of h^2 = r^2 = 0.5 could then be compared.

Table 6.2 Mean reliabilities (r^2) of deregressed breeding values and number of animals (N) for gestation length (GLE) and total number born (TNB)

	GLE		TNB	
Population	r^2	N	r^2	N
DL_{PB}	0.56	1,991	0.41	2,016
DL_{CB}	0.14	1,991	0.17	2,016
LW_{PB}	0.59	2,296	0.43	2,290
LW_{CB}	0.11	2,296	0.14	2,290
CB	0.38	497	0.37	497

DL_{PB} - Dutch Landrace animals with breeding values estimated using purebred offspring
DL_{CB} - Dutch Landrace animals with breeding values estimated using crossbred offspring
LW_{PB} - Large White animals with breeding values estimated using purebred offspring
LW_{CB} - Large White animals with breeding values estimated using crossbred offspring
CB - Cross between DL and LW animals

6.3 Results

DEBV estimated using PB or CB sources of information differed for both traits in both PB populations (e.g. Fig. 6.3).

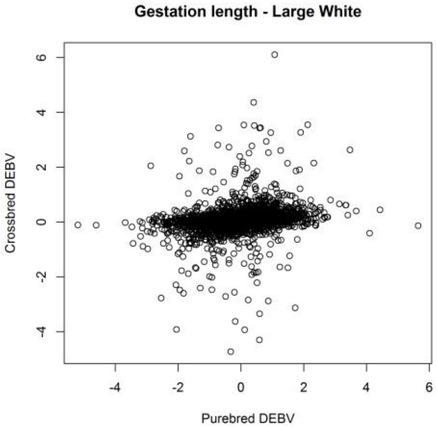

Figure 6.3 Scatter plot of deregressed breeding values. Estimation was done using different sources of information (crossbred or purebred) for gestation length in Large White population

DGV in scenarios 1-4, in which within-population and trait prediction was performed, had good predictive ability. Accuracies for GLE-CB and TNB-CB for both populations ranged from 0.43 to 0.60, whereas for GLE-PB and TNB-PB ranged from 0.43 to 0.62 (Tables 6.3 - 6.4). The regression coefficient of the DEBV on the DGV, which is a measure of bias, was less than 1 for the four traits under study (ranging from 0.43 to 0.82).

DGV accuracy in scenarios 5-10, in which genetic merit of CB animals was predicted using PB population(s) as training set, had different accuracies depending on the source of information used to estimate the breeding values. The regression coefficient of the DEBV on the DGV for all these scenarios was less than 1, whereas the standard error ranged from 0.89 to 1.80.

For scenarios 5-8, where the training set contained only one PB population, accuracies were not different from zero (P>0.05) for CB traits when the training set consisted of DL animals, whereas an accuracy of 0.17 was found for CB traits when the training set consisted of LW animals. For PB traits, this difference between the PB populations was not observed. Accuracies ranged from 0.12 to 0.28 for PB traits for both populations. The corrected accuracies for scenarios 5-8 ranged from zero to 0.35 for CB traits, whereas they ranged from 0.13 to 0.26 for PB traits.

For scenarios 9-10, which contained both parental populations together in the training set, the DGV accuracy was 0.19 for GLE-CB and 0.17 for TNB-CB, whereas it was 0.30 for GLE-PB and 0.22 for TNB-PB. The corrected accuracies for scenarios 9-10 were 0.36 for GLE-CB, 0.30 for TNB-CB, 0.28 for GLE-PB and 0.24 for TNB-PB.

6.4 Discussion

When genomically predicting CB animals, training on DEBV based on phenotypes of CB animals is expected to result in better accuracy than training on DEBV based on phenotypes of PB animals. Therefore, accuracies of predicting breeding values of CB and PB pigs were estimated for two female reproduction traits in ten scenarios aiming to determine whether using phenotypic information from CB animals to estimate the breeding values of PB training animals can improve the accuracy of predicting CB genetic merit.

Results presented are from analyses where animals in the training data were not selected based on the reliability of their DEBV. We have also analyzed all scenarios imposing a cut-off in the reliability of 0.05 for the breeding values (data not shown) and accuracies did not differ greatly from results presented. This suggests that adding this less reliable information, indeed, does not help to increase the accuracy of prediction. On the other hand, the prediction bias did become less when we removed the less reliable DEBV (mean values of the regression coefficient of the DEBV on the DGV of 0.45 when using all data and 0.55 when removing low-reliability DEBV), indicating that DEBV with low reliability affect the genomic prediction resulting in more bias. The standard error of the regression coefficients of the DEBV on the DGV were similar when using all data or when removing low-reliability DEBV.

6.4.1 Within-population prediction

The comparison of PB versus CB data differs between the two traits. For GLE, the accuracy for GLE-PB was greater than for GLE-CB in both breeds. This may be explained by the difference in reliability for DEBV estimated using PB and CB animals. The difference is large for this trait in both DL (0.42) and LW (0.48) (Table 6.2). This difference in reliability is due to the lower number of records of trait generating the CB offspring than the PB offspring (Table 6.1). Conversely, for TNB, the DGV accuracies were greater for TNB-CB than for TNB-PB. For TNB, the differences in reliabilities were smaller than for GLE, both in DL (0.24) and LW (0.29).

The accuracy of prediction for GLE-PB was greater than for TNB-PB possibly due to the greater reliability of DEBV for GLE-PB. There was no strong difference in reliability between GLE-CB and TNB-CB, but still DGV accuracies tended to be greater when DEBV reliabilities were greater.

The regression coefficient of the DEBV on the DGV varied, but was smaller than 1 for all within-population prediction scenarios. This indicates that DGV's variance was overestimated compared to the DEBV, which may occur because the animals to be genotyped were selected based on EBV (Mäntysaari *et al.* 2010).

Previously we performed within-population prediction using the same populations (DL and LW), however the DEBV used as response variable were obtained from the Topigs Norsvin routine genetic evaluation which is a blend between PB and CB records (Hidalgo *et al.* 2015). In this previous study, accuracies ranged from 0.22 to 0.72 for the four traits under study (age at first insemination, litter birth weight, total number of piglets born and litter variation). TNB was one of the studied traits and we obtained higher accuracies in the current study possibly due to the higher number of animals used in the training set.

Table 6.3 Accuracies of prediction for gestation length (GLE) using GBLUP.

	Training				Validation								
	DL		LW		DL		LW		CB	Accuracy			
Scenario	CB	PB	CB	PB	CB	PB	CB	PB	CB	Observed	r^2 =0.5	Slope*	S.E.
1	1,891	-	-	-	100	100	-	-	-	0.46	-	0.48	0.07
2	-	1,891	-	-	-	-	-	-	-	0.56	-	0.64	0.05
3	-	-	2,196	-	-	-	100	-	-	0.48	-	0.48	0.07
4	-	-	-	2,196	-	-	-	100	-	0.62	-	0.84	0.04
5	2,008	-	-	-	-	-	-	-	497	0.00	0.00	0.00	0.93
6	-	2,078	-	-	-	-	-	-	497	0.24	0.23	0.38	0.90
7	-	-	2,296	-	-	-	-	-	497	0.17	0.35	0.35	0.92
8	-	-	-	2,301	-	-	-	-	497	0.28	0.26	0.46	0.89
9	2,008	-	2,296	-	-	-	-	-	497	0.19	0.36	0.31	0.91
10	-	2,078	-	2,301	-	-	-	-	497	0.30	0.28	0.33	0.89

DL - Dutch Landrace, LW - Large White, CB - crossbred between DL and LW, PB - purebred

Accuracies of scenarios 1-4 were obtained by 20-random training-validation sets

r^2 =0.5 - Expected accuracy for a mean reliability of a training set of 0.5 (corrected accuracy)

* - Regression coefficient of the DEBV on the DGV

S.E. - Standard error of the regression coefficient. For scenarios 1-4, the values refers to the confidence interval

Table 6.4 Accuracies of prediction for total number born (TNB) using GBLUP

	Training				Validation								
	DL		LW		DL		LW		CB	Accuracy			
Scenario	CB	PB	CB	PB	CB	PB	CB	PB	CB	Observed	r^2 =0.5	Slope*	S.E.
1	1916	-	-	-	100	-	-	-	-	0.60	-	0.67	0.11
2	-	1916	-	-	-	100	-	-	-	0.49	-	0.64	0.04
3	-	-	2190	-	-	-	100	-	-	0.43	-	0.43	0.06
4	-	-	-	2190	-	-	-	100	-	0.43	-	0.82	0.07
5	2031	-	-	-	-	-	-	-	497	0.00	0.01	0.01	1.80
6	-	2077	-	-	-	-	-	-	497	0.12	0.13	0.33	1.78
7	-	-	2289	-	-	-	-	-	497	0.17	0.31	0.36	1.77
8	-	-	-	2301	-	-	-	-	497	0.17	0.18	0.53	1.77
9	2301	-	2289	-	-	-	-	-	497	0.17	0.30	0.31	1.79
10	-	2077	-	2301	-	-	-	-	497	0.22	0.24	0.56	1.77

DL - Dutch Landrace, LW - Large White, CB - crossbred between DL and LW, PB - purebred

Accuracies of scenarios 1-4 were obtained by 20-random training-validation sets

r^2 =0.5 - Expected accuracy for a mean reliability of a training set of 0.5 (corrected accuracy)

* - Regression coefficient of the DEBV on the DGV

S.E. - Standard error of the regression coefficient. For scenarios 1-4, the values refers to the confidence interval

6.4.2 PB population(s) with different sources of information predicting CB genetic merit

Different breeds and sources of phenotypic information were analyzed in scenarios 5 to 10 for their effect on the accuracy of predicting CB genetic merit. PB traits resulted in greater or equal accuracies than CB traits. One could imagine that using the breeding values estimated based on only CB phenotypes would result in greater accuracies to predict a CB population because the information comes from CB offspring, i.e. it is the same trait. However, there is generally less data available from CB individuals than from PB, resulting in a lower reliability of EBV when they are estimated using CB offspring. This is why greater accuracies were found when PB offspring were used.

The corrected accuracies for CB traits were greater than for PB traits for LW (scenarios 7-8). For DL we were surprised to find DGV accuracies of zero for CB traits (scenario 5), that also resulted in a corrected accuracy of zero. Hence, for DL the CB traits did not outperform PB traits in prediction of CB genetic merit. To explain the greater accuracies found between LW and CB than DL and CB, we set out to compare relatedness between populations. Based on the pedigree relationship, the LW population is more related (0.014) to the CB animals than to the DL population (0.009), which explains this difference in accuracies. It is unlikely, however, that this lower relatedness is the only reason leading to the very low accuracy (zero) between DL and CB for CB traits, as there is still a degree of relationship between these populations. Trying to understand the value of zero found for CB traits of the DL population, we checked the accuracy of prediction for the same trait and population using a pedigree relationship matrix (**A** matrix) instead of the **G** matrix. We observed an accuracy of 0.11, which shows some predictive ability and therefore a similar or higher accuracy would be expected using GBLUP. The genotypes of the animals used in the training set for DL population are the same for both CB and PB sources of information, as we had predictive ability for DL population using DEBV estimated based on PB offspring, it indicates that the genotypes are also sound. Therefore, this result is still a surprise with no clear explanation.

Our accuracies of prediction when using a single population in the training data are in accordance with R. Veroneze (personal communication) who predicted three-way crossbred performance from PB training data and reported values of accuracy ranging from 0.25 to 0.29.

The accuracy of prediction either increased or stayed the same when adding the other parental PB population to predict CB genetic merit (scenarios 9-10). When both parental populations were combined, the increase in accuracies were greater for PB traits than CB traits which reflects the greater reliability of the PB DEBV used as response variable. When both populations had some predictive ability in a single-population prediction of the CB, there was an increase in accuracy from combining both of them in the training set. No increase was seen, as expected, when the DL population with DEBV from CB (scenario 5) was added, as the additional data did not have predictive ability in scenario 5.

In our study, there generally was a small increase in accuracies with the addition of the second parental breed, however this is not always reported. Previous studies (e.g. Harris *et al.*, 2008) showed the opposite, where using both parental breeds to predict the crossbreds did not increase the accuracy of prediction.

The corrected accuracy, for scenarios 5-10, was greater for CB traits than PB traits; except for predicting the CB animals when using only DL (scenario 5). We calculated the genetic correlation between PB and CB performance which was 0.94 for GLE and 0.90 for TNB, i.e. smaller than 1, which explains why a greater accuracy can be expected when predicting based on CB offspring. The genetic correlation between PB and CB performance was high but still, when comparing after correction for DEBV reliability, the CB traits outperformed PB traits. It means that for other important traits, where the genetic correlation between PB and CB performance is lower, the expected advantage of using CB traits for training will be even higher. In light of these results, it is suggested that phenotyping more CB animals is desired to increase DEBV reliability and consequently yield greater DGV accuracies.

Simulation studies have shown that predicting SNP effects using CB information to select PB results in greater crossbred performance (Dekkers 2007; Ibánez-Escriche *et al.* 2009; Toosi *et al.* 2010; Zeng *et al.* 2013). Determining the accuracy of such scenarios in real data is desired, however, a large number of genotyped crossbreds is required. Currently, pig breeding companies may not collect this information because it requires genotyping CB individuals as well as additional identification and individual recording of target traits. Therefore, in order to assess the value of such data when available, we calculated predictive ability (accuracy), in real data, in CB pigs using PB genomic and phenotypic data which gives an idea of the accuracy of genomic prediction between CB and PB populations using real data within the limitations of currently available datasets. Nevertheless, studies determining the accuracy of genomic selection of PB for CB performance and studies considering non-additive effects (e.g. dominance) are still needed.

6.5 Conclusions

DGV accuracies were equal or higher when training data were based on phenotypes of PB offspring compared to phenotypes of CB offspring. The main reason for the superiority of PB data was the greater reliability of the DEBV based on phenotypes of PB offspring. When corrected to the same mean reliability, however, prediction using CB traits outperforms the usage of PB traits, indicating that more CB animals should be phenotyped to increase DEBV reliability and, consequently, accuracy of DGV for CB genetic merit.

6.6 Acknowledgements

AMH currently benefits from a joint grant from the European Commission (within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”), Breed4Food (a public-private partnership in the domain of animal breeding and genomics) and Topigs Norsvin. JWMB was supported by PPP Breed4Food. Topigs Norsvin had a role in study design, data collection, and preparation of the manuscript.

References

Badke YM, Bates RO, Ernst CW, et al (2014) Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3 Genes, Genomes, Genet 4:623–31.

Berry DP, Kearney F, Harris BL (2009) Genomic Selection in Ireland. Interbull Bull 29–34.

Browning SR, Browning BL (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 81:1084–97.

Chen L, Schenkel F, Vinsky M, et al (2013) Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. J Anim Sci 91:4669–78.

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–31.

Dekkers JCM (2007) Marker-assisted selection for commercial crossbred performance. J Anim Sci 85:2104–14.

Ding X, Zhang Z, Li X, et al (2013) Accuracy of genomic prediction for milk production traits in the Chinese Holstein population using a reference population consisting of cows. J Dairy Sci 96:5315–23.

Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 41:55.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.

Groenen MAM, Archibald AL, Uenishi H, et al (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393–8.

Harris BL, Johnson DL, Spelman RJ (2008) Genomic selection in New Zealand and the implications for national genetic evaluation. Proc. Interbull Meet. Niagara Falls, p 325–330

Hayes BJ, Bowman PJ, Chamberlain AC, et al (2009a) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41:51.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009b) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–43.

Hidalgo AM, Bastiaansen JWM, Lopes MS, et al (2015) Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3 Genes, Genomes, Genet 5:1575-1583.

Ibánez-Escriche N, Fernando RL, Toosi A, Dekkers JCM (2009) Genomic selection of purebreds for crossbred performance. Genet Sel Evol 41:12.

Lutaaya E, Misztal I, Mabry JW, et al (2001) Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. J Anim Sci 79:3002–7.

Mäntysaari E, Liu Z, VanRaden P (2010) Interbull validation test for genomic evaluations. Interbull bulletin 41:17-21.

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Moghaddar N, Swan AA, van der Werf J (2014) Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. Genet Sel Evol 46:58.

Ostersen T, Christensen OF, Henryon M, et al (2011) Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genet Sel Evol 43:38.

Ramos AM, Crooijmans RPMA, Affara NA, et al (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.

Toosi A, Fernando RL, Dekkers JCM (2010) Genomic selection in admixed and crossbred populations. J Anim Sci 88:32–46.

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–23.

VanRaden PM, Van Tassell CP, Wiggans GR, et al (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16–24.

Zeng J, Toosi A, Fernando RL, et al (2013) Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol 45:11.

7

Accuracy of genomic prediction of purebreds for crossbred performance in pigs

A.M. Hidalgo[1,2], J.W.M. Bastiaansen[1], M.S. Lopes[1,3], M.P.L. Calus[4], D.J. de Koning[2]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6708WD, the Netherlands; [2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, 750 07, Sweden; [3] Topigs Norsvin, Beuningen 6640AA, the Netherlands; [4] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, 6708WD, the Netherlands

Submitted

Abstract

In pig breeding, as the final product is a crossbred (CB) animal, the goal is to increase the CB performance. This requires different strategies for the implementation of genomic selection from what is currently implemented in e.g. dairy cattle breeding. It seems better to estimate marker effects on the basis of CB performance and subsequently use them to select purebred (PB) breeding animals. The objective of our study was to assess empirically the predictive ability (accuracy) of direct genomic values of PB for CB performance across two traits using CB and PB genomic and phenotypic data. Accuracy of prediction of PB genetic merit for CB performance based on CB training data ranged from 0.23 to 0.27 for gestation length (GLE), whereas it ranged from 0.11 to 0.22 for total number of piglets born (TNB). When based on PB training data it ranged from 0.35 to 0.55 for GLE and from 0.30 to 0.40 for TNB. Our results showed that there is predictive ability from using CB training data to predict PB genetic merit for CB performance. We also showed that using PB was better than using CB training data, mainly due to the structure of our data, which had small to moderate size of the CB training dataset, low relationship between the CB training and the PB validation populations, and a high genetic correlation (>0.90) between the studied traits in PB and CB individuals, thus favoring selection on the basis of PB data.

Key words: crossbreeding, genomic selection, reproduction traits, within-population prediction

7.1 Introduction

Genomic selection is currently primarily applied in purebred (PB) populations in dairy cattle (Hayes *et al.* 2009; VanRaden *et al.* 2009). In pig breeding and other livestock production systems, however, the final product is a crossbred (CB) animal, capitalizing on breed complementarity and heterosis. In pigs, therefore, the final goal is to increase CB performance which may require different strategies for the implementation of genomic selection from what is currently implemented in dairy cattle. Applying the dairy cattle strategy in pig breeding would put the focus on accelerating PB genetic progress, expecting an increase in CB performance as a correlated response. This strategy, however, might be suboptimal as purebred traits are often different from the crossbred traits, i.e. the genetic correlation between PB and CB performance may be considerably lower than 1 (Lutaaya *et al.* 2001; Cecchinato *et al.* 2010).

