Combined purebred and crossbred information for genomic evaluation in pig

Tao Xiang

NNT : 2017IAVF0014

THESE DE DOCTORAT

préparée à l'Institut des sciences et industries du vivant et de l'environnement (AgroParisTech)

pour obtenir le grade de

Docteur de l'Institut agronomique, vétérinaire et forestier de France

Spécialité : Génétique Animale

École doctorale n°581
Agriculture, alimentation, biologie, environnement et santé (ABIES)

par

Tao XIANG

COMBINED PUREBRED AND CROSSBRED INFORMATION FOR GENOMIC EVALUATION IN PIG

Directeur de thèse : **Andres LEGARRA**
Co-directeur de la thèse : **Ole Fredslund CHRISTENSEN**

Thèse présentée et soutenue à Aarhus University, Denmark, le 8 septembre, 2017 :

Composition du jury:

M. Bernt GULDBRANDTSEN, Associate professor, Aarhus University	Rapporteur, Président
M. Mario CALUS, Senior Researcher, Wageningen University	Rapporteur
M. Jörn BENNEWITZ, Professor, University of Hohenheim	Rapporteur
Mme Zulma Vitezica, Associate professor, Institut National Polytechnique de Toulouse	Examinateur
Mme Birgitte ASK, Chief researcher, Danish Pig Research Center, SEGES	Examinateur
M. Andres LEGARRA, Research director, INRA	Directeur de thèse
M. Ole Fredslund CHRISTENSEN, Senior researcher, Aarhus University	Co-directeur de thèse

GenPhySE, INRA
UMR 1388, CS 52627, 31326, Castanet Tolosan, France.

QGG MBG, Aarhus University
Blichers alle 20, 8830 Tjele, Denmark

The thesis is dedicated to the memory of my paternal grandfather and maternal grandfather.

I deeply miss you. 

谨以此论文献给我亲爱的爷爷和外公，深深怀念你们！

Table of Contents

Preface

This thesis is submitted to The Graduate School of Science and Technology (GSST), Aarhus University, Denmark and Doctoral School of Agriculture Food Biology Environment Health (ABIES), AgroParisTech, France in fulfilment of requirements for the double PhD degrees. The PhD project was carried out at the Center for Quantitative Genetics and Genomics (QGG), Department of Molecular Biology and Genetics, Aarhus University, Tjele over the period 2013.09~2015.08 and 2016.09~2017.06, and Génétique Physiologie et Systèmes d'Elevage (GenPhySE) laboratory, Animal Genetics Division, INRA, Toulouse over the period 2015.09~2016.08. This PhD project was funded by Erasmus-Mundus Joint Doctorate program "European Graduate School in Animal Breeding and Genetics (EGS-ABG)", Aarhus University and Danish Pig Research Center, SEGES, Denmark.

My deepest gratitude goes first and foremost to my main supervisors, Ole Fredslund Christensen and Andres Legarra. Thank you very much for your great guidance and constant encouragement during the 4 years! I have been incredibly fortunate to have such two outstanding and gracious supervisors. You gave me enough freedom to explore areas of interest and guidance to a correct direction. You always taught and inspired me with great patience. Your continuous help and support made me more and more confident, independent and be fond of doing research. You two make my graduate experience be one that I will cherish forever!

I am also deeply indebted to my co-authors of papers. They are: Bjarne Nielsen and Tage Ostersen, from Danish Pig Research Center; Guosheng Su and Peipei Ma, from QGG; Zulma Gladis Vitezica, from INP, ENSAT, Université de Toulouse. Thank you for making substantial contributions to this thesis. Your valuable comments and instructions made my PhD study go straightforward.

I would like to express my gratitude to all those who helped me to make the PhD study possible. Particularly, I owe a special debt of gratitude to Per Madsen, for many instructive advices that helped me to solve practical problems in running DMU; Bernt Guldbrandtsen, for great help in programming in Perl and shell scripts; Elise Norberg and Louise Dybdahl Pedersen, on whom I can rely, for giving many useful suggestions and promoting my family reunion; Luc Janss, Jan Lassen, Daniel Sorensen, Beatriz Castro Dias Cuyabano, Pernille Merete Sarup, Goutam Sahana, and Anders Christian Sørensen for giving me a first look into the world of quantitative genetics, linear models, Bayesian theories, gene mapping, and breeding program etc.; Tina Vammen, Cindie Deleuran, and Karin Smedegaard, for expert guidance in coping with practical issues; Mogens Sandø Lund, and Just Jensen for their constructive advices to the thesis through different kinds of discussions; Llibertat Tusell Palomero, Eduardo Manfredi and Jérôme Raoul, for timely help to make my life in France convenient and enjoyable.