In a better strategy, marker effects are estimated on the basis of crossbred performance and subsequently used to select purebred breeding animals. This strategy is expected to give better response to selection in crossbred performance and lower rates of inbreeding in the purebred parental populations compared to within-PB-population selection according to simulation studies (Dekkers 2007; Kinghorn *et al.* 2010; van Grevenhof & van der Werf 2015). The main factors affecting these responses to selection are: i) the size of the training dataset, ii) the relationship of the selection candidates with the training dataset, iii) the genetic correlation between PB and CB performance and iv) the volume of purebred phenotypes versus crossbred phenotypes. An empirical study in sheep of CB training data predicting PB genetic merit based on a mix of CB and PB performance has shown accuracies ranging from 0.05 to 0.41 (Moghaddar *et al.* 2014). Hidalgo *et al.* (2015b) who studied prediction of crossbred pigs using purebred training data, have reported that there is predictive ability, i.e. selection on purebreds would result in response in crossbred pigs. Also, they showed the importance of having a high relationship between PB and CB animals, and that for high accuracies the genetic correlation between PB and CB performance should also be high.

Many studies have been reported for genomic prediction within PB populations (Cleveland *et al.* 2010; Badke *et al.* 2014), however, studies that empirically investigate accuracy of genomic prediction of PB based on CB performance in pigs are lacking. Therefore, the objective of our study was to empirically assess the predictive ability (accuracy) of genomic breeding values of purebred for crossbred performance across two traits (gestation length and total number of piglets born) using CB and PB genomic and phenotypic data.

7.2 Material & Methods

7.2.1 Phenotypic and genotypic data

Phenotypes and genotypes were available from sows with own-performance information of three pig populations: 1,668 Dutch Landrace-based (DL) sows, 2,003 Large White-based (LW) sows and 914 crossbred (CB) animals from an F1 cross between these two purebred lines. The CB animals were roughly 50% from DL sires mated to LW dams and 50% LW sires mated to DL dams. All sows were born from 2005 through 2013. The genotyped CB animals had no fixed family structure and the majority of them were not offspring of the genotyped PB animals, i.e. there was a distance in generations. The 914 CB animals originated from 184 sires and 487 dams. Two female reproduction traits were analyzed: gestation length (GLE) and total number of piglets born (TNB). GLE is the number of days between dates of insemination and farrowing, and TNB is the sum of all piglets born alive and stillborn. Phenotypic information of both traits for parities 2 to 7 was analyzed. First parity information was excluded because genetic correlations between first and later parities were much smaller than 1, i.e. they can be considered different traits (Hanenberg *et al.* 2001). Phenotypes were pre-corrected by subtracting the fixed effects of litter type (purebred or crossbred offspring), whether more than one insemination procedure was performed (yes or no), herd-year-season and for the covariates parity number and TNB (only for GLE). We corrected for service sire as a random effect. Descriptive statistics on the studied traits are in Table 7.1.

Table 7.1 Descriptive statistics of gestation length (GLE) and total number of piglets born (TNB) for purebred and crossbred animals.

	GLE			TNB		
Population	Animals	Records	Mean (sd)	Animals	Records	Mean (sd)
DL	1,615	6,136	116.32 (1.59)	1,668	6,536	15.41 (3.42)
LW	1,904	7,812	115.24 (1.55)	2,003	8,258	16.43 (3.53)
CB	550	2,071	115.38 (1.50)	914	3,598	15.54 (3.43)

DL - Dutch Landrace, LW - Large White, CB - F1 cross between DL and LW
sd - standard deviation

For use as a pseudo-phenotype in the validation set, we estimated the breeding values (EBV) of the PB animals based on the phenotypes of their CB progeny while excluding phenotypes from animals that belonged to the training set. Heritabilities and breeding values were estimated in a single-trait analysis using a repeatability model in ASReml 3.0 (Gilmour *et al.* 2009). These EBVs were deregressed using the methodology proposed by Garrick *et al.* (2009). The reliabilities of the EBV, also required for the deregression procedure, were estimated using the following formula (Gilmour *et al.* 2009):

$$r^2 = 1 - \frac{s_i^2}{(1 + f_i)\sigma_a^2}$$

where s_i is the standard error reported for the EBV of the i^{th} individual, f_i is the inbreeding coefficient, and σ_a^2 is the additive genetic variance. Further, reliabilities of the deregressed breeding values (DEBV) and the weighting factor (w) were also estimated following the methodology proposed by Garrick *et al.* (2009). The formula used to estimate the *w* is:

$$w_i = \frac{1 - h^2}{[c + (1 - r_i^2)/r_i^2]h^2}$$

where w_i is the weighting factor of the i^{th} individual, h^2 is the heritability of the trait, *c* is a scalar assumed to be 0.5, as suggested by Garrick *et al.* (2009), that indicates how much of the genetic variation is not accounted for by the markers, and r_i^2 is the reliability of the DEBV of the i^{th} individual.

7.2.2 Scenarios

We investigated seven scenarios which can be divided into three groups according to the structure of the training and validation data sets (Table 7.2):

- Scenarios 1-3: training and validation data were subsets from the same population, DL, LW and CB respectively, i.e. prediction was within-population. These scenarios serve as references to evaluate whether the genomic data structure can provide reasonable accuracies for “standard” within-population genomic prediction. The validation population contained the 20% youngest genotyped animals of a given population. Selecting the youngest animals as the validation population mimics a typical breeding program where direct genomic values (DGV) are computed for the youngest animals to inform selection decisions.
- Scenarios 4-5: CB phenotypes were used for training and DEBV of PB animals were used for validation. These scenarios determine how well the genetic merit

of PB animals for CB performance can be predicted from CB training data alone. The validation population consisted of the 20% of animals with the most reliable DEBVs of a given PB population. As the DEBV was based on CB progeny only, selection of the 20% of animals with the most reliable DEBVs resulted in a validation set with a reasonable level of DEBV reliability (ranging from 0.23 – 0.64). The distribution of DEBVs of the validation animals did not show any deviations from the distribution of DEBVs of all animals.

- Scenarios 6-7: PB phenotypes were used for training and the DEBV of PB animals for CB performance were used for validation. These scenarios determine how well genetic merit of PB animals for CB performance can be predicted from PB training data alone. This allowed for a comparison between accuracies when using PB or CB data as training set to predict genetic merit of PB animals for CB performance. The validation animals used in these scenarios are the same that were used for scenarios 4-5. Animals that belonged to the validation population were excluded from the training population.

Table 7.2 Description of the scenarios under study.

	Training			Validation				
Sce.	DL	LW	CB	DL	LW	CB	Training phenotypes	Validation phenotypes
1	X			X			Pre-cor. own phenotype	Pre-cor. own phenotype
2		X			X		Pre-cor. own phenotype	Pre-cor. own phenotype
3			X			X	Pre-cor. own phenotype	Pre-cor. own phenotype
4			X	X			Pre-cor. own phenotype	DEBV for CB performance
5			X		X		Pre-cor. own phenotype	DEBV for CB performance
6	X			X			Pre-cor. own phenotype	DEBV for CB performance
7		X			X		Pre-cor. own phenotype	DEBV for CB performance

Sce. - Scenario, DL - Dutch Landrace, LW - Large White, CB - F1 cross between DL and LW, DEBV - Deregressed breeding value, Pre-cor. - Pre-corrected

7.2.3 Genotyping

Animals were genotyped using the Illumina PorcineSNP60 BeadChip (Ramos *et al.* 2009). All genotyped PB animals were breeding animals from nucleus farms that were selected for genotyping because they had phenotypic records from multiple parities on multiple traits. Crossbred animals belonged to farms where phenotypic data from PB and CB animals were recorded to be included in genetic evaluations. Single nucleotide polymorphisms (SNP) with GenCall <0.15, unmapped SNPs and SNPs located on either the X or Y chromosome, according to the Sscrofa10.2 assembly of the reference genome (Groenen *et al.* 2012), were excluded. Quality control (QC) was performed within populations for scenarios where within-population prediction was performed (scenarios 1-3 and 6-7), whereas QC was performed across all populations simultaneously for other scenarios (scenarios 4-5). After within-population QC, 42,360 SNPs for DL, 41,005 SNPs for LW and 43,265 SNPs for CB remained, whereas after across-populations QC, 38,201 SNPs remained out of the initial 64,232 SNPs. QC involved excluding SNPs with call rate <0.95, minor allele frequency <0.01 and strong deviations of Hardy-Weinberg equilibrium (χ^2>600). Individuals with missing genotype frequency >0.05 were also removed. Missing genotypes of the remaining animals were imputed using BEAGLE 3.3.2 (Browning and Browning 2007).

7.2.4 Statistical analyses

Direct genomic values were computed based on the genomic best linear unbiased prediction method (GBLUP) to assess the prediction accuracy in all scenarios applying the following model:

$$\mathbf{y} = \mu + \mathbf{Zg} + \mathbf{Wu} + \mathbf{e},$$

where **y** is a vector of corrected phenotypes, μ is the overall mean, **g** is the vector of random-additive genetic effect assumed to be ~$N(\mathbf{0}, \mathbf{G}\sigma_a^2)$, where **G** is the genomic relationship matrix and σ_a^2 is the additive genetic variance, **u** is the vector of random effect of permanent environment assumed to be ~$N(\mathbf{0}, \mathbf{P}\sigma_p^2)$, where **P** is a diagonal matrix with the number of observations per sow on the diagonal and σ_p^2 is the permanent environmental variance, **Z** and **W** are incidence matrices, and **e** is the vector of random residual effect assumed to be ~$N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where **I** is an identity matrix and σ_e^2 is the residual variance. The random permanent environmental effect was included in the model to account for the repeated observations of a single sow.

The **G** matrix for scenarios 1, 2, 3, 6 and 7 was built according to VanRaden (2008), which was computed as $\mathbf{G} = \mathbf{ZZ}' / 2\sum_{i=1}^{m} p_i q_i$, where **Z** is a matrix of centered genotypes and p_i = 1 - q_i is the allelic frequency for the i^{th} marker based on observed genotypes, and *m* is the number of makers. For scenarios 4 and 5, the **G** matrix was built according to Chen *et al.* (2013) which accounts for differences in allele frequencies between populations. In short, **X** is a matrix with genotype values coded as -1, 0 and 1 for the three SNP genotypes and with dimension *n* x *m* (number of animals x number of SNPs). Matrix **X** includes all animals from both the training and validation sets. The matrix **X** was organized into two blocks: $[\mathbf{X}_1\ \mathbf{X}_2]'$ where $\mathbf{X_1}$ represents the genotypes of line 1 and $\mathbf{X_2}$ the genotypes of line 2. **P** was a matrix of allele frequencies

$[\mathbf{P}_1\ \mathbf{P}_2]'$ corresponding to **X**, each row in **P₁** (or **P₂**) was a replicated row vector **p₁** (or **p₂**) with the frequency of allele *A* for SNP *k* in line 1 (or line 2). The matrix **Z** was computed to set mean values of the allele effects to 0: $[\mathbf{Z}_1\ \mathbf{Z}_2]'$**=X-**2**P+1** where **1** is a matrix of ones. Therefore, the two-population genomic relationship matrix was constructed as:

$$\mathbf{G}=\begin{bmatrix} \dfrac{\mathbf{Z_1 Z_1'}}{2\sum p_{1k}(1-p_{1k})} & \dfrac{\mathbf{Z_1 Z_2'}}{2\sum [p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})]^{1/2}} \\ \dfrac{\mathbf{Z_2 Z_1'}}{2\sum [p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})]^{1/2}} & \dfrac{\mathbf{Z_2 Z_2'}}{2\sum p_{2k}(1-p_{2k})} \end{bmatrix}$$

We used ASREML 3.0 (Gilmour *et al.* 2009) to predict the DGVs with **G** supplied as a user defined matrix. Animals assigned to the validation set had their corrected phenotypes removed before predicting DGV.

We compared our realized DGV accuracy to the predicted accuracy for within-population prediction according to the formula derived by Daetwyler *et al.* (2010):

$$\text{Predicted accuracy} = \sqrt{\frac{N_p h^2}{N_p h^2 + M_e}}$$

where N_p is the number of individuals in the training population, h^2 is the heritability of the trait, M_e accounts for the genomic structure of the population given by the number of independent chromosome segments (M_e=2N_eL/log(4N_eL)). The effective population size (N_e) of each population was calculated according to the method of Gutiérrez *et al.* (Gutiérrez *et al.* 2009), using the software Relax2 (Strandén and Vuori 2006). The genome length (*L*) was set to 19.54 Morgans, an average of four genetic maps studied by Tortereau *et al.* (2012). Wientjes *et al.* (2013) have shown that the formula used to compute M_e based on N_e is appropriate and similar to the one based on the genomic and additive genetic relationship matrix derived by Goddard *et al.* (2011).

The DGV accuracy was computed as the correlation between the DGV and the corrected phenotype/DEBV of the validation set animals divided by the square root of the heritability/reliability of the validation population. Prediction bias was calculated by regressing the validation variables (corrected phenotypes/DEBV) on the prediction variables (DGV).

7.3 Results

The estimated N_e was 123 for DL, 105 for LW and 132 for CB, slightly higher than other studies that found N_e of 91 for Finnish Landrace (Uimari and Tapio 2010) and 74 for Landrace from the USA (Welsh *et al.* 2010).

7.3.1. Within-population prediction

Crossbred animals had the highest h^2 for both GLE and TNB, 0.50 and 0.19, respectively. For GLE, both pure populations had the same heritability of 0.41. Whereas for TNB, Large White had a slightly higher heritability (h^2 = 0.15) compared to Dutch Landrace (h^2 = 0.11).

Scenarios 1-3 evaluated within-population predictions for which accuracies ranged from 0.54 to 0.79 across the two traits and different training sets (Tables 7.3-7.4). For both traits, the highest accuracy was observed in the LW population. The predicted accuracies according to Daetwyler *et al.* (2010) ranged from 0.45 to 0.76. The predicted accuracies were lower than realized accuracies for CB and around the realized values for DL and LW.

The regression coefficient of the corrected GLE phenotype on the DGV was close to 1, whereas for TNB it was considerably different from 1 with values 0.78 for DL, 0.59 for LW and 1.81 for CB. A regression coefficient of the corrected phenotype on the DGV close to 1 indicates unbiasedness, whereas a value lower than 1 indicates inflation of the DGV variance, and a value greater than 1 indicates deflation of the DGV variance.

Table 7.3 Accuracies of genomic prediction for gestation length (GLE) using GBLUP.

		Training			Validation			Accuracy		
Sce.	h^2	DL	LW	CB	DL	LW	CB	Realized	Predicted	Slope*
1	0.41	1,292	-	-	323	-	-	0.64	0.71	0.81
2	0.41	-	1,523	-	-	381	-	0.71	0.76	0.90
3	0.50	-	-	440	-	-	110	0.61	0.53	1.12
4	0.43†	-	-	550	433	-	-	0.27	-	0.31
5	0.23†	-	-	550	-	523	-	0.23	-	0.20
6	0.43†	1,494	-	-	433	-	-	0.55	-	0.50
7	0.23†	-	1,627	-	-	523	-	0.35	-	0.24

Sce. - Scenario, DL - Dutch Landrace, LW - Large White, CB - F1 cross between DL and LW
h^2 - heritability of the trait for the validation population
† - mean DEBV reliability of the validation population
* - regression coefficient of the corrected phenotype/DEBV on the DGV

Table 7.4 Accuracies of genomic prediction for total number of piglets born (TNB) using GBLUP.

		Training			Validation			Accuracy		
Sce.	h^2	DL	LW	CB	DL	LW	CB	Realized	Predicted	Slope*
1	0.11	1,334	-	-	334	-	-	0.54	0.47	0.78
2	0.15	-	1,602	-	-	401	-	0.64	0.59	0.59
3	0.19	-	-	731	-	-	183	0.79	0.45	1.81
4	0.64†	-	-	914	463	-	-	0.11	-	0.26
5	0.39†	-	-	914	-	523	-	0.22	-	0.39
6	0.64†	1,575	-	-	463	-	-	0.30	-	0.69
7	0.39†	-	1,717	-	-	523	-	0.48	-	0.77

Sce. - Scenario, DL - Dutch Landrace, LW - Large White, CB - F1 cross between DL and LW

h^2 - heritability of the trait for the validation population

† - mean DEBV reliability of the validation population

* - regression coefficient of the corrected phenotype/DEBV on the DGV

7.3.2. Prediction of purebreds for crossbred performance

Dutch Landrace animals in the validation population had higher reliabilities of their DEBV, compared to Large White animals, for both traits. For GLE, reliabilities were 0.43 and 0.23, whereas for TNB they were 0.64 and 0.39.

Scenarios 4-5 predicted genetic merit of PB for CB performance based on CB phenotypes and showed low to moderate predictive ability for GLE for the two PB populations with accuracies of 0.27 for DL and 0.23 for LW (Table 7.3). Some, but limited predictive ability was also found for TNB with accuracies of 0.11 for DL and 0.22 for LW (Table 7.4).

Scenarios 6-7 predicted genetic merit of PB for CB performance based on PB phenotypes. These scenarios showed moderate to high predictive ability for GLE for the two PB populations with accuracies of 0.55 for DL and 0.35 for LW (Table 7.3). Moderate predictive ability was also found for TNB with accuracies of 0.30 for DL and 0.40 for LW (Table 7.4).

The regression coefficient of the DEBV on the DGV for scenarios 4-7 was lower than 1 for both GLE and TNB, which indicates that the variance of the DGV was overestimated.

7.4 Discussion

We studied accuracy of genomic prediction for two female reproduction traits in seven scenarios, covering i) within-population genomic prediction in PB and CB pigs, ii) prediction of genetic merit of PB animals for CB performance using a CB training population, and iii) prediction of genetic merit of PB animals for CB performance using a PB training population.

7.4.1. Within-population prediction

Within-population genomic prediction (i) was used as point of reference and showed that the available genomic data resulted in reasonable accuracies in "standard" genomic prediction. Predicted and realized accuracies for within-population prediction slightly differ between each other. However, given that we have only two values per population, predicted and realized accuracies for within-population predictions were in good agreement, with an average difference of only 0.07. The regression coefficient of the corrected phenotype on the DGV for within-population prediction was close to 1 for GLE and considerably differed from 1 for TNB. Luan *et al.* (2009) have observed a trend that the greater the heritability of the trait, the smaller the bias. This trend could also be observed in our results. In general, bias is observed in populations that have undergone selection based on pedigree and phenotypes (Vitezica *et al.* 2011). TNB has been under stronger selection than GLE, which might have led to the more biased predictions.

7.4.2. Comparison between crossbreds and purebreds predicting purebreds for crossbred performance

We evaluated two strategies to select PB animals for CB performance: i) using CB training data and ii) using PB training data. In simulation studies, strategy (i) results in greater response to selection and lower rates of inbreeding (Dekkers 2007; Van Grevenhof and Van Der Werf 2015). This is expected, especially, in the case of having a considerable amount of CB training data, high relationship of the selection candidates with the training dataset, low genetic correlation between PB and CB performance and without major difference in data on PB phenotypes than CB phenotypes. Evaluating our results, however, we find the opposite results when comparing accuracies from scenarios 4-5 to 6-7. Strategy (ii) resulted in higher accuracy than strategy (i). The main reason for this counter-intuitive result lies in the structure of our data which had a small to moderate amount of genotyped CB animals in the training population, distant relationship between the CB training and the PB validation, a high genetic correlation between PB and CB performance (~0.90) and the greater amount of PB versus CB phenotypes.