My sincere thanks would go to my current and former office mates: Dario Fe, Rafael Pimental Maia, Xiujin Li, Silvia Rodriguez Ramilo, Carolina Andrea Garcia-Baccino, Ahmed Sayed Ismael, Grum Gebreyesus, Lu

Cao, and Jette Odgaard Villemoes. Thanks for your enthusiasm and giving me endless happiness. Thanks Jette a lot for translating the summary from English to French!

I would like to acknowledge my Chinese friends and buddies in Denmark and France. You made my life so colorful and wonderful! Excuse me for not listing a long list of your names.

Finally, my heart swells with gratitude to my beloved parents and wife - Fang Fang. I greatly appreciate your continuous support, understanding, tolerance, and endless love. Thanks for having you along the way! I love you!

Tao Xiang (项韬)

June 6, 2017, Tjele

List of Abbreviations

EBV: Estimated breeding value

GBLUP: Genomic best linear unbiased prediction

GLS: Generalized least squares

GS: Genomic selection

LD: Linkage disequilibrium

LL: Landrace

LY: Landrace_Yorkshire crossbreds

MAF: Minor allele frequency

NUH: Number of unique haplotypes

PCA: Principal components analysis

PSH: Proportion of shared haplotypes

QTL: Quantitative trait locus

SEP: Standard error of prediction

SNP: Single nucleotide polymorphism

ssGBLUP: Single-step genomic best linear unbiased prediction

TNB: Total number of piglets born

YL: Yorkshire_Landrace crossbreds

YY: Yorkshire

CHAPTER 1

General Introduction

The “crossbred animals” section in page 10 and “single-step GBLUP method” section in page 15 are expansions of the introduction part in my mid-term qualifying examination progress report.

In this general introduction, the use of crossbred animals in meat production is first introduced. It is well known that crossbred animals perform better than purebred animals due to the heterosis and complementarity. The genetic background of heterosis and researches on heterosis and dominance effects are then reviewed. Genomic selection in pigs, especially for crossbred performance is described in the next paragraph. Among approaches for genomic selection, the single-step method can overcome the issue that not all the involved animals are genotyped. The current development of single-step method is then described. Next, a new concept of metafounder is introduced. With this new concept, single-step method can be executed in another way for crossbred performance. Then studies about genotype imputation and the trait of total number of piglets born are introduced. Finally, the general goals of this thesis are presented.

Crossbred animals

Crossbreeding is used for producing final production animals whose sires and dams originate from different breeds or lines (Falconer and Mackay, 1996). Crossbreeding is used in almost all species of livestock, especially intensively for pig and chicken (Wei, 1992). In the pig industry, usually, breeding companies manage the pure breeds and run the selection program. Purebred animals are bred in nucleus herds, where the environment is controlled in high hygienic status and performance recording systems are standardized (Dufrasne, 2015). Two-way crossbred sows (F1 sows) are produced and fed in the multiplier herds, where environment is controlled at a medium hygienic status, and sold to the sow farmers. Then sow farmers produce the three-way crossbred pigs by using F1 sows and boars that are from a third breed. Three-way crossbred pigs in fattening farms are sold to the slaughterhouses and finally, pork enters the market (Van Arendonk et al., 2010). In Denmark, Danish Landrace (L) and Yorkshire (Y) are used as dam lines to reproduce the F1 sows (either LY or YL) while the Danish Duroc (D) is used as boar line in the three-way crossbreeding programs (Sørensen, 2003). This thesis will concentrate on the two-way crossbreeding system.

The increased performance of crossbred animals compared to purebred animals is attributable to heterosis effects and complementarity between breeds (Wei and van der Werf, 1994). In crossbreeding, heterosis is a specific combining ability of two breeds, instead of two individuals, that lead to the crossbred offspring performs better than the average of the purebred parental breeds (Falconer and Mackay, 1996). Another advantage of using crossbreeding is to utilize breed complementarity. For instance, F1 sows have good maternal abilities whereas purebred boars have good meat production traits that are transmitted to the terminal pig. Crossbreeding can be optimized because it benefits from mating individuals from different breeds to maximize their strengths and minimize their weaknesses (Falconer and Mackay, 1996). In pig industry, pork producers widely use crossbred animals to improve commercial production traits (Hidalgo, 2015). However, the superiorities from the hybrids cannot be retained across generations. Thus, genetic selection programs mostly focus on pursuing genetic process in pure breeds, with the hope that this genetic

progress is also expressed at the crossbred performance level. Programs that select explicitly to improve the crossbred performance rather than purebred performance have always been proposed (Comstock et al., 1949; Hartmann, 1992), but their use is limited due to practical constraints (Wei and van der Werf, 1994).

Heterosis, dominance and inbreeding depression

There are two types of heterosis: individual and maternal heterosis (Falconer and Mackay, 1996). Individual and maternal heterosis originates from the genes of the individual itself and its dam, respectively. The maternal heterosis may influence the offspring through providing an environment related to dams. A typical example is that crossbred sows perform better in mother-care than purebred sows, and their offspring also show better performance than offspring of purebred sows (Van Arendonk et al., 2010). This is one of the reasons why fattening pigs usually have a crossbred mother.

In livestock and plant breeding, dominance has been widely considered as one of the main genetic basis for heterosis (Davenport, 1908; Bruce, 1910; Falconer and Mackay, 1996; Visscher et al., 2000). Dominance is a form of phenotypic robustness to mutations and a consequence of the behavior of the multi-enzyme systems (Bagheri and Wagner, 2004). In terms of gene action, interactions between alleles at the same QTL cause dominance (Su et al., 2012a). Through suppressing expression of the recessive deleterious alleles inherited from one parent and increasing expression of the dominant alleles inherited from another parent in the heterozygous loci, dominance contributes to the phenomenon of heterosis (Bruce, 1910; Jones, 1917). More specific, Falconer and Mackay (1996) formulated that, when parental populations are in Hardy-Weinberg equilibrium and random mating holds between the sires and dams, the heterosis in one locus in the F1 crossbred animals is equal to the dominance effect multiplied by the square of the differences of allele frequencies between the parental populations. Data analysis demonstrated that heterosis is mostly positive (Anous and Mourad, 1993; Mavrogenis, 1996; Shikano and Taniguchi, 2002). However, if the dominance effect in the heterozygous loci is negative, heterosis may be negative. Negative heterosis means the performance of the heterozygotes is lower than the average of both homozygotes (Hedgecock et al., 1995). With many loci, the overall heterosis is the effect summed over all the interactions between alleles, both due to dominance and epistasis (Falconer and Mackay, 1996). However, epistasis is commonly considered as playing a secondary or minor role in heterosis (Luo et al., 2001; Li et al., 2008) and the epistasis effects are ignored in this thesis.

According to studies, accuracies of estimated breeding values increase if dominance effects are included in the animal breeding genetic evaluation models (De Boer and Hoeschele, 1993; Zeng et al., 2013; Moghaddar et al., 2014; Sun et al., 2014). Theoretically, estimating dominance genetic effects accurately is beneficial for estimating allele substitution effects and improving the accuracies of estimated breeding values in genomic prediction (Toro and Varona, 2010). Compared with additive genetic variances, although dominance

variations are smaller, they are non-negligible (Misztal, 1997; Esfandyari, 2016). For purebred populations, dominance variation is expected to account for around 10% of total genetic variation (Toro and Varona, 2010), although it varies dramatically depending on the traits and species. In crossbred animals, dominance variation is expected to be larger than that in purebred animals (Su et al., 2012a) and the inclusion of dominance effects in the model is expected to yield higher accuracies in genetic evaluation for crossbred performance than for purebred performance (Lo et al., 1997; Su et al., 2012a). However, dominance effects haven't been frequently used in traditional animal breeding models because it is very difficult to estimate dominant genetic parameters and effects accurately and the computational complexity for the inverse matrix of dominance relationships is high (Henderson, 1985). To estimate the dominance effects, usually, a large amount of dataset including a large ratio of full sibs is required (Misztal et al., 1998), but this is rarely the case.

With the continuously declining costs of genotyping, SNP markers are available for many specials. Estimates of dominant variations and dominance effects become feasible through using SNP markers. Recently, some studies were carried out on dominance effects, but mainly for purebred performances (Su et al., 2012a; Zeng et al., 2013; Ertl et al., 2014; Lopes et al., 2015). Limited number of studies tried to extend it to crossbred performance (Hidalgo, 2015; Esfandyari et al., 2016; Vitezica et al., 2016). For the crossbred performance, they used univariate genomic models, with variance components estimated based on either purebred genomic information or crossbred animals only. However, their conclusions on the needs of including dominance effects explicitly in the model are not consistent. More studies are needed to investigate the role of dominance in crossbred performance.

For modern animal breeding systems, because the earliest known ancestors arose from a common base population (VanRaden, 1992) and high intensity of long-term selection are kept processing (Rauw et al., 1998), all animals within a population are related and inbred to some extent. Inbreeding can be described as deviation of genotypic frequencies towards lower heterozygosity from expected proportions under Hardy-Weinberg. Genealogical inbreeding is due to mating of related individuals. The effect of inbreeding is to increase the number of homozygous loci per animal and increase the frequency of homozygote genotypes in an inbred population (Keller and Waller, 2002). The increased homozygotes lead to higher chances of deleterious alleles becoming homozygous and expressing themselves and the performance of associated traits decreased. This is known as the inbreeding depression (Charlesworth and Charlesworth, 1987). Inbreeding coefficient of an individual is the probability that at a random locus, both alleles are identical by descent (Falconer and Mackay, 1996). Several algorithms exist for calculating pedigree-based inbreeding coefficients (Tier, 1990; Meuwissen and Luo, 1992). With the availability of SNP markers, inbreeding coefficient of an individual can be easily calculated, directly from the genotypes, as the fraction of homozygous markers; for a general overview, see Silio et al. (2014). To distinguish such inbreeding

coefficient from the pedigree based one, SNP-based inbreeding coefficient is termed as "genomic inbreeding coefficient" (Leutenegger et al., 2006).