The number of animals used in the training dataset for scenarios 6-7 was greater than for scenarios 4-5. To enable a fair comparison between those scenarios, we have randomly selected PB animals in the training data to match the number of animals used in the training of scenarios 4-5 (Table 7.5). These predictions were repeated 20 times, generating 20-random training PB populations; accuracy presented was the average of the 20 realized accuracies. Comparing the results when the same training data size is used, accuracy was still higher for PB training data rather than CB data. The differences, however, were not as great as when all PB animals were used.

Table 7.5 Accuracies of genomic prediction using GBLUP with a restriction in number of training animals.

		Training		Validation			
Trait	r^2	DL	LW	DL	LW	Accuracy†	Slope*
GLE	0.43	550	-	433	-	0.41	0.46
GLE	0.23	-	550	-	523	0.26	0.23
TNB	0.64	914	-	463	-	0.26	0.72
TNB	0.39	-	914	-	523	0.40	0.73

GLE – Gestation length, TNB – Total number of piglets born, DL - Dutch Landrace, LW - Large White, CB - F1 cross between DL and LW

r^2 - mean DEBV reliability of the validation population

† - estimate obtained by 20 random-training populations

* - regression coefficient of the DEBV on the DGV

The higher accuracies found with PB training data can, in part, be explained by the relationship between training and validation. The PB training population is more closely related to the validation population than the CB training population. In addition, the genetic correlation between PB and CB performance was not equal to, but close to unity, being 0.94 for GLE and 0.90 for TNB (Hidalgo *et al.* 2015a). While this high genetic correlation is not a disadvantage for using CB animals as training population, it also did not help to offset the other disadvantages our CB training populations suffered. Simulations that showed greater response for selecting PB for CB performance based on CB training rather than based on PB training assumed genetic correlation lower than 0.8 (Dekkers 2007; Van Grevenhof and Van Der Werf 2015).

To achieve high levels of accuracy, greater numbers of CB animals need to be included in the training set. The number of training animals with phenotypes should always be larger when the target is to predict CB compared to PB genetic merit for PB animals assuming a genetic correlation between PB and CB performance lower than 1, because the CB training animals always have at least half of their haplotypes originating from another population. In our dataset the genetic correlation was high. The additional number of animals needed may not be so large in this case.

The relationship between training and validation also plays an important role in the accuracy of prediction. In our study, there was a distance in generations resulting in a low number of CB animals with close relationship to the PB genotyped animals. In a scenario where there is a close relationship between CB and PB animals, greater accuracies are expected (Van Grevenhof and Van Der Werf 2015). It is also important to highlight that genomic selection for traits that are hardly measured in PB pigs, such as robustness in commercial conditions, CB phenotypic information might play a major role as PB phenotyping for such traits may be lacking or the data volume may be extremely limited.

Further studies to compare the accuracy of genomic prediction of PB genetic merit based on CB performance with a larger training set are needed, even though in the current study using a small size training set already showed that there is predictive ability. In addition to larger training sets, other more complex genomic models that include breed-specific effects of SNP alleles or dominance (Ibánez-Escriche *et al.* 2009; Zeng *et al.* 2013) can be employed. These models have been shown to outperform an additive model in specific cases, e.g. with high dominance levels or when the number of SNPs is small relative to the size of the training population. Further studies that develop more complex models and include larger datasets to predict PB genetic merit based on CB performance are needed and studies are underway (Bastiaansen *et al.* 2014). In addition, genomic prediction of traits for which the genetic correlation between PB and CB performance is much lower than 1 should clarify if CB training data can outperform PB training data to predict genetic merit of PB for CB performance. At the moment phenotypes and genotypes on CB animals are limiting such further analyses.

7.5 Conclusions

Predictive ability was observed for genomic prediction of PB genetic merit for CB performance, which is interesting for production systems that have the CB performance as their final breeding goal and also for breeding programs that have limited data in PB animals. Prediction of PB genetic merit for CB performance using PB training was more accurate than using CB training data for GLE and TNB which we expect to be specific for the current dataset with high genetic correlations between CB and PB performance for the studied traits (~0.90) and the low relationship between the CB training and the PB validation populations. These results, however, are encouraging and it seems worth the effort and cost to produce better datasets to investigate the prediction of CB performance in PB lines from CB genotyped and phenotyped animals, especially for traits with low genetic correlation between PB and CB performance and for traits which phenotypes are scarce in purebreds, e.g. disease related traits.

7.6 Acknowledgements

AMH currently benefits from a joint grant from the European Commission (within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”), Breed4Food (a public-private partnership in the domain of animal breeding and genomics) and Topigs Norsvin.

References

Badke YM, Bates RO, Ernst CW, et al (2014) Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3 Genes, Genomes, Genet 4:623–31.

Bastiaansen JWM, Bovenhuis H, Lopes MS, et al (2014) SNP effects depend on genetic and environmental context. Proc. 10th WCGALP. p 356

Browning SR, Browning BL (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 81:1084–97.

Cecchinato A, de los Campos G, Gianola D, et al (2010) The relevance of purebred information for predicting genetic merit of survival at birth of crossbred piglets. J Anim Sci 88:481–490.

Chen L, Schenkel F, Vinsky M, et al (2013) Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. J Anim Sci 91:4669–78.

Cleveland MA, Forni S, Garrick DJ, Deeb N (2010) Prediction of genomic breeding values in a commercial pig population. Proc. 9th WCGALP. p 266

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–31.

Dekkers JCM (2007) Marker-assisted selection for commercial crossbred performance. J Anim Sci 85:2104–14.

Garrick DJ, Taylor JF, Fernando RL (2009) Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol 41:55.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.

Goddard ME, Hayes BJ, Meuwissen THE (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 128:409–421.

Groenen MAM, Archibald AL, Uenishi H, et al (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393–8.

Gutiérrez JP, Cervantes I, Goyache F (2009) Improving the estimation of realized effective population sizes in farm animals. J Anim Breed Genet 126:327–32.

Hanenberg EHAT, Knol EF, Merks JWM (2001) Estimates of genetic parameters for reproduction traits at different parities in Dutch Landrace pigs. Livest Prod Sci 69:179–186.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–43.

Hidalgo AM, Bastiaansen JWM, Lopes MS, et al (2015a) Accuracy of genomic prediction using deregressed breeding values estimated from purebred and crossbred offspring phenotypes in pigs. J Anim Sci 93:3313–3321.

Hidalgo AM, Bastiaansen JWM, Lopes MS, et al (2015b) Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3 Genes, Genomes, Genet 5:1575–1583.

Ibánez-Escriche N, Fernando RL, Toosi A, Dekkers JCM (2009) Genomic selection of purebreds for crossbred performance. Genet Sel Evol 41:12.

Kinghorn BP, Hickey JM, Werf JHJ Van Der (2010) Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals. Proc. 9th WCGALP. p 36

Luan T, Woolliams JA, Lien S, et al (2009) The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. Genetics 183:1119–26.

Lutaaya E, Misztal I, Mabry JW, et al (2001) Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. J Anim Sci 79:3002–7.

Moghaddar N, Swan AA, van der Werf J (2014) Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. Genet Sel Evol 46:58.

Ramos AM, Crooijmans RPMA, Affara NA, et al (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.

Strandén I, Vuori K (2006) RelaX2: Pedigree analysis program. Proc. 8th WGCALP. p 27-30.

Tortereau F, Servin B, Frantz L, et al (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC Genomics 13:586.

Uimari P, Tapio M (2010) Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds. J Anim Sci 89:609–614.

Van Grevenhof IEM, Van Der Werf JHJ (2015) Design of reference populations for genomic selection in crossbreeding programs. Genet Sel Evol 47:1–9.

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–23.

VanRaden PM, Van Tassell CP, Wiggans GR, et al (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16–24.

Vitezica ZG, Aguilar I, Misztal I, Legarra A (2011) Bias in genomic predictions for populations under selection. Genet Res 93:357–66.

Welsh CS, Stewart TS, Schwab C, Blackburn HD (2010) Pedigree analysis of 5 swine breeds in the United States and the implications for genetic conservation. J Anim Sci 88:1610–1618.

Wientjes YCJ, Veerkamp RF, Calus MPL (2013) The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621–31.

Zeng J, Toosi A, Fernando RL, et al (2013) Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol 45:11.

8

Evaluation of genomic prediction of purebreds for crossbred performance in pigs accounting for dominance effects

A.M. Hidalgo[1,2], J. Zeng[3], R.L. Fernando[3], M.S. Lopes[1,4], J.C.M. Dekkers[3]

[1] Animal Breeding and Genomics Centre, Wageningen University, Wageningen, 6708WD, the Netherlands; [2] Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, 750 07, Sweden; [3] Department of Animal Science, Iowa State University, Ames, 50011, USA; [4] Topigs Norsvin, Beuningen 6640AA, the Netherlands

Submitted

Abstract

Background

In pig breeding, the final product is a crossbred pig capitalizing on heterosis, which has dominance as one of its main causes. Our objectives were: 1) estimate variance components for crossbred pigs using an additive or an additive plus dominance model for two traits, gestation length (GLE) and total number of piglets born (TNB), 2) test whether the dominance model was better than the additive model for prediction and for response to selection of purebreds for crossbred performance. This test was based on a training population of F1 crossbreds from a Dutch Landrace-based (DL) and a Large White-based (LW) population, using either pedigree or genotypes, and 3) to estimate the additive genetic variance in crossbred performance that is explained by each parental purebred population. We used three methods: pedigree-based BLUP (PBLUP), genomic BLUP (GBLUP) and a genomic variable selection model (BayesC). For each method, we used an additive model (PBLUP_{A}, GBLUP_{A} and BayesC_{A}) or an additive plus dominance model (PBLUP_{AD}, GBLUP_{AD} and $\text{BayesC}_{\text{AD}}$).

Results

For both traits, PBLUP_{AD} estimated no dominance variance. GBLUP_{AD} estimated dominance effects that accounted for 33% of the genetic variance of GLE but none for TNB. Accuracies to predict purebreds for crossbred performance using an additive model ranged from 0.114 to 0.235 for GLE and from 0.060 to 0.134 for TNB across populations and methods. Using the dominance model accuracies ranged from 0.114 to 0.237 for GLE and from 0.063 to 0.129 for TNB across populations and methods. For GLE in the LW population GBLUP_{AD} was 4% less accurate than GBLUP_{A}. In general $\text{BayesC}_{\text{AD}}$ resulted in greater accuracy than BayesC_{A}. GBLUP for TNB and pedigree-based analyses for both traits resulted in the same response to selection across populations. BayesC generally resulted in slightly greater selection responses than the other methods. PBLUP showed higher response to selection than GBLUP and BayesC for TNB in LW. We also estimated that genes from the LW population explained 1.25-fold and 1.75-fold more of the additive genetic variance in crossbred performance than genes from the DL population for GLE and TNB, respectively, but these differences were not statistically significant.

Conclusions

Genotype-based analyses are better at capturing dominance variance than pedigree-based analyses. The dominance model can slightly improve the accuracy of genomic prediction and response to selection over the additive model for a trait that has dominance variation. In addition, there are indications that the proportion of genetic variance in crossbred performance differs between the parental purebreds used in the cross.

Keywords

breed-specific effects, genomic selection, heterosis, non-additive effects, *Sus scrofa*

8.1 Background

In pig breeding multiple purebred (PB) sire and dam lines are used to produce a three- or four-way cross finisher pig with superior performance in economically important traits (Lutaaya *et al.* 2001; Merks and De Vries 2002). The crossbred (CB) pigs are raised on commercial farms with poorer management and biosecurity conditions than nucleus farms. This difference in conditions between commercial and nucleus farms is often reflected in the traits. The genetic correlation for the same trait between PB and CB performance is not always 1 (Lutaaya *et al.* 2001; Zumbach *et al.* 2007; Cecchinato *et al.* 2010) mainly because of genotype-by-environment interactions between commercial and nucleus farms, and non-additive effects. Therefore one strategy is to use CB animals in the training set to estimate breeding values and select PB breeding animals for CB performance (Dekkers 2007; Van Grevenhof and Van Der Werf 2015). This strategy is expected to give a higher response in CB performance than within-PB-population selection (Dekkers 2007; Ibánez-Escriche *et al.* 2009; Kinghorn *et al.* 2010; Toosi *et al.* 2010; Zeng *et al.* 2013; Van Grevenhof and Van Der Werf 2015) especially when the genetic correlation between PB and CB performance is substantially less than 1. This higher response is because using CB information in the training population considers genotype-by-environment effects and differences between PB and CB individuals. Using CB data in the training population also allows breeding for traits for which phenotypes are scarce in PB animals, because these traits cannot be evaluated in nucleus herds, such as disease traits (Ibañez-Escriche and Gonzalez-Recio 2011).

Another advantage of predicting CB performance is that heterosis needs to be accounted for in the predictions. One of the main causes of heterosis is likely to be dominance (Xiao *et al.* 1995; Falconer and Mackay 1996; Charlesworth and Willis 2009) making dominance an important component of crossbreeding programs. Estimating dominance effects using a pedigree-based analysis is difficult because it requires large amount of data on many full-sib families (Vitezica *et al.* 2013). Genomic information is better at estimating dominance than pedigree information because it uses heterozygosity of SNP genotypes (Vitezica *et al.* 2013). In addition, to maximize the performance of CB progeny, it is important to consider the frequencies of alleles in the breed that PB animals are crossed with to produce the CB progeny when estimating marker allele substitution effects and breeding values of PB animals for crossbred performance (Dekkers and Chakraborty 2004).

The breeding value of PB animals for CB performance using high-density SNP genotypes can be estimated using the dominance model by Zeng *et al.* (2013). This model estimates dominance while using the allele frequency of the other PB breed used to produce the CB progeny to calculate the allele substitution effects. Using simulation they showed that the dominance model had higher response to selection in CB performance than the additive model when the trait is affected by dominance and no retraining is done. Using CB animals to predict PB for CB performance with an additive model estimates accuracies ranging from 0.11 to 0.27 (Hidalgo *et al.*, personal communication). However, no study has been reported that predicts PB for CB performance accounting for dominance. In addition, most pedigree-based studies on CB performance assume the contribution of the parental breeds to the additive genetic variance in CB performance is equal to the proportion of genes contributed by each breed to the cross. However, this equality may not occur because of parental imprinting and maternal effects, among others. Therefore our objectives were 1) to estimate components of variance for CB animals using an additive, and an additive plus dominance models for two female reproduction traits in pigs; 2) to assess empirically the performance of the dominance model compared with the additive model for prediction and for response to selection of PB for CB performance for the same traits using a CB training population; and 3) to evaluate the contribution of each PB population to the additive variance in CB performance.

8.2 Material & Methods

8.2.1 Data

This experiment followed the regulations of the Netherlands law for the protection of animals. Phenotype and SNP genotype data were available on pigs from three populations: 402 Dutch Landrace-based purebreds (DL), 288 Large White-based purebreds (LW) and 914 CB pigs from an F1 cross between these two PB populations. The CB pigs were roughly 50% DL sires/LW dams and 50% LW sires/DL dams. The PB animals were breeding animals from nucleus farms. The CB animals were managed on five farms that recorded phenotypic data on PB and CB animals to be included in genetic evaluation. The pigs from the three populations were born between 2005 and 2013. The CB animals had no fixed family structure and most were not offspring of the PB animals analyzed in this study. Two female reproduction traits were analyzed: gestation length (GLE) and total number of piglets born (TNB). GLE is the interval between insemination and farrowing in days, and TNB is the sum of all piglets born alive and stillborn in the same litter. Phenotypes of both traits for parities 2 to 7 were analyzed. First parity records were excluded because genetic correlations between first and later parities are significantly lower than 1 (Irgang *et al.* 1994; Hanenberg *et al.* 2001; Oh *et al.* 2005) and are different traits.

The response variables used to estimate variance components and for training were pre-corrected phenotypes instead of the original phenotypes of the genotyped CB animals. Pre-corrections were obtained by fitting a single trait pedigree-based linear model using *ASReml v3.0* (Gilmour *et al.* 2009) with a larger data set that included all genotyped CB animals (914), the CB offspring of the genotyped PB animals and the CB contemporaries (76,866). Phenotypes were pre-corrected by subtracting estimates of the fixed effects of parity number, whether more than one insemination was performed (yes or no), and herd-year-season, as well as for the covariate of TNB (only for GLE). We also corrected for estimates of the random effect of service sire (estimated without pedigree). The response variable used for validation was the pre-corrected mean performance of the CB offspring of PB animals, which is described in Table 8.1. To calculate this mean, we used the CB offspring of the genotyped PB animals and the CB contemporaries but did not use the phenotypes of the 914 CB animals that belonged to the training set. From the 402 DL animals used for validation, 75 were parents of the genotyped CB individuals, whereas, from the 288 LW animals used for validation, 28 were parents of the genotyped CB individuals. The CB offspring of PB animals used to calculate the validation response variable were housed in a total of 187 farms, from which 4 were also farms housing the genotyped CB animals used for training. These overlapping numbers show that there were CB individuals used for training which had close relationships and were raised in the same environment as CB individuals used in the calculation of the mean CB offspring used for validation.

Table 8.1 Description of the crossbred data used to compute the mean performance of the crossbred offspring of purebred animals (validation response variable)

	GLE		TNB	
Description	DL	LW	DL	LW
Nr. of purebred animals	235	144	402	288
Nr. of crossbred offspring	21,426	10,162	52,685	25,250
Nr. of records of crossbred offspring	59,876	26,236	169,275	76,658
Mean nr. of offspring per purebred animal	91.2	70.6	131.1	87.7

GLE - gestation length, TNB - total number of piglets born, DL – Dutch Landrace, LW – Large White

8.2.2 Genotyping

Animals were genotyped using the Illumina PorcineSNP60 BeadChip (Ramos *et al.* 2009). We excluded single nucleotide polymorphisms (SNPs) with a GenCall <0.15, call rate <0.95, minor allele frequency <0.01 and strong deviations of Hardy-Weinberg equilibrium (χ^2>600). We also removed unmapped SNPs and SNPs located on the X and Y chromosomes based on the Sscrofa10.2 assembly of the reference genome (Groenen *et al.* 2012). This quality control was done across all populations simultaneously, leaving 38,201 of the initial 64,232 SNPs. Individuals with more than 5% missing genotypes were also removed. Missing genotypes of the remaining animals were imputed using BEAGLE 3.3.2 (Browning and Browning 2007) separately for each population.