Genomic selection in (crossbred) pigs

Genomic selection was initially put forward by Meuwissen et al. (2001) and became feasible for the pig industry after the release of the commercial 60K SNP chip in 2009 (Knol et al., 2016). In genomic selection, genome wide distributed SNPs are used to capture the genetic variances for traits and the QTLs are assumed to be in LD with at least one of these SNP markers (Meuwissen et al., 2001). Factors that potentially affect the accuracy of genomic selection are: the density of SNP markers (Calus et al., 2008; Solberg et al., 2008), the LD between SNP and QTLs (Meuwissen et al., 2001; Wientjes et al., 2013), relationships between animals in the reference population and validation population (Habier et al., 2007; Hayes et al., 2009), the size of reference population (Goddard and Hayes, 2007; VanRaden et al., 2009), etc. It has been proved that the reliability in genomic prediction is higher than the pedigree-based prediction (BLUP), but the advantages vary according to the traits and species (Hayes et al., 2009; Tusell et al., 2013).

Genomic selection has been successfully applied in purebred performance (Hayes et al., 2009; Lillehammer et al., 2011), but has been rarely investigated in crossbred performance. Due to the genotype-by-environment interactions, the presence of non-additive genetic effects and the allele frequencies are dissimilar in different breeds (Wei and Steen, 1991; Dekkers, 2007), the genetic correlation of breeding values between purebred and crossbred performances (r_{pc}) is usually lower than 1 (Wei and van der Werf, 1994; Lutaaya et al., 2001). The performance of purebred parents cannot be used to predict the performance of their crossbred offspring accurately when the r_{pc} is considerably lower than 1 (Dekkers, 2007). Ideally, to implement genetic evaluation for crossbred performance, collecting data from both purebred and crossbred animals is required (Wei and van der Werf, 1994). However, due to the complexities of collecting pedigree information and high costs of collecting phenotypes from crossbred animals, it is rare to have access to both purebred and crossbred information.

Several studies have been carried out on crossbred performance by using genomic selection. Dekker (2007) found crossbred selection resulted higher genetic gains in crossbred performance and lower rates of inbreeding than purebred selection or combined purebred and crossbred selection. They concluded that estimates of marker effects on crossbred performance enable an effective genomic selection for crossbred performance.

Ibáñez-Escriche et al. (2009) investigated crossbred performance by using phenotypes and SNP genotypes from crossbred animals and then applied the estimated SNP effects to genotypes obtained from purebred animals to predict their crossbred breeding values. Additionally, because effects of SNP markers may be

breed specific, they applied a model with breed-specific SNP effects to fit crossbred phenotypes. However, they concluded that when the marker density was high and purebred lines were closely related, a model that fit breed-specific SNP effects did not perform better than a model that fit across-breed SNP effects.

These conclusions were further confirmed in real dataset by Lopes (2016). He found accuracies of estimated breeding values from a model with breed-specific effects were equal to or only slightly higher than those from a model with across-breed effects, but prediction of crossbred sows was more accurate when training population consisted of crossbred animals rather than purebred animals.

The above mentioned approaches require collecting genotypes and phenotypes from crossbred animals, which is costly. Esfandyari et al. (2015) explored the possibilities of improving crossbred performance by using reference population consisted of only purebred animals. They compared the accuracy to that obtained by using a reference population consisting of only crossbred animals. However, results showed that to optimize the genomic selection on purebred animals for their crossbred performance, marker effects were better estimated based on crossbred data than on purebred data.

Hidalgo et al. (2016) also compared accuracies of genomic selection for crossbred performance by using reference populations consisting of either only crossbred or only purebred animals. They found that it is possible to predict crossbred performance by using crossbred training data alone, but the accuracies were lower than those from purebred training data, which were opposite to the above mentioned results. Nevertheless, Hidalgo et al. (2016) attributed their results to the data structure that (1) the reference size of crossbred population was small; (2) the low relationships between purebred and crossbred animals; (3) the high r_{pc}(>0.90) lead to favor the purebred training population; and thus they announced that the results were not general.

All of the studies above commonly indicate that genomic selection offers opportunities for selecting purebreds for crossbred performance in pigs. All these studies used genomic models that assume that all animals belong to a single population, and variance components were only estimated based on either the genotyped purebred or crossbred animals. Wei and Van der Werf (1994) multiple-trait model was more sophisticated than these single-trait models, because not all the phenotypic records were used in these single-trait models and the potential genotype by environment or genotype by genotype interactions cannot be accounted as well in the single-trait models as in the multiple-trait model. Therefore, on the basis of the Wei and Van der Werf model, Christensen et al. (2014) incorporated genomic information to the model and extended the marker-based relationship matrices to the non-genotyped animals. This model can evaluate both purebred and crossbred performances simultaneously and integrate all the genomic and phenotypic information, which is a breakthrough. This method is an extension of a single-step BLUP method (Legarra et al., 2009; Christensen and Lund, 2010) from purebred performance to combined purebred and crossbred

performances. This method had not been evaluated in real dataset before this thesis, although Tusell et al. (2016) successfully applied a simplified Christensen model, which is also an extension of Wei and Van der Werf model, but only including purebred genotypes, in crossbred performance.

Single-step GBLUP method

Approaches of genomic selection generally require that all the involved individuals are genotyped (Meuwissen et al., 2001; VanRaden, 2008), which is currently unfeasible due to the restriction of high cost and practical constraints (Legarra et al., 2009). There is often no phenotype for the genotyped individual and vice versa. Legarra et al. (2009) and Christensen and Lund (2010) in parallel proposed a genomic evaluation method - "single-step genomic BLUP", which can handle the situation where only parts of animals are genotyped. The method applies an integrated relationship matrix (**H** matrix) for all animals by blending the information of pedigree and genomic markers:

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} + \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{G} \\ \mathbf{G}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} \end{bmatrix},$$

where **A** is the pedigree-based numerator relationship matrix, and matrices $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, $\mathbf{A}_{22}$ are submatrices of **A** matrix; subscript 1 and 2 indicated non-genotyped and genotyped individuals, respectively.

In the traditional BLUP method, pedigree-based numerator relationship matrix (**A** matrix) accounts for the family relationships, while in the classical GBLUP method, marker-based genomic relationship matrix (**G** matrix) accounts for relationships among all the genotyped individuals. GBLUP method performs better than traditional BLUP because the **G** matrix is an improved estimator of true relationships among the genotyped individuals compared to the **A** matrix (VanRaden, 2008). If all the involved animals are genotyped, the single-step method is the same as GBLUP method; if none of animals are genotyped, the single-step method becomes the traditional BLUP method. This method is termed as "single-step" method because it can handle both genotyped and non-genotyped animals at once. The influence of marker genotypes from genotyped animals on the non-genotyped animals is via the numerator relationships (Legarra et al., 2014). Alternatively, in a multiple-step genomic evaluation, usually, traditional genetic evaluation is processed as the first step and the estimated breeding values are used to create pseudo phenotypes (e.g: deregressed proofs) for the genomic evaluation step (Mäntysaari et al., 2011). Such multiple-step method could cause loss of information and lead to inaccuracies and biases (Legarra et al., 2014), which can be avoided by using single-step method (Aguilar et al., 2010; Forni et al., 2011).

Single-step method was shown to have the power of producing at least the same accuracies of GEBVs as GBLUP for genotyped pigs, but higher accuracies for estimating the EBVs in non-genotyped pigs than pedigree-based method (Christensen et al., 2012; Guo et al., 2014). Therefore, pig breeding companies Topigs Norsvin and PIC nowadays use single-step method as the standard tool for routine genetic evaluation.

Also, since October 2011, a routine single-step genomic evaluation system has been set up in DanAvl in Denmark. Single-step genomic evaluation was also tested in other species, such as dairy cattle (Gao et al., 2012), dairy sheep (Baloche et al., 2014) and broiler chicken (Simeone et al., 2012). However, all of these studies were carried out on purebred performance.

As mentioned above, crossbred performance is vital in pig industry, but the existing genomic selection methods for crossbred performance commonly require genotyping all the pigs. Therefore, Christensen et al. (2014) developed a single-step method for genomic evaluation of both purebred and crossbred performance in a two-way crossbreeding system. In this novel approach, two breed-specific partial relationship matrices (García-Cortés and Toro, 2006) across purebred animals and crossbred gametes were used. This approach was later termed as "partial genetic approach" by Christensen et al. (2015). Later, Christensen et al. (2015) further developed single-step method for three-way crossbred animals. Genomic evaluation for the three-way crossbred animals used either partial genetic approach or a "common genetic approach", which was developed based on a new concept of metafounder (Legarra et al. 2015). Overall, single-step genomic method is theoretically available for both purebred and crossbred performance.