8.2.3 Relationship matrices

For each trait we ran two analyses, one with the CB animals with the DL population and the other with the CB animals with the LW population. Therefore, all genetic relationship matrices were built using information on two populations simultaneously: the CB population and the PB population under study in the given analysis (DL or LW). Pedigree-based additive relationship matrices were computed using *ASReml v3.0* (Gilmour *et al.* 2009). Pedigree-based dominance matrices were computed according to Cocherkham (1954) using the package *Synbreed* (Wimmer *et al.* 2015) implemented in *R* (R Development Core Team 2013). Genomic additive relationship matrices (**G** matrix) were built according to VanRaden (2008) as $\mathbf{G} = \mathbf{ZZ}^{'}/2\sum \mathrm{p_i q_i}$, where **Z** is a matrix of centered genotype codes (0/1/2) and $\mathrm{p_i} = 1 - \mathrm{q_i}$ is the allele frequency for the i^{th} SNP based on observed genotypes. To account for differences in allele frequencies between PB and CB populations, we built **G** following Chen *et al.* (2013). Briefly, the **G** matrix is a 2x2 block matrix, with the diagonals derived using the allele frequencies in each of the populations and the off-diagonal blocks derived using the combination of allele frequencies between the two populations. The genomic dominance relationship matrices were computed based on Vitezica *et al.* (2013) with modifications to account for the differences in allele frequency between the CB and PB populations:

$$\mathbf{D} = \begin{bmatrix} \dfrac{\mathbf{M_{d1}M_{d1}^{'}}}{\sum\left[\mathrm{p_{1j}(1-p_{1j})}\right]^2} & \dfrac{\mathbf{M_{d1}M_{d2}^{'}}}{\sum[\mathrm{p_{1j}(1-p_{1j})p_{2j}(1-p_{2j})}]} \\ \dfrac{\mathbf{M_{d2}M_{d1}^{'}}}{\sum[\mathrm{p_{1j}(1-p_{1j})p_{2j}(1-p_{2j})}]} & \dfrac{\mathbf{M_{d2}M_{d2}^{'}}}{\sum\left[\mathrm{p_{2j}(1-p_{2j})}\right]^2} \end{bmatrix}$$

We computed the genomic **D** matrices using a matrix **P** that included all animals from both the training and validation sets. The matrix **P** was organized into two blocks: $[\mathbf{P}_1\ \mathbf{P}_2]'$ where $\mathbf{P}_1$ are the allele frequencies of the CB animals and $\mathbf{P}_2$ the allele frequencies of the PB animals. The dimensions of matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ were n x m where n was the number of animals and m was the number of SNP. For example, each row in $\mathbf{P}_1$ was a replicated row vector $\mathbf{p}_1$. The *j*th element in $\mathbf{p}_1$ was denoted as $\mathrm{p}_{1\mathrm{j}}$, was the frequency of allele *A* of the *j*th SNP for the CB population. Matrix $\mathbf{M}_{\mathrm{d}1}$ was an n x m (number of CB animals x number of SNP) matrix, with the element for the *i*th individual at the *j*th SNP calculated as:

$$\mathbf{M}_{\mathbf{d1}\mathrm{i,j}} = \begin{cases} -2\mathrm{p}_{1\mathrm{j}}^2 \\ 2\mathrm{p}_{1\mathrm{j}}\left(1 - \mathrm{p}_{1\mathrm{j}}\right) \\ -2\left(1 - \mathrm{p}_{1\mathrm{j}}\right)^2 \end{cases} \text{for genotypes} \begin{cases} (\mathrm{AA}) \\ (\mathrm{AB}) \\ (\mathrm{BB}) \end{cases}$$

The $\mathbf{M}_{\mathrm{d}1}$ matrix computed the coefficients of dominance for the CB. The $\mathbf{M}_{\mathrm{d}2}$ was computed similarly to $\mathbf{M}_{\mathrm{d}1}$, except that the $\mathrm{p}_{2\mathrm{j}}$ was used in place of $\mathrm{p}_{1\mathrm{j}}$, therefore the matrix $\mathbf{M}_{\mathrm{d}2}$ had dimensions n x m (number of PB animals x number of SNP).

8.2.4 Methods and statistical models

We used three methods to estimate breeding values (EBV) of PB for CB performance: pedigree-based best linear unbiased prediction (PBLUP), genomic best linear unbiased prediction (GBLUP) and a Bayesian variable selection method (BayesC) (Habier *et al.* 2011). For each method, we used both an additive model ($\mathrm{PBLUP_A}$, $\mathrm{GBLUP_A}$ and $\mathrm{BayesC_A}$) and an additive plus dominance model ($\mathrm{PBLUP_{AD}}$, $\mathrm{GBLUP_{AD}}$ and $\mathrm{BayesC_{AD}}$). For PBLUP and GBLUP, we used *ASReml v3.0* (Gilmour *et al.* 2009). Phenotypes of animals from the validation were removed. The additive and dominance models were also used to estimate variance components for both traits based on the genotyped CB animals. For $\mathrm{PBLUP_{AD}}$ and $\mathrm{GBLUP_{AD}}$ we supplied the pedigree-based **D** matrix ($\mathbf{D}_{\mathrm{ped}}$), and the genomic **G** and **D** matrices ($\mathbf{D}_{\mathrm{gen}}$) as a user defined matrix to *ASReml v3.0* (Gilmour *et al.* 2009).

PBLUP. The pedigree-based additive model ($\mathrm{PBLUP_A}$) used for analysis was:

$$\mathbf{y} = \mu + \mathbf{Zu} + \mathbf{Wpe} + \mathbf{e}$$

where **y** is a vector of pre-corrected phenotypes, μ is the overall mean, **u** is the vector of random additive genetic effects assumed to be ~$N(\mathbf{0}, \mathbf{A}\sigma_{\mathrm{u}}^2)$, where **A** is the numerator relationship matrix and σ_{u}^2 is the additive genetic variance, **pe** is the vector of random permanent environmental effects assumed to be ~$N(\mathbf{0}, \mathbf{PE}\sigma_{\mathrm{pe}}^2)$, where **PE** is a diagonal matrix with the number of observations per sow on the diagonal and σ^2_{pe} is the permanent environmental variance, **Z** and **W** are incidence matrices, and **e** is the vector of random residual effects assumed to be ~$N(\mathbf{0}, \mathbf{I}\sigma^2_e)$, where **I** is an identity matrix and σ^2_e is the residual variance. A random permanent environmental effect was included in the model to account for repeated observations on a sow.

The pedigree-based dominance model (PBLUP$_{AD}$) used in the analysis was:

$$\mathbf{y} = \mu + \mathbf{Z}_a\mathbf{u} + \mathbf{Z}_d\mathbf{d} + \mathbf{Wpe} + \mathbf{e}$$

where **y**, μ, **u, pe**, **W** and **e** are the same as the additive model, **d** is the vector of dominance effects assumed to be ~$N(\mathbf{0}, \mathbf{D}_{ped}\sigma^2_d)$, where $\mathbf{D}_{ped}$ is the pedigree-based dominance relationship matrix, σ^2_d is the dominance variance, and $\mathbf{Z_a}$ and $\mathbf{Z_d}$ are incidence matrices.

GBLUP. The models used for GBLUP$_A$ and GBLUP$_{AD}$ were the same as PBLUP$_A$ and PBLUP$_{AD}$, respectively, except that the variances and covariances of **u** and **d** were based on genomic information rather than pedigree. Therefore, **u** is the vector of additive genetic effects assumed to be ~$N(\mathbf{0}, \mathbf{G}\sigma^2_u)$, where **G** is the genomic additive relationship matrix and **d** is the vector of dominance effects assumed to be ~$N(\mathbf{0}, \mathbf{D}_{gen}\sigma^2_d)$, where $\mathbf{D}_{gen}$ is the genomic dominance relationship matrix.

BayesC. We used a modified version of *GenSel* (Fernando and Garrick 2008) to estimate the additive and dominance effects of SNP, which were then used to predict the EBV of the PB animals for CB performance. *GenSel* does not allow a repeatability model so the response variable used in this method was the mean of the pre-corrected phenotypes of each CB animal, in this case we corrected also for the estimate of the permanent environmental effect. The residual variance was weighted using the keyword "rinverse" of *GenSel* that attributes values for the diagonal matrix of the residual variances. The weights (w) were calculated according to Falconer and Mackay (1996):

$$w = \frac{1 + r(n - 1)}{n}$$

where r is the repeatability of the trait and n is the number of observations per animal. The repeatabilities used for GLE (0.60) and TNB (0.29) was computed as the proportion of phenotypic variance explained by genetics and permanent environmental effects based on GBLUP$_A$ using the genotyped CB animals.

For the additive model (BayesC$_A$) effects were estimated with the following model:

$$y_i = \mu + \sum_{j=1}^{m} \delta_j X_{ij} \alpha_j + e_{ij}$$

where y_i is the mean pre-corrected phenotype of the *i*th animal, μ is the overall mean, δ_j is the inclusion indicator variable of the *j*th SNP (0/1), X_{ij} is the copy number of a given allele of the *j*th SNP (0/1/2), α_j is the allele substitution effect of the *j*th SNP, and e_{ij} is the residual effect of the *i*th animal. Conditional on σ^2_α, the variance of random substitution effects common to all SNP, α_j had a mixture prior of a normal distribution and a point mass at zero:

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

We used a probability π of 0.99, i.e. 1% of the SNP were expected to have non-zero effects. A scaled inverse Chi-square distribution with degrees of freedom v_α= 4 and scale parameter S^2_α was specified as a prior for $\sigma^2_\alpha \sim v_\alpha S^2_\alpha \chi^{-2}_{v_\alpha}$. More details on the calculation of the scale parameter is in Fernando *et al.* (2007).

For the dominance model (BayesC$_{AD}$), effects were estimated with the following model:

$$y_i = \mu + \sum_{j=1}^{m} \left(\delta_{aj} X_{ij} a_j + \delta_{dj} K_{ij} d_j\right) + e_{ij}$$

where y_i, μ, X_{ij} are the same as the additive model, K_{ij} is the indicator variable for the heterozygous genotype of the *j*th SNP (0/1), a_j is the additive effect, d_j the dominance effect of the *j*th SNP, e_j is the residual, and δ_{aj} and δ_{dj} are the inclusion indicator variables of the *j*th SNP (0/1) for the a and d, respectively. Conditional on π_a (the probability that a_j is zero) and σ^2_a (the variance of a_j when it is non-zero), the prior for a_j is a mixture of a point mass at zero and a normal, similar to the additive model.

The dominance effect $\mathrm{d_j}$ has a similar mixture prior independent of $\mathrm{a_j}$, given π_d and σ_d^2, with corresponding definitions:

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

We used a probability π_d of 0.99, i.e. 1% of the SNP were expected to have non-zero dominance effects. The variance components σ_a^2 and σ_d^2 were assumed to have independent scaled inverse Chi-square distributions with their own degrees of freedom and scale parameters. The scale parameters S_d^2 and S_d^2 were computed as functions of the additive and dominance genetic variances which used the estimates from GBLUP$_{\mathrm{AD}}$. See Zeng *et al.* (2013) for additional details.

For both models the EBV of PB animals was calculated as:

$$\mathrm{EBV} = \sum_{j=1}^{m} Z_{ij}\widehat{\alpha}_j^r$$

where Z_{ij} is the marker genotype and $\widehat{\alpha}_j^r$ is the estimated allele substitution effect of the j^{th} SNP in breed *r*. For the dominance model, $\widehat{\alpha}_j^r$ was calculated based on the estimates of a and d for that SNP and the allele frequency in the other breed that generated the CB animals ($p_j^{r'}$):

$$\widehat{\alpha}_j^r = \widehat{\alpha}_j + \left(1 - 2p_j^{r'}\right)\widehat{d}_j$$

See Zeng *et al.* (2013) for further details.

8.2.5 Accuracy, bias, and response to selection

The training population consisted of genotyped CB pigs with pre-corrected phenotypes. The validation population consisted of genotyped PB pigs from either the DL or LW population, with the pre-corrected mean performance of their CB offspring (Table 8.1). The accuracy was computed as the correlation between the EBV and the mean performance of the CB offspring of the validation animals. Prediction bias was calculated by regressing the validation variables (mean performance of the CB offspring) on the predicted values (EBV). As the expectation of the regression coefficient is 0.5 because the validation variable comes from the offspring, we multiplied the regression coefficient by 2 so that a value of 1 indicates no bias. The validation animals belonged to multiple generations so we expected a genetic trend. To account for this, the EBV of the validation animals were deviated from the average EBV by year of birth.

Observed response to selection of PB for CB performance was calculated by selecting a given proportion of the validation PB animals (from 0.1 to 0.5) based on their EBV. The mean CB performance of the progeny of all validation PB animals was subtracted from the mean CB performance of the progeny of the selected animals giving the observed response to selection. We also calculated the expected response to selection (R) using the following formula:

$$R = ir\sigma_a$$

where i is the intensity of selection, r is the accuracy of the EBV and σ_a is the additive genetic standard deviation of CB performance for the trait in the parental PB population. The additive genetic standard deviation was estimated based on the contribution of each PB population to the genetic variance in CB performance based on pedigree information. This estimate is described in the next section.

8.2.6 Crossbred genetic variance explained by each parental PB population

We calculated the CB genetic variance explained by each PB population based on pedigree information by fitting the phenotypes of the genotyped CB animals to the model:

$$\mathbf{y} = \mu + \mathbf{Z}_{DL}\mathbf{u}_{DL} + \mathbf{Z}_{LW}\mathbf{u}_{LW} + \mathbf{Wpe} + \mathbf{e}$$

where **y** is a vector of pre-corrected phenotypes of the CB individuals, μ is the overall mean, $\mathbf{u}_{\mathrm{DL}}$ and $\mathbf{u}_{\mathrm{LW}}$ are vectors of random additive genetic effects for the DL and LW populations assumed to be ~$N(\mathbf{0}, \mathbf{A}_{\mathbf{pop}}\, \sigma^2_{u_{pop}})$, where $\mathbf{A}_{\mathbf{pop}}$ is the numerator relationship matrix based only on either population DL or LW, and $\sigma^2_{u_{pop}}$ is the additive genetic variance for each PB population, **pe** is the vector of random permanent environmental effects assumed to be ~$N(\mathbf{0}, \mathbf{PE}\sigma^2_{pe})$, where **PE** is a diagonal matrix with the number of observations per sow on the diagonal and σ^2_{pe} is the permanent environmental variance, $\mathbf{Z}_{DL}$, $\mathbf{Z}_{LW}$ and **W** are incidence matrices, and **e** is the vector of random residual effects assumed to be ~$N(\mathbf{0}, \mathbf{I}\sigma^2_{e})$, where **I** is an identity matrix and σ_e^2 is the residual variance. As the number of genotyped CB animals used in this analysis was relatively small. To improve accuracy of the estimate we included CB animals that were not genotyped but had phenotypes (CB_{all}) in the analysis. The CB_{all} data sets had 1,412 pigs for GLE and 2,273 pigs for TNB. To test whether the CB variances explained by each of the PB population were significantly different, we ran the same model but forcing the genetic variances to be the same and compared the two models using the Akaike's Information Criterion (AIC) (Akaike 1974), which was calculated as:

$$AIC = 2k - 2\ln(L)$$

where L is the maximized value of the likelihood function for the model, and k is the number of estimated parameters in the model. The number of estimated parameters was 4 when genetic variances for were estimated separately and 3 when they were forced to be the same. We also verified whether the difference in CB variances explained by the PB populations was not due to sire versus dam effects, e.g. due to imprinting, maternal, or sex chromosome effects. Therefore, we analyzed with the CB animals split into the reciprocal crosses: 1) DL sires and LW dams, and 2) LW sires and DL dams, applying the same model used for the calculation of the CB genetic variance explained by each parental PB population.

8.3 Results

Allele frequencies were different between the DL and LW populations with a correlation of 0.22.

8.3.1 Additive and dominance variances

Variance components for CB performance were estimated for GLE and TNB with PBLUP and GBLUP (Table 8.2) using the genotyped CB animals. Pedigree-based estimates of dominance variance were zero for both traits so $PBLUP_A$ and $PBLUP_{AD}$ had the same variance components. $GBLUP_{AD}$, dominance effects accounted for 33% of the genetic variance of GLE but no dominance variance was found for TNB. Therefore, $GBLUP_A$ and $GBLUP_{AD}$ had the same variance component estimates for TNB. For GLE, $GBLUP_{AD}$ had slightly lower estimates of additive, permanent environmental and residual variances a lower narrow-sense heritability and a higher broad-sense heritability than $GBLUP_A$.

Table 8.2 Estimates of variance components for crossbred performance using genotyped crossbred animals estimated for GLE and TNB. Variance components are for the additive (A), and additive + dominance (AD) models using pedigree and genomic BLUP for each trait.

	$PBLUP_A$		$PBLUP_{AD}$				$GBLUP_A$		$GBLUP_{AD}$			
Trait	V_A	V_P	V_A	V_D	V_G	V_P	V_A	V_P	V_A	V_D	V_G	V_P
GLE	0.85	1.69	0.85	0.00	0.85	1.69	0.89	1.79	0.70	0.35	1.06	1.76
TNB	1.72	9.92	1.72	0.00	1.72	9.93	1.91	10.13	1.91	0.00	1.91	10.12

GLE - gestation length, TNB - total number of piglets born, V_A - additive genetic variance, V_D - dominance variance, V_G - genetic variance, V_P - phenotypic variance

8.3.2 Accuracy and bias

Accuracies to predict PB for CB performance using an additive model ranged from 0.114 to 0.235 for GLE and from 0.060 to 0.134 for TNB across populations and methods (Table 8.3). Using dominance model, accuracies were very similar and ranged from 0.114 to 0.237 for GLE and from 0.063 to 0.129 for TNB across populations and methods (Table 8.3). For both traits, $PBLUP_A$ and $PBLUP_{AD}$ resulted in the same accuracies due to no dominance variance. Similarly, $GBLUP_A$ and $GBLUP_{AD}$ had no difference in accuracies for TNB. For GLE using $GBLUP_{AD}$ reduced the accuracy by 4% for the LW population compared with using $GBLUP_A$. BayesC was the method that showed the greatest differences in accuracy between the additive and dominance models for GLE in the LW population. In general, $BayesC_{AD}$ resulted in higher accuracy than $BayesC_A$, up to 5% higher for GLE and TNB in the LW population (Table 8.3). The correlation between EBV based on the different models across methods was either 1 or close to unity (Table 8.4).

Table 8.3 Accuracy and bias of genomic prediction of breeding values of purebreds for crossbred performance using different methods and models for GLE and TNB.

	Training	Validation			PBLUP		GBLUP		BayesC	
Trait	Size	Size	Pop.	Model	Acc.	Bias	Acc.	Bias	Acc.	Bias
GLE	550	235	DL	A	0.233	0.61	0.233	0.44	0.235	0.49
GLE	550	235	DL	AD	0.233	0.61	0.233	0.50	0.237	0.52
GLE	550	144	LW	A	0.114	0.35	0.164	0.37	0.173	0.43
GLE	550	144	LW	AD	0.114	0.35	0.157	0.43	0.182	0.52
TNB	914	402	DL	A	0.089	0.50	0.122	0.44	0.134	0.47
TNB	914	402	DL	AD	0.089	0.50	0.122	0.44	0.129	0.40
TNB	914	288	LW	A	0.105	0.60	0.083	0.31	0.060	0.22
TNB	914	288	LW	AD	0.105	0.60	0.083	0.31	0.063	0.21

GLE - gestation length, TNB - total number of piglets born, DL - Dutch Landrace, LW - Large White, Pop. - population, Acc. – accuracy, A - additive model, AD - additive + dominance model, Bias - regression coefficient of the corrected phenotype on the EBV

Bias ranged from 0.21 to 0.61 across traits, populations and methods (Table 8.3). The regression coefficients were always substantially lower than 1 indicating that the EBV were overestimated. Pedigree-based analyses were generally less biased than genotype-based analyses. For GLE accounting for dominance effects in the prediction model reduced bias.