Metafounder

Christensen (2012) summarized two issues raised by the single-step genomic evaluation method. First, how to choose the allelic frequencies used in the genomic relationship matrix **G**. Theoretically, matrix **G** depends crucially on the assumed base allelic frequencies (Toro et al., 2011). The allelic frequencies in the base population of the pedigree should be used (VanRaden, 2008), but these frequencies are rarely available (Christensen, 2012). The assumed base allelic frequencies change as the start of pedigree changes arbitrarily (Legarra et al., 2014). Second, single-step GBLUP requires that both genomic and pedigree-based relationship matrices (**G** and **A** respectively) refer to the same (base) population so that the average of breeding values and genetic variances of base populations are comparable across the different measures of relationships (Legarra, 2016), but this is difficult in practice (Legarra et al., 2015). Thus, the second problem is how to achieve the compatibility between genomic relationships and pedigree-based relationships.

To overcome these two issues, Christensen (2012) modelled the likelihood of genotypes given the pedigree as a quantitative trait and then he marginalized (integrated out) the allelic frequencies from this likelihood. This resulted in an arbitrary reference for genomic relationship matrix with the allelic frequency fixed at 0.5 and referring the pedigree-based relationships to this base population. As a byproduct, the founders of base population become related, described by a parameter γ that determines the relationship and inbreeding of individuals in the base population. The assumption that founders of base populations are related is in accordance with the observed relatedness across base individuals based on the marker information from

genotyped animals (Ter Braak et al., 2010; VanRaden et al., 2011), although this assumption is contradictory to the usual assumption that base animals are unrelated. Then, instead of inferring the base allelic frequencies, ssGBLUP was altered to infer γ and another scaling parameter s, which can be interpreted as a counterpart of the heterozygosity of the markers in the base population (García-Baccino et al.,2017), that makes pedigree-based and genomic relationship matrices compatible (Christensen, 2012). García-Baccino et al. (2017) concluded that in a single population, methods of generalized least squares and maximum likelihood can efficiently estimate an unbiased γ in single-step approach with one metafounder. Overall, contrary to the usual adjustment of genomic relationships to be compatible with pedigree-based relationships (Vitezica et al., 2011; Christensen et al., 2012), the pedigree-based relationship matrix needs to be adjusted to be compatible with marker-based matrix.

On the basis of Christensen (2012), Legarra et al. (2015) showed an equivalent idea about relationships within and across base populations as Christensen (2012) and developed a new concept of metafounders. A metafounder is a generalization of unknown parent groups (Kennedy, 1991) and can be understood as a finite-size pool of gametes, from which the founders of pedigree are drawn. The advantages of using the concept of metafounder over Christensen's idea (2012) are (1) With metafounders added to the pedigree, regular methods to build and invert pedigree relationships in the new matrix $\mathbf{A}(\gamma)$ can be used with only minor modifications, and $\mathbf{A}(\gamma)$ is still a very sparse matrix; (2) genomic relationships and pedigree-based relationships are automatically compatible by construction if γ is estimated from marker genotypes (Christensen, 2012); (3) when analyzing multiple populations simultaneously, the concepts of metafounders can be easily extended from one to multiple base populations with relationships $\boldsymbol{\Gamma}$ (a matrix instead of a scalar) across base populations. In short, the two mentioned issues in ssGBLUP can be resolved conveniently by the use of 0.5 allelic frequencies and metafounders.

Simulation studies on purebred performance showed that ssGBLUP with metafounders performed more accurate and less biased than the regular ssGBLUP method (Christensen, 2012; García-Baccino et al., 2017), but how genomic evaluation with metafounders perform in real data and for several populations and crosses was unknown before the thesis.

Genotype imputation

Genotype imputation is defined as the prediction of genotypes that are not genotyped (Marchini and Howie, 2010). Although the development of genotyping technologies has made it feasible to genotype animals in large scale, genotyping is still costly. Thus, to reduce the cost of genotyping, a possible way is to genotype a large number of animals by low-density SNP chips, while a limited number of individuals are genotyped with a high-density chip and regarded as a reference population for imputation. Then imputation is processed from low density to high density (Habier et al., 2009).

Imputation is generally considered as an initial step for genomic selection. As mentioned above, accuracies of genomic selection are affected by marker density and reference population size. Imputation can improve the call rate for SNP markers and individuals and thus it improves accuracies of genomic selection (Su et al., 2012b). Moreover, imputing missing genotypes that are not called by genotyping techniques is also required prior to genomic selection (Hickey et al., 2012a). For crossbred performance, genotypes come from multiple breeds and populations and sometimes are obtained from different chips. To combine different datasets, imputation should be used to infer those missing genotypes and it is essential to get same amount of markers for animals in different populations (Ma, 2013).

Several kinds of software have been developed for genotype imputation. Among them, some methods are based on the construction of a library of inferred haplotypes in the reference populations, and then the missing markers in the imputed animals are filled in according to the existing markers aligning to the inferred haplotypes (Ma et al., 2013). These methods depend crucially on local LD pattern across markers. Software based on such algorithms are e.g: Beagle (Browning and Browning, 2009), IMPUTE2 (Howie et al., 2009) and fastPHASE (Scheet and Stephens, 2006). Other imputation software rely on the combination of these libraries with the use of pedigree information, such as AlphaImpute (Hickey et al., 2012b) or FImpute (Sargolzaei et al., 2011). Ma et al. (2013) compared those software in imputing genotypes in dairy cattle and concluded that Beagle and IMPUTE2 are most accurate and robust for imputing genotypes from low density panel (3K) to moderate density panel (54K).

To measure the accuracies of imputation, most studies used correct rates between imputed and true genotypes, which is defined as the proportion of correctly imputed alleles (Zhang and Druet, 2010; Brøndum et al., 2012). However, this accuracy is allele-frequency dependent and favors markers with low minor allele frequencies. Therefore, correlation coefficients between true genotypes and imputed ones, which do not suffer these problems, are considered as a better way of measuring imputation accuracies (Hickey et al., 2012a).

It is rare to genotype crossbred animals, especially with moderate or high density chips. In this thesis, 8K SNP marker genotypes in crossbred pigs and 60K SNP marker genotypes in purebred pigs were provided by Danish Pig Research Centre. To implement genomic prediction for crossbred performance, imputation from low density (8K) to moderate density (60K) in crossbred animals is needed. However, the extent of LD between SNP markers and QTL may differ between crossbred and purebred populations (Dekkers, 2007). Thus, when using an imputation method based on local phasing haplotypes (e.g: Beagle), the performance of genotype imputation may be different in crossbred and purebred populations. For instance, Ventura et al. (2014) reported that in taurine beef cattle, accuracies of imputation for purebreds was much higher than those for crossbreds, but this has not been evaluated in crossbred pigs before this thesis.

The trait: Total Number of Piglets Born

Litter size is considered as a vital reproduction trait in pig production (Guo et al., 2014). It depends on both the ovulation rate of the dam and the embryonic survival of the offspring (Johnson et al., 1999). Lund et al. (2002) stated that litter size at weaning is the most important reproduction trait in pig production because of its high economic importance. However, due to cross fostering in pig farms, it is difficult to measure litter size at weaning. Instead, litter size at birth or the total number of piglets born (TNB), has been used as an alternative (Lund et al., 2002). The TNB in the first parity is considered as a different trait from later parities, because the genetic correlations between first and later parities are significantly lower than 1 (Irgang et al., 1994; Hanenberg et al., 2001). The trait of TNB is well known as a lowly heritable trait. Rothschild et al. (1998) reviewed that the average heritabilities were about 0.11 for TNB in both Landrace and Yorkshire populations, in line with results from Nielsen et al. (2013). Guo et al. (2014) also reported heritability of 0.11 for Landrace, but 0.09 for Yorkshire. Su et al. (2007) reported even lower heritablities for Landrace (0.07) and Yorkshire (0.05). For traits with low heritability, conventional selection cannot efficiently increase the genetic gain, which leaves room for the application of genomic selection. Genomic selection increases the genetic gain by increasing the accuracies of EBVs in pigs. The genetic gain in purebred populations was reported to increase up to 55% when compared to the conventional selection (Lopes, 2016).

Outline of this PhD thesis

This thesis aims at investigating genomic evaluation in pigs for crossbred performance for TNB. More specifically, the aims of this PhD project are:

- First, apply single-step genomic evaluation method for crossbred performance in different scenarios with data recordings and genotypes in Danish Landrace, Yorkshire and F1 crossbred pig populations;
- Second, investigate the impact of non-additive genetic effects on improvement of genomic evaluation for crossbred performance.

In chapter 2, performance of genotype imputation in low density panels were compared using different imputation strategies in both purebred and crossbred populations. Based on the optimal strategy, imputation from low density to moderate density panels was then investigated by a pedigree-based simulated dataset. This chapter demonstrated that imputation for crossbreds work as well as for purebreds. After these steps, genotypes were available at the same density for all the three populations, and for the further genomic selection.

In chapters 3 and 4, the single-step GBLUP method was applied in both purebred and crossbred datasets, focusing on evaluating genetic ability for crossbred performance. In chapter 3, a three-trait animal model that

can incorporate marker genotypes was applied to investigate both purebred and crossbred performances in different scenarios. Additive genetic effects in crossbred animals were split into two breed-specific gametic effects. Two breed-specific partial relationship matrices were used to account for the family relationships across purebred breed animals and breed-specific purebred gametes in crossbred animals. This method required estimating the breed origin of crossbred alleles, a difficult task. Therefore, the same dataset was revisited in chapter 4 by using a similar animal model, but on the contrary, instead of using two relationship matrices, one relationship matrix with metafounders was used to relate all the involved animals in the three populations. This method did not need the tracing of crossbred alleles for the breed origins.

Non-additive genetic effects play an important role in crossbred animals, but they cannot be implemented by single-step GBLUP approaches yet. In chapter 5, joint genomic evaluation of purebreds and crossbreds with GBLUP including additive genetic effects, dominance genetic effects and inbreeding depression was investigated on genotyped animals using precorrected data.