Table 8.4 Correlation (standard error) between estimated breeding values (EBV) of purebred validation animals obtained using the additive and the additive plus dominance model by method, trait and population

Trait	Population	PBLUP	GBLUP	BayesC
GLE	DL	1.000 (7.87e-6)	0.986 (0.011)	0.983 (0.012)
GLE	LW	1.000 (1.05e-5)	0.993 (0.010)	0.983 (0.015)
TNB	DL	1.000 (1.2e-5)	1.000 (1.2e-5)	0.994 (0.005)
TNB	LW	1.000 (1.7e-5)	1.000 (1.7e-5)	0.996 (0.005)

GLE - gestation length, TNB - total number of piglets born, DL - Dutch Landrace, LW - Large White

8.3.3 Response to selection

PBLUP$_A$ and PBLUP$_{AD}$ had the same responses to selection across traits and populations because no dominance variance was estimated (Figs. 8.1 and 8.2). Similarly, GBLUP$_A$ and GBLUP$_{AD}$ resulted in the same responses for TNB because no dominance variance was estimated. For GLE, EBV estimated using GBLUP$_{AD}$ had lower response to selection than EBV estimated using GBLUP$_A$ when the proportion selected was 0.1. However, response to selection on EBV from GBLUP$_{AD}$ was greater than that of GBLUP$_A$ when the proportion selected was greater than 0.1. Response to selection using BayesC$_{AD}$ was higher than BayesC$_A$ for GLE in the LW population but not in the DL population. For TNB, response to selection using BayesC$_{AD}$ was slightly lower than using BayesC$_A$ for both PB populations. Responses to selection using BayesC were slightly higher than other methods across traits and populations. Response to selection using PBLUP, however, was higher than using GBLUP and BayesC for TNB in the LW population. Observed and expected responses were similar across traits, populations and methods (Figs. 8.1 and 8.2).

8.3.4 Crossbred genetic variance explained by each parental PB population

To investigate potential reasons for the severe biases in EBV we observed (Table 8.3), we estimated the CB genetic variance that is explained by each PB population based on pedigree information using an additive model. Using the genotyped CB animals, for GLE, the LW population explained 5 times as much CB variance than the DL population (Table 8.5); this difference was significant (Table 8.6). For TNB, the LW explained 1.5-fold more CB variance than the DL (Table 8.5), however, this difference was not significant (Table 8.6). When performing the same analysis but including CB animals that were not genotyped but had phenotype (CB_{all}), the estimates were more accurate. The differences of CB variance explained by each of the PB populations for GLE became smaller and not significant, whereas for TNB the difference was maintained but still not significant (Tables 8.5 and 8.6). Nevertheless, we were interested in verifying whether the numerical difference in CB variances explained by the PB populations was not due to sire versus dam effects. We performed separate analyses with the CB animals split into the reciprocal crosses and showed that the numerical difference in CB variance explained by the two PB populations was still present, indicating that it was due to breed effects and not to a sire or a dam genetic effect (Table 8.5).

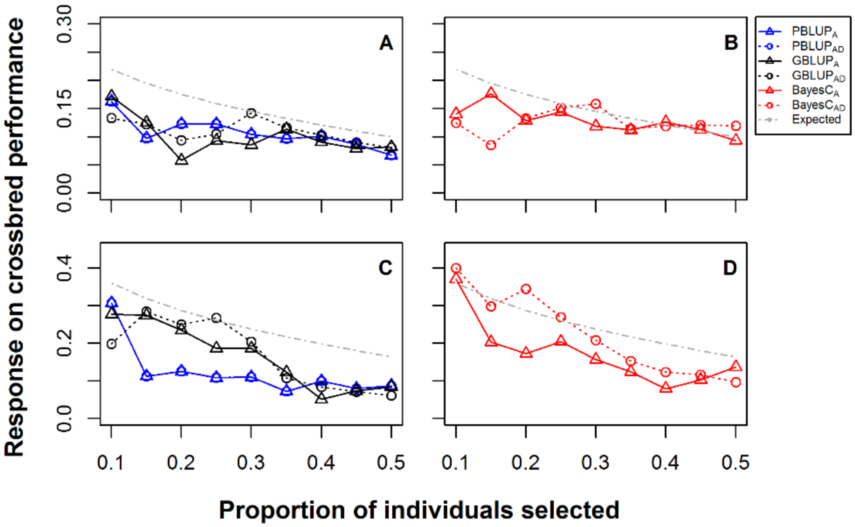

Figure 8.1 Observed and expected response to selection in crossbred performance against proportion of purebred individuals selected based on EBV for crossbred performance. Includes additive and dominance models for different estimation methods for gestation length. (A) For GBLUP and PBLUP in the Dutch Landrace, (B) for BayesC in the Dutch Landrace, (C) for GBLUP and PBLUP in the Large White and (D) for BayesC in the Large White.

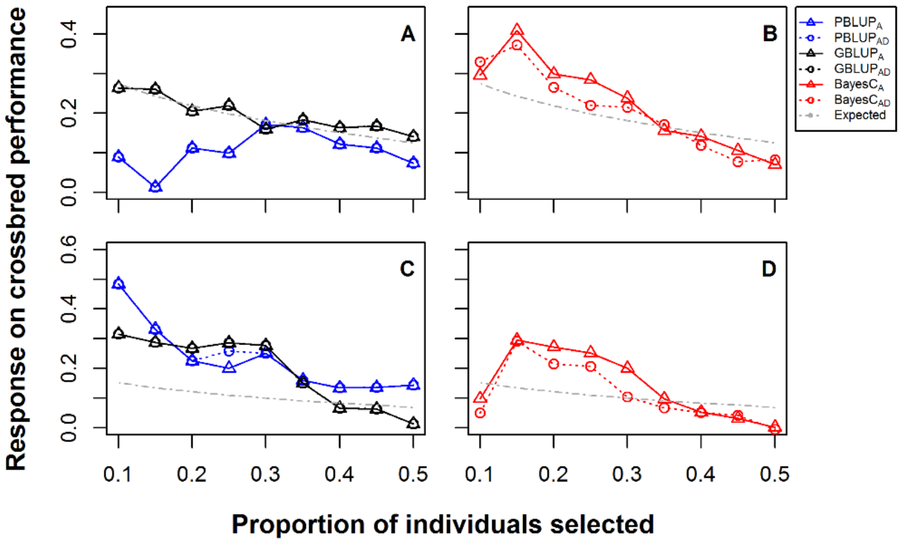

Figure 8.2 Observed and expected response to selection in crossbred performance against proportion of purebred individuals selected based on EBV for crossbred performance. Includes additive and dominance models based on different methods for total number of piglets born. (A) For GBLUP and PBLUP in the Dutch Landrace, (B) for BayesC in the Dutch Landrace, (C) for GBLUP and PBLUP in the Large White and (D) for BayesC in the Large White.

Table 8.5 Estimates of the crossbred genetic variance that is explained by each purebred population using all crossbreds or split into the reciprocal crosses.

		CB		CB_{all}		CB_{all} (DL♂ x LW♀)		CB_{all} (LW♂ x DL♀)	
Trait	Parameter	Variance (SE)	Nr. animals	Variance (SE)	Nr. animals	Variance (SE)	Nr. animals	Variance (SE)	Nr. animals
GLE	DL	0.07 (0.04)	550	0.16 (0.04)	1412	0.14 (0.06)	655	0.15 (0.06)	757
	LW	0.35 (0.08)		0.20 (0.04)		0.22 (0.06)		0.23 (0.07)	
	Perm	0.55 (0.06)		0.51 (0.04)		0.39 (0.06)		0.60 (0.07)	
	Residual	0.71 (0.03)		1.08 (0.03)		0.96 (0.04)		1.17 (0.04)	
	Phenotypic	1.68 (0.09)		1.95 (0.06)		1.71 (0.08)		2.15 (0.10)	
TNB	DL	0.34 (0.15)	914	0.32 (0.11)	2273	0.32 (0.14)	1416	0.40 (0.21)	857
	LW	0.52 (0.18)		0.56 (0.14)		0.50 (0.19)		0.62 (0.22)	
	Perm	1.81 (0.22)		2.14 (0.18)		2.21 (0.26)		2.13 (0.28)	
	Residual	7.23 (0.20)		8.37 (0.17)		8.20 (0.22)		8.53 (0.26)	
	Phenotypic	9.90 (0.27)		11.39 (0.23)		11.23 (0.29)		11.68 (0.38)	

GLE - gestation length, TNB - total number of piglets born, DL – Dutch Landrace, LW – Large White, CB – genotyped CB individuals from the current study, CB_{all} – genotyped CB animals and CB animals not genotyped but phenotyped, SE – standard error

Table 8.6 Comparison of statistical models based on Akaike's Information Criterion (AIC). Used to test the significance of the difference in estimates of crossbred variance contributed by each parental purebred population

Trait	Data set	Nr. of animals	Model	AIC
GLE	CB	550	Different variances	2,256.88
GLE		550	Same variance	2,263.74
TNB		914	Different variances	11,481.98
TNB		914	Same variance	11,480.46
GLE	CB_{all}	1,412	Different variances	5,851.32
GLE		1,412	Same variance	5,849.56
TNB		2,273	Different variances	4,096.00
TNB		2,273	Same variance	4,095.64

GLE - gestation length, TNB - total number of piglets born, CB – genotyped CB individuals from the current study, CB_{all} – genotyped CB animals and CB animals not genotyped but phenotyped

8.4 Discussion

We estimated components of variance for CB animals using an additive and an additive plus dominance model for two female reproduction traits in pigs. For the same traits, we also studied prediction of PB for CB performance. For these predictions, we assessed whether including dominance improves the accuracy of prediction and response to selection using three prediction methods. All these analyses were performed using either pedigree or high-density genotypes. Lastly, we evaluated the CB additive variance specific to each parental PB populations.

8.4.1 Dominance variance

Heterosis has been reported in many studies for litter size (Bondoc *et al.* 2001; Cassady *et al.* 2002; Bidanel 2010; Nwakpu and Onu 2011) and gestation length (Cassady *et al.* 2002; Nwakpu and Onu 2011) in pigs. Also, dominance variance has been detected for number of piglets born alive (Culbertson *et al.* 1998). Based on these reports, one could expect dominance variance for both traits that were studied here. Pedigree-based analyses, however, resulted in zero estimates of dominance variance for both traits. This can be explained by the limited size of our CB data set and the limited number of full-sibs. In our data, a total of 360 individuals for GLE and 532 for TNB had at least one full-sib included in the analyses. Estimation of dominance effects using a pedigree-based analysis is typically not performed because it is not very precise, as it requires large amounts of data on many full-sib families (Vitezica *et al.* 2013).

Using genotype data, on the other hand, we found dominance variance for GLE but not for TNB. Dominance variance can be estimated more precisely with genomic than with pedigree information, because it is possible to determine whether the assessed locus is heterozygous (Vitezica *et al.* 2013). These results demonstrate that GLE is affected by dominance, whereas TNB is a trait that may not be affected by dominance in the studied CB population. In this case, if there is heterosis for TNB in the pig populations under study, epistasis rather than dominance could be the cause of it (Hill and Mäki-Tanila 2015). In pigs, Nishio and Satoh (2014) studied two traits, which names were not revealed (T4 and T5), and found dominance to explain 24% and 15% of the genetic variance in PB lines. In PB dairy cattle, Sun *et al.* (2014) found that dominance explained around 10% of the genetic variance for a range of traits. One would expect greater dominance variance in CB than in PB populations (Nishio and Satoh 2014), so the lack of evidence of dominance variance for TNB in this study is surprising.

8.4.2 Accuracy and bias

Dominance variance was zero for pedigree-based analyses and for TNB in the genotype-based analyses (Table 8.2). Including dominance, therefore, was not expected to change accuracies of EBV of PB for CB performance for these cases and our results agreed with this expectation for PBLUP and GBLUP. For BayesC, however, dominance effects were still estimated even when dominance variance was estimated to be zero by GBLUP$_{AD}$, which decreased accuracy for TNB for the DL population and increased accuracy for the LW population (Table 8.3). The estimation of dominance effects using BayesC might have been because we used the mean of the pre-corrected phenotypes of each CB animal as input instead of the repeated observations. When we used the mean of the pre-corrected phenotypes as input we found dominance variance using BayesC$_{AD}$ but not when we used repeated observations for PBLUP$_{AD}$ or GBLUP$_{AD}$. Therefore as BayesC detected dominance variance we estimated non-zero dominance effects for a trait that has no dominance variance. Therefore it is likely that the decrease in accuracy for the DL population and the increase in accuracy for the LW population are due to chance.

We estimated that dominance accounted for a third of the genetic variance for GLE based on GBLUP (Table 8.2). Consequently we expected an increase in accuracy of prediction when including dominance in the genomic prediction model. For the DL population, however, there was no change in accuracy, while we found a slight decrease in accuracy when dominance was included for GBLUP in the LW population (Table 8.3). Ertl *et al.* (2014) observed a similar results with accuracy decreasing for

EBV of milkability in dairy cattle when dominance was included in the GBLUP. Inclusion of dominance effects, however, increased the accuracy of prediction for GLE in both populations for BayesC, as expected.

Genotype-based predictions (GBLUP and BayesC) generally had higher accuracies than pedigree-based predictions (PBLUP; Table 8.3). These higher accuracies are because genotype-based predictions use SNP genotypes which estimate the relationship coefficients among pigs better than pedigree-based prediction (Forni *et al.* 2011; Tusell *et al.* 2013; Hidalgo *et al.* 2015). BayesC had slightly higher accuracies than GBLUP. We also observed that the increase in accuracy by including dominance in the model was higher for BayesC than for GBLUP. This higher increase was possibly because BayesC accounted for the allele frequency of the mates when generating the CB pigs rather than just using the allele frequency of the same parental population when estimating the allele substitution effects (Dekkers and Chakraborty 2004; Zeng *et al.* 2013). The allele frequencies in the DL and LW populations were quite different (correlation = 0.22), showing that the allele frequency of the other parental population might be relevant in the genomic prediction. To test this, we estimated the genomic predictions from BayesC based on allele substitution effects estimated using either the same or other breed's allele frequencies. For GLE, the accuracy increased for DL when using the same breed's allele frequency (0.237 when using LW frequencies compared with 0.245 when using DL frequencies), whereas for the LW, the accuracy decreased when using the same breed's allele frequency (0.182 when using LW frequencies compared with 0.174 when using DL frequencies). For TNB, the accuracies did not differ between using the same or other breed's allele frequencies in estimation of allele substitution effects. Therefore, we could not confirm that using the other breed's allele frequency is always beneficial when predicting variances using a dominance model.

We found large dominance variance for GLE when using genomic information but the gain in accuracy of prediction was rather small (Tables 8.2 and 8.3). The correlation between EBV based on the different models across methods was either 1 or close to unity (Table 8.4). This means that there was no or only small differences in the predictive ability between the A and AD models. Also, the correlation between allele substitution effects estimates from BayesC$_A$ and BayesC$_{AD}$ was high (0.90 for GLE, 0.98 for TNB), indicating that little difference in accuracy would be expected between models. Su *et al.* (2012) also found large non-additive genetic variances and a relatively small increase in accuracy of prediction of PB Durco pigs.

Bias was the same (PBLUP and GBLUP) or slightly decreased (BayesC; Table 8.3) when dominance variance was not estimated (Table 8.2). For example, for TNB across methods and when using PBLUP for GLE. For cases where dominance variance was found (Table 8.2), GBLUP_{AD} and $\text{BayesC}_{\text{AD}}$ had less biased predictions for GLE than GBLUP_{A} and BayesC_{A} (Table 8.3). Nishio and Satoh (2014) also found regression coefficients smaller than 1 (ranging from 0.63 to 0.71) and less bias when dominance was included in the prediction model.

The validation accuracies are estimated with standard errors, therefore some of these results may be due to random chance. Also, the reported accuracies are correlations of EBV with means of pre-corrected phenotypes rather than the true breeding value.

8.4.5 Response to selection

Response to selection evaluates the effectiveness of genomic prediction which is the final goal of a breeding program. Generally, studies report accuracies as an indicator of response to selection, however, studying response to selection based on different proportions of selected individuals is interesting because it is possible to compare multiple points between models, traits and populations.

Responses to selection on EBV from the pedigree-based models and for TNB for the GBLUP method agreed with the results found for the dominance variance and accuracies of prediction for both PB populations. Since no dominance variance and no improvement in accuracy of prediction from including dominance was found for this situations, there was no improvement in response to selection (Figs. 8.1 and 8.2). For GLE, however, including dominance using GBLUP generally increased response to selection more than the additive model (Fig. 8.1). Using $\text{BayesC}_{\text{AD}}$ versus BayesC_{A} increased response to selection for GLE in the LW population (Fig. 8.1), which agrees with the results for accuracy of prediction (Table 8.3). For the DL population, however, depending on the proportion selected, either BayesC_{A} or $\text{BayesC}_{\text{AD}}$ had a higher response. For TNB, in general, using the dominance model did not increase response to selection due to the lack of dominance variance.

8.4.6 Crossbred genetic variance explained by each parental PB population

The amount of genetic variance in CB performance that was explained by each parental PB population differed between PB populations and data sets (Table 8.5). When we used the genotyped CB animals (550 for GLE and 914 for TNB), we found that the LW population explained a significantly larger portion of CB variance than the DL population for GLE (Tables 8.5 and 8.6). When we included CB animals that were not genotyped but were phenotyped (CB_{all}) and had more accurate estimates, the difference in CB variance for GLE that was explained by the two PB populations was lower and was not statistically significant (Tables 8.5 and 8.6). Nevertheless, we were interested in verifying whether the numerical difference in CB variances explained by the PB populations was not due to sire versus dam effects, e.g. due to imprinting, maternal or sex chromosome effects. The numerical difference in CB variance explained by the two PB populations was still present in the analyses using the reciprocal crosses (Table 8.5), indicating that it was due to breed effects and not to a sire or a dam genetic effect. This numerical difference, which was not confirmed using a larger data (CB_{all}), could have been due to breed-specific allele effects. SNP allele effects can differ between breeds due to, for instance, epistatic effects, differences in allele frequency or differences in linkage disequilibrium between markers and QTL between breeds (Ibánez-Escriche *et al.* 2009). Other studies (Zumbach *et al.* 2007; Dufrasne *et al.* 2013) reported different variances contributed by parental populations to CB performance. They, however, treated these as sire or dam effects instead of breed effects. These studies did not analyze using the reciprocal crosses, which makes it impossible to disentangle whether the effects are due to sire, dam or breed. Although the large data set (CB_{all}) found the non-significant differences in CB variance explained by each of the PB populations, it would be beneficial to do the same study using additional populations and traits. If allele effects are breed-specific, a genomic model that includes breed-specific effects of SNP alleles is needed (Ibánez-Escriche *et al.* 2009). With the current data, we cannot trace the origin of the CB alleles, preventing us from doing such analysis. Studies that will enable determination of the breed of origin of alleles, however, are underway (Bastiaansen *et al.* 2014).

8.5 Conclusions

Genotype-based analyses have greater ability to capture dominance variance than pedigree-based analyses. For a trait with zero dominance variance, the additive and the dominance models were similar. For a trait that has non-zero dominance variance, the dominance model can slightly improve accuracy of prediction, response to selection and reduce bias over the additive model. Our data contained individuals from many generations with no fixed family structure; using a larger and better designed data set may result in higher accuracies and response to selection for both additive and dominance models. In addition, we found indications that the contribution of parental breeds differed in the proportion of the crossbred genetic variance they explained, but these differences were not statistically significant in a larger population.

8.6 Acknowledgements

AMH currently benefits from a joint grant from the European Commission (within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG"), Breed4Food (a public-private partnership in the domain of animal breeding and genomics) and Topigs Norsvin.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19:716–723.

Bastiaansen JWM, Bovenhuis H, Lopes MS, et al (2014) SNP effects depend on genetic and environmental context. Proc. 10th WCGALP. p 356

Bidanel J (2010) Biology and genetics of reproduction. In: Rothschild MF (ed) The genetics of the pig. CABI, pp 218–233

Bondoc OL, Santiago CAT, Tec JDP (2001) Least-square analysis of published heterosis estimates in farm animals. Philipp J Vet Anim Sci 27:12–26.

Browning SR, Browning BL (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 81:1084–97.

Cassady JP, Young LD, Leymaster KA (2002) Heterosis and recombination effects on pig growth and carcass traits. J Anim Sci 80:2286–2302.

Cecchinato A, de los Campos G, Gianola D, et al (2010) The relevance of purebred information for predicting genetic merit of survival at birth of crossbred piglets. J Anim Sci 88:481–490.

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nat Rev Genet 10:783–96.

Chen L, Schenkel F, Vinsky M, et al (2013) Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. J Anim Sci 91:4669–78.

Cockerham CC (1954) An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present. Genetics 39:859–882.

Culbertson MS, Mabry JW, Misztal I, et al (1998) Estimation of dominance variance in purebred Yorkshire swine. J Anim Sci 76:448–451.

Dekkers JC, Chakraborty R (2004) Optimizing purebred selection for crossbred performance using QTL with different degrees of dominance. Genet Sel Evol 36:297–324.

Dekkers JCM (2007) Marker-assisted selection for commercial crossbred performance. J Anim Sci 85:2104–14.

Dufrasne M, Misztal I, Tsuruta S, et al (2013) Estimation of genetic parameters for birth weight, preweaning mortality, and hot carcass weight of crossbred pigs. J Anim Sci 91:5565–5571.

Ertl J, Legarra A, Vitezica ZG, et al (2014) Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet Sel Evol 46:40.

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Longman, Harlow, England

Fernando RL, Garrick DJ (2008) GenSel - User manual for a portfolio of genomic selection related analyses.

Fernando RL, Habier D, Stricker C, et al (2007) Genomic selection. Acta Agric Scand A Anim Sci 57:192–195.

Forni S, Aguilar I, Misztal I (2011) Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet Sel Evol 43:1.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.

Groenen MAM, Archibald AL, Uenishi H, et al (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393–8.

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

Hanenberg EHAT, Knol EF, Merks JWM (2001) Estimates of genetic parameters for reproduction traits at different parities in Dutch Landrace pigs. Livest Prod Sci 69:179–186.

Hidalgo AM, Bastiaansen JWM, Lopes MS, et al (2015) Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3 Genes, Genomes, Genet 5:1575–1583.

Hill WG, 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.

Ibánez-Escriche N, Fernando RL, Toosi A, Dekkers JCM (2009) Genomic selection of purebreds for crossbred performance. Genet Sel Evol 41:12.

Ibañez-Escriche N, Gonzalez-Recio O (2011) Review. Promises, pitfalls and challenges of genomic selection in breeding programs. Spanish J Agric Res 9:404–413.

Irgang R, Favero JA, Kennedy BW (1994) Genetic parameters for litter size of different parities in Duroc, Landrace, and Large White sows. J Anim Sci 72:2237–2246.

Kinghorn BP, Hickey JM, Werf JHJ Van Der (2010) Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals. Proc. 9th WCGALP. p 36

Lutaaya E, Misztal I, Mabry JW, et al (2001) Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. J Anim Sci 79:3002–7.

Merks JWM, De Vries AG (2002) New sources of information in pig breeding. Proc. 7th WCGALP. pp 3–10

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

Nwakpu P, Onu P (2011) Heterosis for litter size traits in native by two exotic inbred pig crosses. Agric Biol J North Am 2:1340–1346.

Oh SH, Lee DH, See MT (2005) Estimation of genetic parameters for reproductive traits between first and later parities in pig. Asian-Australasian J Anim Sci 19:7–12.

R Development Core Team (2013) R: A language and environment for statistical computing. http://www.R–project.org/.

Ramos AM, Crooijmans RPMA, Affara NA, et al (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.

Su G, Christensen OF, Ostersen T, et al (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 PM, Cole JB, O'Connell JR (2014) Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PLoS One 9:e103934.

Toosi A, Fernando RL, Dekkers JCM (2010) Genomic selection in admixed and crossbred populations. J Anim Sci 88:32–46.

Tusell L, Pérez-Rodriguez P, Forni S, et al (2013) Genome-enabled methods for predicting litter size in pigs : a comparison. Animal 7:1739–1749.

Van Grevenhof IEM, Van Der Werf JHJ (2015) Design of reference populations for genomic selection in crossbreeding programs. Genet Sel Evol 47:1–9.

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–23.

Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195:1223–1230.

Wimmer AV, Auinger H, Albrecht T, et al (2015) Synbreed: framework for the analysis of genomic prediction data using R.

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

Zeng J, Toosi A, Fernando RL, et al (2013) Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol 45:11.

Zumbach B, Misztal I, Tsuruta S, et al (2007) Genetic correlations between two strains of Durocs and crossbreds from differing production environments for slaughter traits. J Anim Sci 85:901–908.

9

General discussion

9.1 Introduction

With the development of high-throughput and cost-effective genotyping methods, exploiting genomic information became an indispensable approach for major breeding companies. Pig production relies on crossbreeding, hence, the use of genomic data for selection for crossbred performance needs to be carefully assessed. Implementation of genomic selection in crossbreeding schemes cannot be a simple copy of what is applied in breeding programs for purebred performance.

For the research presented in this thesis, I used genomic information from purebred and crossbred pigs. I have detected genomic regions associated with gestation length and with androstenone level by genome-wide association and fine-mapping analyses. Further, I studied potential pleiotropic effects of the androstenone level QTL on chromosome 6 on production and reproduction traits. To investigate the potential and peculiarities of applying genomic selection in a crossbreeding setting, I evaluated and showed that there is predictive ability between purebred and crossbred pigs. Consequently, genomic selection in purebred pigs will result in gains in the performance of crossbreds. In this Chapter, I discuss the relevance of my findings in a broader context. I will discuss how to integrate individual genetic markers with genomic selection, as well as different strategies for applying genomic selection in pig breeding using genotypes and phenotypes of purebred and crossbred animals.

9.2 Integrating individual genetic markers with genomic selection

For qualitative traits, DNA tests were developed, starting some 25 years ago, which allowed selection against an undesired condition or phenotype. For example, a recessive allele (HAL 1843^{TM}) in the porcine ryanodine receptor (*RYR1*) gene that causes malignant hyperthermia in stressful conditions (Fujii *et al.* 1991). When a single locus is controlling the trait, a DNA test is an effective tool for selection. The majority of the production traits in livestock, however, are continuously distributed (quantitative) because many quantitative trait loci (QTL) are controlling the trait. Due to the high number of loci affecting the trait, individual QTL only explain a proportion of the total genetic variance.

Because of the typically small effects, selection based only on individual markers was not applied in pig breeding companies. This was in contrast with the expectations that were set after the initial boom of genetic markers (Ibáñez-Escriche *et al.* 2014).

Genetic markers that explain part of the variance and are in linkage disequilibrium with a QTL, were incorporated into the genetic evaluation using customized SNP panels (Van Eenennaam *et al.* 2014). Such markers were used as complementary tool (Ibáñez-Escriche *et al.* 2014) resulting in marker-assisted BLUP (MA-BLUP) being applied by pig breeding companies. Like most QTL, the QTL regions for gestation length identified in Chapter 2 also explained a relatively small proportion of the genetic variance, 1.12% for the Dutch Landrace and 0.77% for the Large White pigs. Further, in Chapter 3, I fine-mapped a previously identified QTL region for androstenone level that also explained a small proportion of phenotypic variance, 6% in the Duroc population (Duijvesteijn *et al.* 2010). These results are concordant with the vast literature that reported 13,030 QTL for 663 traits usually with small effects (Animal QTLdb, http://www.animalgenome.org/QTLdb).

With the development of methods that allow to perform genomic prediction based on a large number of genetic markers (Meuwissen *et al.* 2001), and after the availability of commercial SNP chips, genomic selection (GS) became the center of attention for animal and plant breeders. Since then, GS has been implemented in dairy cattle (VanRaden *et al.* 2009) and it was shown to result in higher accuracies than traditional genetic evaluations (BLUP) (Hayes *et al.* 2009b). The main positive point of GS lies in its ability to capture the infinitesimal nature of the majority of economically important traits, which was exactly the main cause for the limited success of marker-assisted selection. In GS, all markers have their effects estimated without the need to know the biological meaning. All that is needed is a training population and sufficient computational power to run the genomic evaluation. The training population, which is phenotyped and genotyped, has to have sufficient size (Misztal 2011) and preferably be related to the selection candidates.

Even though only few causative mutations have been identified so far, such significant markers will continue to be identified. Further developments in genotyping technology resulted in a reduction of costs, enabling the production of commercial high-density (HD) SNP chips (e.g. Illumina Bovine HD 770k SNP chip). Therefore, with more animals genotyped, which increases the sample size, and with the genome more densely covered with markers, which leads to a smaller distance between the SNP and the causative mutation, a more precise detection of QTL is expected. Genome-wide association studies (GWAS) using HD SNP panels have been performed in cattle (e.g. Purfield *et al.* (2015)). In pigs, a HD SNP chip has been recently developed with approximately 660,000 SNP, however, GWAS with this HD SNP chip are still lacking. The ultimate level of genotypic information is the sequence data. Sequencing determines the order of all nucleotides of the DNA of a given organism. Therefore, sequence data contain the causative mutations of the trait. A GWAS using sequence data, hence, is expected to find the causative mutation (Meuwissen and Goddard 2010). There have been efforts to increase the numbers of sequenced animals (e.g. Daetwyler *et al.* (2014)), to enable GWAS with sequenced individuals. The approach that has been taken is to perform a GWAS using HD SNP chip genotype data and then focus on the identified peaks, performing a region-wise association study (RWAS) using imputed sequence data (Sahana *et al.* 2014; Wu *et al.* 2015). This method was able to refine previously detected QTL regions, however, it was not able to identify the causative mutation, mainly because of strong blocks of linkage disequilibrium. Another factor that might be hampering the identification of the causative mutation is that imputation is not 100% accurate, especially for rare variants and small reference panels.

As these significant regions on the genome are still being found and described, it is of interest to integrate the significant markers in the genomic evaluation. This integration is relevant because, while the causative mutations are not detected, these significant markers provide knowledge regarding the genetic architecture of the trait. Although the effects found are not large, they might add to the prediction accuracy and thus should be explored. Integrating these markers into the genomic evaluation would be a form of marker-assisted genomic prediction. Here, the marker genotype (0, 1 or 2) is fitted as a fixed effect in the genomic prediction model (MA-GBLUP). The outcome of this analysis is an estimate of estimated breeding value (EBV) of the animal and an estimate of the marker's allele substitution effect. After that, multiplying the estimate of the marker effect by the animal's genotype (0, 1 or 2) and adding this value to the EBV results in the animal's EBV from MA-GBLUP. MA-GBLUP offers the possibility to apply the results described in Chapters 2 and 3 to within-population genomic predictions as described in Chapters 5-7.

Before implementing MA-(G)BLUP it is important to know the effect of the QTL on all traits in the breeding goal. Hence, assessing pleiotropic effects of that marker on other traits is recommended to avoid unfavorable effects due to pleiotropy and/or due to genetic hitchhiking. Grindflek *et al.* (2011) found markers on the pig genome affecting simultaneously the levels of boar taint compounds (e.g. androstenone) and of sex hormones. Given that the androstenone markers have an unfavorable impact on sex hormones, the use of such markers for selection would be challenging. I showed in Chapter 4, however, that selection for the marker on chromosome 6 that reduces androstenone level will have no unfavorable effect on production and reproduction traits studied. Therefore, the use of that marker to reduce androstenone level in a breeding program becomes of interest.

To show whether integrating significant markers with genomic prediction is relevant, I performed a MA-GBLUP analysis using the most significant marker of each population described in Chapter 2 and the marker studied in Chapter 4. Markers were: rs81308021 for androstenone level in the Duroc, rs81366467 for gestation length in the Dutch Landrace and rs344547786 for gestation length in the Large White. Individuals from three pig populations were used: 833 Duroc, 1,615 Dutch Landrace and 1,904 Large White animals. These animals were genotyped using the Illumina PorcineSNP60 BeadChip (Ramos *et al.* 2009) and quality control was performed on the genotypes according to the methods described in Chapter 5. After quality control, 41,289 SNP for the Duroc, 42,360 SNP for the Dutch Landrace and 41,005 SNP for the Large White remained out of the initial 64,232 SNP. We analysed the data using ASReml 3.0 (Gilmour *et al.* 2009) with the model:

$$\mathbf{y} = \mu + \mathbf{b}_1\mathbf{SNP} + \mathbf{Zu} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of pre-corrected phenotypes, μ is the overall mean, $\mathbf{b}_1$ is the vector of regression coefficients of each SNP, **SNP** is the incidence vector for $\mathbf{b}_1$ with genotypic information (0, 1 and 2), **Z** is the incidence matrix for **u**, **u** is the vector of random additive genetic effects, assumed to be $\sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$, where **G** is the genomic relationship matrix, and e is the residual error, assumed to be $\sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where **I** is an identity matrix. The accuracy of prediction was estimated as the correlation between the EBV and the corrected phenotype in a set of validation animals. The validation population consisted of the 20% youngest genotyped animals of a given population. Phenotypes were corrected for fixed effects as described in Chapter 5. Prediction results of MA-GBLUP were compared to the results obtained from using the traditional genetic evaluation (BLUP), marker-assisted BLUP (MA-BLUP) and genomic evaluation (GBLUP) (Table 1).

Table 9.1 Accuracies of prediction for androstenone level (AND) and gestation length (GLE) using different methods.

Trait	Breed	$N_{training}$	$N_{validation}$	Accuracy† (Bias*)			
				BLUP	MA-BLUP	GBLUP	MA-GBLUP
AND	DU	666	167	0.39 (1.43)	0.42 (1.29)	0.43 (1.01)	0.45 (1.07)
GLE	DL	1,292	323	0.29 (0.73)	0.31 (0.79)	0.41 (0.81)	0.42 (0.81)
GLE	LW	1,523	381	0.41 (1.11)	0.41 (1.11)	0.46 (0.90)	0.46 (0.90)

DU - Duroc, DL - Dutch Landrace, LW - Large White, N - number of animals

† - Correlation between EBV and pre-corrected phenotype

* - Regression coefficient of the phenotype on the EBV

MA-GBLUP resulted in the highest accuracy for all three analyses (Table 1). In the Large White population, no difference was observed from either including or excluding the marker as fixed effect for gestation length when comparing BLUP with MA-BLUP, nor when comparing GBLUP with MA-GBLUP. This result in the Large White population is probably due to the minor allele frequency (MAF) of the most significant marker being very low (0.01) (Chapter 2), which means that the majority of the animals had the same genotype. Therefore adding the same marker effect to the EBV of the vast majority of the animals would not affect the accuracy. For androstenone level in the Duroc, and for gestation length in the Dutch Landrace, there was an increase in accuracy when the significant marker information was used. The increase in accuracy for MA-BLUP over BLUP was greater than for MA-GBLUP over GBLUP. As BLUP uses only pedigree information, fitting the most significant marker as fixed effect can differentiate animals with regard to the QTL, leading to a possible increase in accuracy. The increase in accuracy of MA-GBLUP over GBLUP was not as great because GBLUP already accounts for the significant marker in the **G** matrix. However, even when the same genotypic information is present in the **G** matrix, fitting the significant marker separately as a fixed effect still resulted in higher accuracy of prediction because the marker effect is better captured by the model. Fitting the marker as a separate fixed effect is not expected to lead to lower accuracies, even if the marker is a false-positive. In such a case, the effect estimated would be zero, accuracy would remain the same, and thus no harm would be done to the prediction. An issue will occur when trying to fit more markers as fixed effects than the number of animals. In this case, estimation problems occur because of a lack of degrees of freedom to fit all effects simultaneously by least squares (Lande and Thompson 1990). However, markers with large effects are not so common, therefore this issue is not likely to become a problem for the MA-GBLUP model. The regression coefficients of the phenotype on the EBV were in general close to 1 in all analyses included in Table 9.1, which indicates unbiased predictions. Less bias was observed for MA-BLUP than for BLUP, and for the genomic models GBLUP and MA-GBLUP compared with MA-BLUP and BLUP. These analyses were performed in purebred animals, therefore I can predict that MA-GBLUP would result in greater response to selection in the pure lines over GBLUP. In a breeding program where the goal is to select purebred animals for purebred performance, MA-GBLUP is therefore recommended for traits with known significant marker(s). To extrapolate to prediction of crossbred performance, MA-GBLUP would be beneficial for both purebred and crossbred performance when the QTL is the same for purebred and crossbred performance. If the interest is to select purebred animals for crossbred performance, as is the case in pig breeding, I would expect that using MA-GBLUP could improve accuracy of prediction as long as the marker is affecting the crossbred population.

9.3 Genomic selection in pigs

Genomic selection has been introduced in dairy cattle breeding aiming to improve performance of purebred animals (VanRaden *et al.* 2009). In pigs, however, the end product is a crossbred animal which may require different strategies for the implementation of GS from what is currently applied in dairy cattle. In pig breeding, specialized sire and dam lines are kept in the breeding stock and crossed to produce a three-way or four-way cross finisher pig (Merks and De Vries 2002).

In this thesis, I have analyzed androstenone level and reproduction traits. Reproduction traits generally have low heritability, but gestation length has moderate heritability. Genomic selection has a large added value for low-heritability traits (Muir 2007; Calus *et al.* 2008) because the accuracies of these traits are usually low as they depend on the heritability of the trait (Falconer and Mackay 1996; Muir 2007; Visscher *et al.* 2008). For production traits, which generally have higher heritabilities, traditional genetic evaluation already provides EBV with high accuracy, therefore the added value of GS is less. In addition to heritability, other factors affect the value of GS, e.g. the time at which traits are measured. GS can have a great positive impact on the accuracy of EBV for meat-quality traits, which are measured after slaughter therefore usually measured on relatives of selection candidates. Also, GS is expected to have a larger impact on sex-limited traits, traits that are difficult (expensive) to record, and on traits that are recorded late in life (Muir 2007). This positive impact occurs because the accuracy of traditional genetic evaluation is limited for these traits.

In this section, I will discuss different strategies of genomic selection in pigs and their perspectives. The use of within-, across- and multi-population predictions will be discussed, along with the use of crossbred information for genomic prediction.

9.3.1 Within-population prediction

Pig breeders have focused on the estimation of breeding values of purebred animals using data obtained also from purebred animals which are kept in nucleus farms. In other words, the selection is applied to improve purebred genetic merit with an expectation for a response in crossbreds. In Chapters 5 and 6, results of within-population genomic predictions are presented which showed considerably high accuracy of prediction. Within population, genomic prediction generally performed better than traditional genetic evaluation based on pedigree, which is also observed in other studies in pigs (e.g. Tusell *et al.* (2013)). Therefore genomic prediction, within-population, is recommended when the aim is to increase purebred performance. In practice, breeding companies currently perform within-population genomic prediction by applying the single-step approach (Misztal *et al.* 2009). This approach is preferred by breeding companies because current data sets still contain a large amount of data on phenotyped animals that are not genotyped. With the single-step approach, these records can still be used together with phenotyped and genotyped individuals to estimate the breeding values. Additionally, the pipeline for implementing the single-step approach is similar to the traditional genetic evaluation that was in use previously. The only major change is the replacement of the average numerator relationship matrix (**A** matrix) with an **H** matrix which contains the pedigree-genomic relationships (Legarra *et al.* 2009).

Once within-population genomic prediction is implemented, accounting for the genetic architecture of the trait might be relevant. Weighting the **G** matrix increases the accuracy of prediction (Zhang *et al.* 2010; Tiezzi and Maltecca 2015; Veroneze 2015). A practical problem is accounting for the genetic architecture in genomic evaluations would require a separate analysis for every single trait because a different **G** matrix would have to be built for each trait. To avoid this problem, using the MA-GBLUP methodology, described above, is a way of accounting for the markers with large effect in a multi-trait genomic evaluation without the need of constructing separate **G** matrices for each trait.

9.3.2 Across-population prediction

In pig breeding, multiple dam and sire lines are kept in the breeding stock. It is possible that a training dataset is not available for a specific line or that a design might be desired in which training data would only be produced in some of the lines. In such cases, performing across-population prediction could be a good strategy (Hayes *et al.*, 2009). Across-population prediction involves using population A as training dataset to predict population B. Studies in cattle have shown that training in one population to predict another results in accuracies close to zero (Harris *et al.*, 2008; Hayes *et al.*, 2009; Chen *et al.*, 2015). This low accuracy has been attributed to the different marker-QTL linkage disequilibrium phase across populations (De Roos *et al.* 2009). In pigs, we have also found accuracies close to zero for across-population predictions (Chapter 5). Therefore, under the current circumstances of a low number of animals, genotyped with around 60,000 SNP, I would not recommend across-population prediction. No matter what the reason for the application of across-population prediction would be, constraints in expenses or genomic breeding program design, the results are not encouraging. Instead, I would perform within-population genomic prediction for the line that has a training population and continue the pedigree-based genetic evaluation for the other line. In the future, when more animals are sequenced and possibly more causative mutations are identified, across-population prediction might yield better accuracies.

9.3.3 Multi-population prediction

An alternative to across-population prediction is to have, in the training set, some animals from the same population that will be predicted, and increase the size of the training set by combining populations A and B. The increase in accuracy from multi-population prediction is highly dependent on the relationship between the combined populations (De Roos *et al.* 2009). Many studies on multi-population prediction were performed in dairy cattle and have been reviewed by Lund *et al.* (2014). Generally, there is an increase in accuracy when the same breeds from different countries are combined, whereas this increase is minor when the breeds are only distantly related. Multi-population prediction in pigs, using Dutch Landrace and Large White animals plus the cross between these two populations was performed in Chapter 5. Results showed that adding the other population in the training set did not improve the accuracy compared with within-population prediction. The main reason for that was that the Dutch Landrace and Large White populations are only distantly related. Predicting the F1 cross using a multi-population training data set, which contained the F1 cross plus both parental populations, was advantageous over within-population prediction when genetic correlation between purebred and crossbred performance was high (>0.9). The parental populations are closely related to the F1 which appears to have a positive impact on accuracy of multi-population prediction (Chapter 5). Also, having a high genetic correlation between purebred and crossbred performance is relevant in boosting the accuracy of multi-population prediction. Thus, multi-population prediction in pig breeding can be recommended when predicting crossbred animals, given that populations are closely related and/or the genetic correlation between purebred and crossbred performance is 1 or close to unity.

9.3.4 Using crossbred information for genomic prediction

The final goal in pig breeding is to improve performance of the commercial crossbred pigs, taking advantage of heterosis and breed complementarity (Visscher *et al.* 2000). Crossbred pigs are mostly raised in farms at the commercial level which have lower management and biosecurity conditions compared with nucleus farms. This difference in conditions between commercial and nucleus farms is often reflected in the traits (Dekkers 2007). The same trait when measured in a commercial crossbred animal is not genetically the same as when it is measured in a purebred animal at a nucleus farm. This difference between the traits is reflected in genetic correlations below 1.0, even when the same trait is measured in purebred and crossbred animals. Lutaaya *et al.* (2001) found genetic correlations of 0.62 for growth rate, and 0.32 and 0.70 for backfat thickness between purebred and crossbred phenotypes. Whereas Cecchinato *et al*. (2010) found genetic correlation of 0.25 for piglet survival at birth. A strategy has been proposed in which crossbred animals are used in the training population to subsequently select purebred breeding animals for crossbred performance. This strategy is expected to give a higher response in crossbred performance compared with within-purebred-population selection (Dekkers 2007; Kinghorn *et al.* 2010; Van Grevenhof and Van Der Werf 2015). Besides the increase in response at the crossbred level, using crossbred data in the training population is also appealing because it allows breeding for traits for which phenotypes are scarce in purebreds. Some traits cannot be evaluated in nucleus herds, such as disease traits (Ibañez-Escriche and Gonzalez-Recio 2011).

The strategy of maximizing response to selection of purebreds for crossbred performance by using a crossbred training population has only been evaluated in simulation studies (Dekkers 2007; Kinghorn *et al.* 2010; Van Grevenhof and Van Der Werf 2015). The main issue in performing empirical studies is the need of phenotypes and genotypes of crossbred animals. The collection of these data is costly because this requires, besides genotyping, the individual recording of phenotypes on animals that are kept in group-housing systems and often have no pedigree information. Breeding companies were hesitant to make such investments.

Recently, however, crossbred data for genomic selection in pigs is becoming increasingly important.

In Chapter 5, data on purebred animals were used to predict performance of crossbreds. At the time, the number of genotyped crossbreds was not large enough to be used as a training population. Accuracies of predicting crossbred performance ranged from 0.11 to 0.31 for traits in which the genetic correlation between purebred and crossbred performance ranged from 0.88 to 0.90. These accuracies were not as great as accuracies for within-purebred-population, but they show the predictive ability between purebred and crossbred pigs. For the trait whose accuracy of prediction was zero, a low genetic correlation between purebred and crossbred performance was found (0.31) which is in line with this low accuracy. The predictive ability found for predicting crossbreds with purebred training data indicates that selection in the purebreds will result in a response in the crossbreds when the genetic correlation between purebred and crossbred performance is high.

In Chapter 5, the response variable for genomic prediction was a deregressed breeding value from a routine genetic evaluation. This breeding value was estimated based on records from a mix of purebred and crossbred animals. In practice, there is no problem with the use of a breeding value from a routine genetic evaluation in the evaluation. For research purposes however, it is important to investigate how the choice for purebred, crossbred, or a mix of data used to estimate the breeding values for genomic prediction affects accuracy. In Chapter 6, therefore, we looked into the source of phenotypic information used to estimate the breeding values for the training data set. Training on breeding values of purebred animals estimated using crossbred performance, resulted in more accurate prediction of crossbred genetic merit than training on breeding values of purebred animals estimated using purebred performance; as long as the breeding values that were used as response variable have the same reliability. Likewise, in a simulation study, Esfandyari *et al.* (2015) showed that selecting purebred animals based on crossbred performance data rather than on purebred performance data resulted in a greater response to selection in the performance of crossbred animals.

The results from Chapters 5 and 6 were promising and showed the ability of purebred data to predict performance of crossbred pigs. Thereafter, I wanted to test whether the use of crossbreds in the training population results in greater accuracies than solely using purebreds to select purebreds for crossbred performance. This analysis became possible because more data on crossbred animals became available

(Chapter 7). There was predictive ability when using crossbred phenotypes as training data, however, the accuracies were lower than from using purebred phenotypes. Results of simulation studies (e.g. Dekkers (2007)) that showed greater accuracy from using data on crossbreds rather than on purebred animals in the training population were not confirmed by my results. This discrepancy is explained by the high genetic correlation (>0.90) between purebred and crossbred performance for the traits studied in this thesis. The simulations studies consider a lower genetic correlation between purebred and crossbred performance (0.70 - 0.80) (Dekkers 2007; Van Grevenhof and Van Der Werf 2015). Further studies with other traits with lower genetic correlation between purebred and crossbred performance need to be carried out. I would expect that with lower genetic correlations between purebred and crossbred performance, the benefits from using crossbreds as training population would increase in comparison with using purebreds. With a breeding goal in which all traits have high genetic correlation between purebred and crossbred performance, there would be no need for a crossbred training population, current practice with purebred training would suffice. However, not all traits will have a correlation close to 1, as has been shown by other studies in pigs (Lutaaya *et al.* 2001; Cecchinato *et al.* 2010).

Although greater response to selection is observed in simulation studies from the use of crossbred data for training, these scenarios need to be carefully assessed. Factors such as the reliability of field records and the generation lag could hinder genomic prediction (Ibañez-Escriche and Gonzalez-Recio 2011). As phenotypes will be recorded in crossbreds from commercial farms, the recording system must be well designed and correctly applied because the large number of crossbred animals might be a hindrance to data collection compared with nucleus farms. On top of that, the difference in generations between purebred selection candidates and crossbred pigs, might hamper the genetic gain of genomic selection based on crossbreds. Thus, there is a need for studying whether the additional genetic gains promised by simulations can be confirmed by empirical studies. The additional genetic gains must offset the disadvantages mentioned above.

Using crossbred pigs in the training population to select purebreds for crossbred performance also has an effect on the purebred genetic progress. When genetic correlation between purebred and crossbred performance is high, one will still observe purebred genetic progress. If, however, the genetic correlation is low, one can expect less genetic progress in purebred, or even negative values. With crossbred training populations, the evaluation of breeding program performance will need to shift from analyzing the genetic progress in purebreds to monitoring the improvement of crossbred performance.

9.3.5 Using dominance information for genomic prediction

Dominance is important in crossbreeding schemes as it is the likely basis of heterosis (Xiao *et al.* 1995; Falconer and Mackay 1996; Charlesworth and Willis 2009). Therefore, using a model that accounts for dominance is expected to be beneficial for genomic prediction with a crossbred training population. Hence, I have evaluated genomic prediction when dominance effects are accounted for in the model using a crossbred training population (Chapter 8).

Some studies have reported dominance variance estimates using real pig data and pedigree-based models (Culbertson *et al.* 1998; Norris *et al.* 2010). Estimates of dominance variance are not so precise because they require massive amounts of data especially on full-sib families (Vitezica *et al.* 2013). Dominance variance estimates from pedigree information were found to be zero for gestation length and total number of piglets born (Chapter 8). With genomic information, dominance variance can be estimated more precisely based on heterozygosity of SNP genotypes (Vitezica *et al.* 2013). Studies using genomic data in purebred pigs, showed that non-additive effects are relevant factors contributing to the genetic variation of the studied traits (Su *et al.* 2012; Nishio and Satoh 2014). In addition, they also showed that accounting for the dominance effects improved accuracy of genomic prediction, compared to accounting only for additive effects. Using genomic data from crossbred pigs I showed that, for a trait with dominance variation, accounting for dominance effects can slightly improve genomic predictions compared with accounting only for additive effects (Chapter 8) similar to the reports on purebred pigs mentioned above. Even though there was a slight improvement in prediction from adding the dominance effect, I expect that the inclusion of non-additive effects in routine genetic evaluations is still a long time ahead of us, if breeding companies will ever include them at all. It has been shown that breeding programs should focus on additive effects as they account for more than 50%, and often even 100% of the genetic variation (Hill *et al.* 2008).

Besides a dominance model, a model accounting for breed-specific effects of marker alleles may be relevant in prediction of crossbreeding performance (Ibánez-Escriche et al. 2009). I have found indications that the proportion of genetic variance in crossbred performance differs between the parental purebreds that contributed to the cross (Chapter 8). Such a model, however, needs to be empirically investigated before implementation in breeding programs can be considered.

9.4 Concluding remarks

In the first part of this thesis I describe research that detected genetic markers significantly associated with gestation length, fine-mapped a QTL region for androstenone level, and studied potential pleiotropic effects. I expect that GWAS will continue to be performed because they provide scientifically relevant results, especially with the greater statistical power when more animals will be sequenced or genotyped using HD SNP chips. With more markers, the physical distance between marker and the causative mutation will be shortened, therefore, QTL regions can be fine-mapped. However, finding the causative mutation will require more than just a GWAS using denser genotyping or sequence data. Linkage disequilibrium plays a major role in GWAS and one may require addition functional evidence to distinguish associated variants. The results of GWAS can be incorporated in a MA-GBLUP, to increase the accuracy of genomic prediction compared with GBLUP.

In the second part of this thesis I describe genomic prediction using purebred and crossbred pigs, which is a subject that is highly relevant for pig breeding. Although little has been reported so far, efforts to have more data on crossbred animals have been ongoing and contributed to the analyses performed in this thesis. I have shown that there is predictive ability from using phenotypes of crossbred animals to predict the genetic merit of purebred animals for crossbred performance. Even though the results obtained did not confirm the simulation results, I expect that for other traits with low genetic correlation between purebred and crossbred performance, the simulation results will be confirmed. If confirmed in empirical studies, the use of crossbred training populations for genomic selection will be implemented by breeding companies. The implementation of crossbred training population will, at least in the foreseeable future be without accounting for non-additive effects. Reasons for omitting non-additive effects from prediction models are the large proportion of the total genetic variance explained by additive effects, the increased computational power required to generate for example a genomic dominance matrix, and the negligible added-value to accuracy shown so far from adding dominance to genomic prediction.

References

Animal QTLdb. http://www.animalgenome.org/QTLdb. Accessed 21 Aug 2015

Calus MPL, Meuwissen THE, De Roos APW, Veerkamp RF (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–61.

Cecchinato A, de los Campos G, Gianola D, et al (2010) The relevance of purebred information for predicting genetic merit of survival at birth of crossbred piglets. J Anim Sci 88:481–490.

Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. Nat Rev Genet 10:783–96.

Chen L, Vinsky M, Li C (2015) Accuracy of predicting genomic breeding values for carcass merit traits in Angus and Charolais beef cattle. Anim Genet 46:55–59.

Culbertson MS, Mabry JW, Misztal I, et al (1998) Estimation of dominance variance in purebred Yorkshire swine. J Anim Sci 76:448–451.

Daetwyler HD, Capitan A, Pausch H, et al (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46:858–865.

De Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. Genetics 183:1545–53.

Dekkers JCM (2007) Marker-assisted selection for commercial crossbred performance. J Anim Sci 85:2104–14.

Duijvesteijn N, Knol EF, Merks JWM, et al (2010) A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC Genet 11:42.

Esfandyari H, Sørensen AC, Bijma P (2015) Maximizing crossbred performance through purebred genomic selection. Genet Sel Evol 47:1–16.

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics, 4th edn. Longman, Harlow, England

Fujii J, Otsu K, Zorzato F, et al (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. Science 253:448–451.

Gilmour AR, Gogel BJ, Cullis BR, Thompson R (2009) ASReml user guide release 3.0.

Grindflek E, Lien S, Hamland H, et al (2011) Large scale genome-wide association and LDLA mapping study identifies QTLs for boar taint and related sex steroids. BMC Genomics 12:362.

Harris BL, Johnson DL, Spelman RJ (2008) Genomic selection in New Zealand and the implications for national genetic evaluation. Proc. Interbull Meet. Niagara Falls, p 325–330

Hayes BJ, Bowman PJ, Chamberlain AC, et al (2009a) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol 41:51.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009b) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–43.

Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4:e1000008.

Ibánez-Escriche N, Fernando RL, Toosi A, Dekkers JCM (2009) Genomic selection of purebreds for crossbred performance. Genet Sel Evol 41:12.

Ibáñez-Escriche N, Forni S, Noguera JL, Varona L (2014) Genomic information in pig breeding: Science meets industry needs. Livest Sci 166:94–100.

Ibañez-Escriche N, Gonzalez-Recio O (2011) Review. Promises, pitfalls and challenges of genomic selection in breeding programs. Spanish J Agric Res 9:404–413.

Kinghorn BP, Hickey JM, Werf JHJ Van Der (2010) Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals. Proc. 9th WCGALP. p 36

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

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

Lund MS, Su G, Janss L, et al (2014) Genomic evaluation of cattle in a multi-breed context. Livest Sci 166:101–110.

Lutaaya E, Misztal I, Mabry JW, et al (2001) Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. J Anim Sci 79:3002–7.

Merks JWM, De Vries AG (2002) New sources of information in pig breeding. Proc. 7th WCGALP. p 3–10

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

Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Misztal I (2011) FAQ for genomic selection. J Anim Breed Genet 128:245–246.

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.

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

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

Norris D, Varona L, Ngambi JW, et al (2010) Estimation of the additive and dominance variances in SA Duroc pigs. Livest Sci 131:144–147.

Purfield DC, Bradley DG, Evans RD, et al (2015) Genome-wide association study for calving performance using high-density genotypes in dairy and beef cattle. Genet Sel Evol 47:47.

Ramos AM, Crooijmans RPMA, Affara NA, et al (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.

Sahana G, Guldbrandtsen B, Thomsen B, et al (2014) Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. J Dairy Sci 97:7258–7275.

Su G, Christensen OF, Ostersen T, et al (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One 7:e45293.

Tiezzi F, Maltecca C (2015) Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. Genet Sel Evol 47:1–13.

Tusell L, Pérez-Rodriguez P, Forni S, et al (2013) Genome-enabled methods for predicting litter size in pigs : a comparison. Animal 7:1739–1749.

Van Eenennaam AL, Weigel KA, Young AE, et al (2014) Applied Animal Genomics: Results from the Field. Annu Rev Anim Biosci 2:105–139.

Van Grevenhof IEM, Van Der Werf JHJ (2015) Design of reference populations for genomic selection in crossbreeding programs. Genet Sel Evol 47:1–9.

VanRaden PM, Van Tassell CP, Wiggans GR, et al (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92:16–24.

Veroneze R (2015) Linkage disequilibrium and genomic selection in pigs. PhD Thesis. Wageningen University

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 PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9:255–266.

Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195:1223–1230.

Wu X, Lund MS, Sahana G, et al (2015) Association analysis for udder health based on SNP-panel and sequence data in Danish Holsteins. Genet Sel Evol 47:50.

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

Zhang Z, Liu J, Ding X, et al (2010) Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. PLoS One 5:1–8.

Summary

Summary

In the last decade, high-throughput genomic information became available for most livestock species. Efficient use of this information is important for the competitiveness of a breeding company. Application of genomic selection (GS) in pigs, may require different strategies from what is currently applied in dairy cattle because the end product in pig production is a crossbred animal. In this work, I explored the use of genomic information for the genetic improvement of purebred and crossbred pigs. Firstly, working mainly in purebred animals, regions affecting gestation length (Chapter 2) and androstenone level (Chapter 3) were detected in the pig genome by genome-wide association and fine-mapping. Also, potential pleiotropic effects of the androstenone level quantitative trait locus (QTL) on reproductive traits were studied (Chapter 4). Secondly, we investigated the potential of GS in pig breeding by determining the accuracy of genomic prediction using different strategies. These strategies varied in training and validation populations, selected from multiple purebred lines and their crossbred offspring, different data types and models.

Genome-wide association study (GWAS) identified two QTL regions for gestation length, one in the Dutch Landrace and one in the Large White (Chapter 2). Three associated SNP were detected in a QTL region spanning 0.52 Mbp on *Sus scrofa* chromosome (SSC) 2 in Dutch Landrace and for the Large White, four associated SNP were detected in a region of 0.14 Mbp on SSC5. The region of a previously detected QTL for androstenone level on SSC6 was fine-mapped, narrowing the region down from 3.75 Mbp to 1.94 Mbp and identifying a candidate mutation in *SULT2A1* (Chapter 3). This fine-mapped region was evaluated for possible pleiotropic effects on production and reproduction traits in pigs (Chapter 4). No unfavorable pleiotropic effects were found, indicating that using the studied marker for selection would not unfavorably affect the other relevant traits.

In the later chapters I have investigated the potential of different strategies for the implementation of GS in pig breeding when the aim is to improve crossbred performance. Within-population prediction was showed considerably high accuracy of prediction (Chapters 5 and 6) while across-population prediction, evaluated in Chapter 5 had accuracies close to zero. Multi-population prediction, where combinations of Dutch Landrace and Large White animals plus their cross were used as training showed that adding data from other populations did not improve the

accuracy except when predicting the F1 cross with records from both parental populations added to the F1 training data. When only purebred data was used, there was some predictive ability for crossbred performance (Chapter 5). In the first study the training data contained a mix of records measured on purebred and crossbred animals. In Chapter 6, therefore, the source of training data was clearly separated into purebred and crossbred records. Training on breeding values of purebred animals that were estimated using crossbred offspring performance, resulted in more accurate prediction of their crossbred genetic merit compared with training on breeding values of those same animals, estimated using purebred offspring performance. Genotyped and phenotyped crossbreds in the training population were expected to have higher accuracies when predicting genetic merit for crossbred performance. However, in Chapters 5 and 6 we did not test this strategy because sufficient genotyped crossbred were lacking at that time. Later, with more crossbred data, we evaluated this strategy and the accuracies were not improved over the use of genotyped and phenotyped purebreds (Chapter 7) mainly due to the high genetic correlation between purebred and crossbred performance for the studied traits. Finally, the inclusion of dominance in the model, with a crossbred training population was evaluated. For a trait that had dominance variation, accounting for dominance effects can be slightly beneficial for genomic prediction compared with a model that accounts only for additive effects.

Finally, in Chapter 9, the relevance of the findings was discussed, how breeders can benefit from the combination of genomic selection with the information of individual QTL. To finalize, I make suggestions for future studies and how breeders can make use of the results generated in the thesis.

Acknowledgements

Acknowledgements

This PhD degree would not be possible without the help and contribution of many people. I am very grateful to everybody that directly or indirectly contributed to the completion of this thesis.

Thanks to my three supervisors that provided me knowledge, confidence and friendship to finalize this PhD degree.

DJ, thanks for everything, especially during the time I spent in Uppsala. It has always been nice to chat with you in any occasion, always with nice jokes and emoticons via Skype ☺. Even though apparently you were not that happy when I was singing while getting to my office early in the morning… "Too easy this PhD life" hehe. Now instead of keeping the piggies always pinky I will try to keep the cows always spotted ☺.

John, besides the technical genetic part, be sure that I learnt a lot on how to look at a manuscript in a more critical editorial way because of you. Your comments and modifications are always useful and I really look up to your writing style. At the same time I learnt a lot during your daily supervision that I will take and apply to my researcher life. I hope you didn't get annoyed by the many times I asked whether you had revised my manuscript. Thanks for the support given to me during these four years regarding trip to the US, writing more papers, job applications, having holidays, everything. You always gave me the support to believe that was possible and indeed it was…

Martien, thank you for the support even when I changed the direction of the PhD project deviating from your main field of expertise. You are very busy yet your door was always open for me whenever I had doubts or comments. Thanks for being my promoter in such a unique opportunity that I had of doing my PhD abroad.
Thanks to my committee members (opponents) for reading the thesis and hopefully we will have a fruitful discussion during the defense ceremony.

I very much appreciate the EGS-ABG initiative including all coordinators, secretaries and participants of this amazing program. Thanks to the European Union and Topigs Norsvin for funding my 4 years of PhD studies.

Thanks to Wageningen University and the Swedish University of Agricultural Sciences for the opportunity to study in such renowned universities.

Thanks to Ada, Lisette and Monique from ABGC as well as to Helena, Harriet and Monica from the Animal Breeding and Genetics department at SLU for all the help with paperwork.

To all staff and people that I met from WUR and SLU, I learnt a lot during QDGs, seminar series, TLMs, genomics meetings, coffee- and lunch-breaks, lab meetings, etc.

Thanks to Topigs people for being always very open to receive me in the company, to provide data and to discuss results, doubts, etc. You are part of a nice company and wish you all the success! I really appreciated the way PhD candidates are treated there, taking part in meetings, colloquium with the doors always open to receive us.

Thanks Jack Dekkers for receiving me in Ames as a visiting scholar, also Jian and Rohan. It was a great time spent in Ames learning a lot, getting to know very nice people and facing a very nice cold weather! Thanks to the whole Animal Breeding and Genetics group in Ames for receiving me very well during the 3 months! Valeu Nicola pela ajuda também, tocar um violão, cantar um Raimundos, torcer pro São Paulo e ter conhecido a sua sensacional família!

Obrigado a todo o pessoal que convivi como amigo ou estudante durante meus estudos na Universidade Estadual de Maringá e da Universidade Federal de Viçosa, vocês sempre me apoiaram e me deram forças! Um agradecimento especial para o Elias que despertou meu interesse pelo melhoramento! Também agradeço o Paulo Sávio, Simone e o Fabyano por todo o período que estive em Viçosa e após também, além da ótima relação de amizade que tenho com vocês!

Muito obrigado ao Brenno Harry Oliveira Silva por desenvolver a criação e design gráfico da capa! Obrigado também ao Cleibe Pinto por ter supervisionado todo esse processo!

These 4 years of PhD life passed by in a blink of an eye. One of the main reasons for that was the amazing group of people that I had the opportunity to know and to

build up very nice friendship. There are too many awesome people that I have met in Wageningen and Uppsala, as well as in conferences, EGS-ABG, courses, etc. Thanks for everything, I really enjoyed spending plenty of time discussing scientific matters, the PhD life, having barbecues, kebabs, ribs, Chinese restaurant, playing games, pub-quizing and the list goes on.

Special thanks to a bunch of people, Nancy, words are inadequate to express my gratitude of having your friendship! Hehehe! We really had a nice time during these 4 years in both Wageningen and Uppsala. Thanks for all the conversations that we had about life, PhD, papers, jokes, etc. Thank you very much for you English lessons and singing performances! I wish you all the best! Sandrine muito obrigado pela sua amizade, tanto na Holanda como na Suécia, apesar da sua personalidade forte que eu sempre soube contorná-la com uma boa risada, você sempre quer o bem das pessoas e sei que sempre posso contar com você, assim como você pode contar comigo! Katrijn thanks for all the talks, parties, rides, singing, sewing, trips, dinners, house renting, and especially your friendship! Gus, Mr. Munni, thanks for a true friendship. You really helped me to mingle and to develop my social skills. All the dinners at your place, the amazing hamburger with cheese inside! Johnny Cash songs! Australian day! All the room sharing during conferences and drinking from the bottle sideways! All the Aussie lessons! Get a dog up ya! Mathieu thanks for all the fish and French lessons, nice talks, songs, trips and for an OTH friendship! Naomi, thanks a lot for your friendship as well as a lot of Guinness, M&Ms, Dutch lessons that I will have with your daughters now that we are neighbours hehe! Juanma, the Spanish guy! Thanks for the friendship, always sharing some good stories, good food and losing football matches on the pro-nights hehe! ☺. Claudia valeu pela amizade, por sempre ter assunto bom pra conversar, um ótimo humor e pelo seu amor pelo Brasil heheheheh! Obrigado Renata e Lucas por dividirem um bom tempo por aqui na Holanda, revivendo os tempos de Viçosa hehe! Thanks also to my office-mates Dianne, Marzieh, Yvonne, always sharing some food, jokes and scientific questions! Thanks to the “PhD mafia” in Uppsala, Merina, Thu, Agnese, Berihu, Ahmed, Chrissy, Bingjie, Josh, Alberto, Sangeet, Fabiana, Iris, Axel, Jonas! Some nice table-tennis matches, charades, a lot of singing and of course good food! Each and every one of you really made every weekend fun in Uppsala!

Marcola, pensa num cara parceiro e que sempre estava presente pro que der e vier nesses 4 anos por aqui! Ainda me lembro de quando você veio pra Wageningen para nos encontrarmos de novo aqui na Holanda e eu comi o croquete com mostarda hehehe! Todas as viagens que fizemos e ainda faremos, discussões científicas e não científicas, academia, ótimas conversas e refeições, ou seja, qualquer coisa. Não tenho muitas palavras, apenas um obrigado por ser seu amigo!

Gabriel, big G! Since the first days in Wageningen you were always with a smile on your face and ready to share a nice conversation. I really enjoy spending time with you! Even though you were in Denmark for 2 years whenever we would meet it seemed that we had seen each other just a little while ago... Thanks a lot for your friendship and be sure que esta amizade é sincera!

Família, não tem nada mais importante do que isso! Muitíssimo obrigado por todo o carinho, ajuda, suporte, conversas, visitas durante não só esse período do doutorado, mas durante minha vida toda! Muito obrigado por me permitirem ir atrás das minhas conquistas sempre me apoiando! Eu realmente sou muito sortudo de ter uma família tão boa!

Davi, obrigado por sempre querer falar no Skype já que as vezes não sou muito adepto da comunicação hehe! Bruna, também muito obrigado! Tia Ete, obrigado pela força e por sempre estar interessada em como eu estou e com as coisas vão por aqui! Mami e Papi, realmente não tenho palavras para vocês, apenas de que tudo que sou é fruto do que aprendi com vocês! Estou julgando que sou algo bom hehe! De verdade, muito obrigado por tudo!

Vica, os 4 anos se passaram! Ainda lembro de quando me disse que valia a pena tentar! E valeu! Obrigado pelo amor, amizade, sinceridade, tudo que você me proporciona! Agora começando uma nova empreitada nas nossas vidas juntos! Muito obrigado também a sua família pelo apoio que sempre me(nos) dão! ☺

Curriculum vitae

About the author

André Marubayashi Hidalgo was born on the 15th of October 1987 in Maringá, Brazil. Once as a kid he said: “I will have a farm with all animals of the world in it”. This admiration for animals led him to pursue his BSc diploma in Animal Science at the Universidade Estadual de Maringá. He obtained his Bachelor in 2009 with the thesis entitled “Genetic characterization of egg weight, egg production and age at first egg in quails” under the supervision of Prof. Dr. Elias Nunes Martins. In 2010, he started his MSc studies in Animal Breeding and Genetics at the Universidade Federal de Viçosa, Brazil. In 2011, he obtained his MSc diploma with the thesis entitled “Fine mapping and single nucleotide polymorphism effects estimation on pig chromosomes 1, 4, 7, 8, 17 and X” under the supervision of Prof. Dr. Paulo Sávio Lopes. In 2011, André was selected for the PhD project entitled “Genetic analysis of performance traits in pigs: exploiting genomic information in crossbreeding schemes” in the European Graduate School in Animal Breeding and Genetics. The project was a collaboration between Wageningen University, Swedish University of Agricultural Sciences and Topigs Norsvin. In September 2011, therefore, André started his PhD studies in Wageningen, the Netherlands. From April 2013 to September 2014, André took part of his PhD studies in Uppsala, Sweden. Early 2015, he spent 3 months doing research in the group of Jack Dekkers in Ames, USA. During these 4 years, André wrote the PhD thesis entitled “Exploiting genomic information on purebred and crossbred pigs”. In October 2015, he started a job in a cattle improvement company, CRV, as genomics researcher following his interest in animal breeding and applied science.

Peer-reviewed publications

Hidalgo, AM; Lopes, MS; Harlizius, B; Bastiaansen, JWM. Genome-wide association study reveals regions associated with gestation length in two pig populations. Anim Genet, in press.

Hidalgo, AM; Bastiaansen, JWM; Lopes, MS; Harlizius, B; Groenen, MAM; De Koning, DJ. Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3 - Genes Genom Genet, 5:1575-1583, 2015.

Hidalgo, AM; Bastiaansen, JWM; Lopes, MS; Veroneze, R; Groenen, MAM; De Koning, DJ. Accuracy of genomic prediction using deregressed breeding values estimated from purebred and crossbred offspring phenotypes in pigs. J Anim Sci, 93(7): 3313-3321, 2015.

Hidalgo, AM; Bastiaansen, JWM; Harlizius, B; Knol, EF; Lopes, MS; De Koning, DJ; Groenen, MAM. Asian low-androstenone haplotype on pig chromosome 6 does not unfavorably affect production and reproduction traits. Anim Genet, 45:6, 2014.

Hidalgo, AM; Bastiaansen, JWM; Harlizius, B; Megens, HJ; Madsen, O; Crooijmans, RPMA; Groenen, MAM. 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, 2014.

Hidalgo, AM; Silva, LP; Mota, RR; Martins, EN. Canonical-correlation analysis applied to selection-index methodology in quails. Livest Sci, 169: 35-41, 2014.

Hidalgo, AM; Lopes, PS; Paixão, DM; Silva, FF; Bastiaansen, JWM; Paiva, SR; Faria, DA; Guimarães, SEF. Fine mapping and single nucleotide polymorphism effects estimation on pig chromosomes 1, 4, 7, 8, 17 and X. Genet Mol Biol 36(4), 2013.

Hidalgo, AM; Martins, EN; Santos, AL; Quadros, TCO; Ton, APS; Teixeira, R. Genetic characterization of egg weight, egg production and age at first egg in quails. R Bras Zootecn, 40, p. 95-99, 2011.

Veroneze, R; Lopes, MS; **Hidalgo, AM**; Guimarães, SEF; Silva, FF; Harlizius, B. Lopes, PS; Knol, EF; Van Arendonk, JAM; Bastiaansen, JWM. Accuracy of genome-enabled prediction exploring purebred and crossbred pig populations. J Anim Sci, 93(10): 4684-4691.

Mota, RR; Marques, LFA; Lopes, PS; Silva, LP; **Hidalgo, AM**; Leite, CDS; Torres, RA. Random regression models in the evaluation of the growth curve of Simbrasil beef cattle. Genet Mol Res, 12(1): 528-536, 2013.

Paixão, DM; Carneiro, PLS; Paiva, SR; Sousa, KRS; Verardo, LL; Braccini Neto, J; Pinto, APG; **Hidalgo, AM**; Nascimento, C; Périssé, IV; Lopes, PS; Guimarães, SEF. Detection of quantitative trait loci on chromosomes 1, 2, 3, 12, 14, 15, X in pigs: performance characteristics. Arq Bras Med Vet Zootec, 65(1): 213-220, 2013.

Paixão, DM; Carneiro, PLS; Paiva, SR; Sousa, KRS; Verardo, LL; Braccini Neto, J; Pinto, APG; **Hidalgo, AM**; Nascimento, C; Périssé, IV; Lopes, PS; Guimarães, SEF. Mapping of QTL on chromosomes 1, 2, 3, 12, 14, 15 and X in pigs: characteristics carcass and quality of meat. Arq Bras Med Vet Zootec, 64(4): 974-982, 2012.

Santos, AL; Scapinello, C; Martins, EN; Granzotto, F; Paula, MC; **Hidalgo, AM**. Genetic evaluation of weight gain and feed-to-gain ratio of White New Zealand rabbits raised in different environments. Acta Sci Anim Sci, 32, 2010.

Training and education

Training and supervision plan

The basic package (9 ECTS)	year	credits
Welcome to the EGS-ABG	2011	2.0
Course on philosophy of science and/or ethics	2011	1.5
Early fall research school “Animal breeding and society”	2012	2.0
WIAS introduction course	2012	1.5
EGS-ABG fall research school	2013	2.0
Scientific exposure (11 ECTS)	year	credits
International conferences (5.1 ECTS)		
63rd EAAP annual meeting, Bratislava (Slovakia), 27-31.08	2012	1.2
64th EAAP annual meeting, Nantes (France) 26-30.08	2013	1.2
10th WCGALP, Vancouver (Canada), 17-23.08	2014	1.5
66th EAAP annual meeting, Warsaw (Poland) 31.08-04.09	2015	1.2
Seminars and workshops (0.6 ECTS)		
WIAS science day, Wageningen	2012	0.3
WIAS science day, Wageningen	2013	0.3
Presentations (5.0 ECTS)		
63rd EAAP annual meeting, Bratislava (Slovakia), ORAL	2012	1.0
64th EAAP annual meeting, Nantes (France), ORAL	2013	1.0
WIAS science day, Wageningen (Netherlands), POSTER	2013	1.0
10th WCGALP, Vancouver (Canada), ORAL	2014	1.0
66th EAAP annual meeting, Warsaw (Poland), ORAL	2015	1.0
In-depth studies (35 ECTS)	year	credits
Disciplinary and interdisciplinary courses (19 ECTS)		
Sequence data analysis training school	2012	1.5
Next generation sequencing - applications in animal breeding and genetics	2012	5.0
Identity by descent (IBD) approaches to genomic analysis of genetic traits	2012	1.2
Advanced methods and algorithms in animal breeding with focus on GS	2012	1.5
Genetic analysis using ASReml 4.0	2014	1.5
Introduction to statistical methods in quantitative genetics and breeding	2014	4.0
Advanced quantitative genetics for animal breeding	2014	3.0
Introduction to theory and implementation of genomic selection	2014	1.35
Advanced statistics courses (3 ECTS)		
Statistics for the life sciences	2012	2.0
Advanced statistics course: design of experiments	2012	1.0
PhD students' discussion groups (1 ECTS)		
Quantitative genetics discussion group	2012	1.0
MSc level courses (12 ECTS)		
Genomics (ABG-30306)	2011	6.0
Genetic improvement of livestock (ABG-31306)	2011	6.0

Professional skills support courses (6 ECTS)	year	credits
Techniques for writing and presenting a scientific paper	2012	1.2
Teaching and supervising thesis students	2012	1.0
Writing grant proposals	2015	2.0
High-impact writing course	2015	1.3
Survival guide to peer review	2015	0.3
Research skills training (5 ECTS)	year	credits
Introduction to R for statistical analysis	2012	0.6
Getting started in ASReml	2013	0.3
External training period: SLU, Sweden	2013	2.0
External training period: ISU, USA	2015	2.0
Didactic skills training (2 ECTS)	year	credits
Supervising practicals and excursions		
Genomics (ABG-30306)	2012	1.0
Genomics (ABG-30306)	2014	1.0
Education and training total (68 ECTS)		

Colophon

Colophon

This work was supported the European Commission (within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”), Breed4Food (a public-private partnership in the domain of animal breeding and genomics) and Topigs Norsvin.

The cover of this thesis was designed by Brenno Harry Oliveira Silva with the supervision of Cleibe Pinto.

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