Finally, in chapter 6, a general discussion on the findings in a broad context was presented. Perspectives of shortages of this thesis and possible future improvements were also discussed.

References:

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93: 743-52.

Anous, M., and M. Mourad. 1993. Crossbreeding effects on reproductive traits of does and growth and carcass traits of kids. Small Ruminant Research 12: 141-9.

Bagheri, H. C., and G. P. Wagner. 2004. Evolution of dominance in metabolic pathways. Genetics 168: 1713-35.

Baloche, G., A. Legarra, G. Sallé, H. Larroque, J-M. Astruc, C. Robert-Granié, and F. Barillet. 2014. Assessment of accuracy of genomic prediction for French Lacaune dairy sheep. Journal of Dairy Science 97: 1107-16.

Browning, B. L., and S. R. Browning. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics 84: 210-23.

Bruce, A. 1910. The Mendelian theory of heredity and the augmentation of vigor. Science 32: 627-8.

Brøndum, R. F., P. Ma, M. S. Lund, and G. Su. 2012. Short communication: Genotype imputation within and across Nordic cattle breeds. Journal of Dairy science 95: 6795-6800.

Calus, M., A. De Roos, and R. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178: 553-61.

Charlesworth, D., and B. Charlesworth. 1987. Inbreeding depression and its evolutionary consequences. Annual review of ecology and systematics 18: 237-68.

Christensen, O. F. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genetics Selection Evolution 44: 37.

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

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

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal 6: 1565-71.

Comstock, R. E., H. Robinson, and P. Harvey. 1949. Breeding procedure designed to make maximum use of both general and specific combining ability. Agronomy Journal.

Davenport, C. B. 1908. Degeneration, albinism and inbreeding. Science 28: 454-55.

De Boer, I., and I. Hoeschele. 1993. Genetic evaluation methods for populations with dominance and inbreeding. Theoretical and Applied Genetics 86: 245-58.

Dekkers, J. 2007. Marker-assisted selection for commercial crossbred performance. Journal of Animal Science 85: 2104-14.

Dufrasne, M. 2015. Genetic improvement of pig sire lines for production performances in crossbreeding. PhD Thesis, Université De Liège Gembloux, Belgium.

Ertl, J., J. Ertl, A. Legarra, Z. G. Vitezica, L. Varona, C. Edel, R. Emmerling, and K. U. Götz, K. U. 2014. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genetics Selection Evolution 46: 40.

Esfandyari, H. 2016. Genomic selection for crossbred performance. PhD thesis, Wageningen University, the Netherlands.

Esfandyari, H., P. Bijma, M. Henryon, O. F. Christensen, and A. C. Sørensen. 2016. Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. Genetics Selection Evolution 48: 9.

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

Falconer, D., and T. Mackay. 1996. Introduction to Quantitative Genetics. 4 ed. Harlow: Longmans Green. UK.

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43: 1.

Gao, H., O. F. Christensen, P. Madsen, U. S. Nielsen, Y. Zhang, M. S. Lund, and G. Su. 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genetics Selection Evolution 44: 8.

Garcia-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, Z. G. Vitezica, and R. J. Cantet. 2017. Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. Genetics Selection Evolution, 49: 34.

García-Cortés, L. A., and M. Á. Toro. 2006. Multibreed analysis by splitting the breeding values. Genetics Selection Evolution 38: 601-15.

Goddard, M. E., and B. Hayes. 2007. Genomic selection. Journal of Animal Breeding and Genetics. 124: 323-30.

Guo, X., O. F. Christensen, T. Ostersen, Y. Wang, M. S. Lund, and G. Su. 2015. Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. Journal of Animal Science 93: 503-12.

Habier, D., R. Fernando, and J. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389-97.

Habier, D., R. L. Fernando, and J. C. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182: 343-53.

Hanenberg, E., E. Knol, and J. Merks. 2001. Estimates of genetic parameters for reproduction traits at different parities in Dutch Landrace pigs. Livestock Production Science 69: 179-186.

Hartmann, W. 1992. Evaluation of the potentials of new scientific developments for commercial poultry breeding. World's Poultry Science Journal 48: 17-27.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41: 1.

Hedgecock, D., D. J. McGoldrick, and B. L. Bayne. 1995. Hybrid vigor in Pacific oysters: an experimental approach using crosses among inbred lines. Aquaculture 137: 285-98.

Henderson, C. 1985. Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. Journal of Animal Science 60: 111-17.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012a. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Science 52: 654-63.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. van der Werf, and M. A. Cleveland. 2012b. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genetics Selection Evolution 44: 9.

Hidalgo, A., J. Bastiaansen, M. Lopes, M. Calus, and D. Koning. 2016. Accuracy of genomic prediction of purebreds for cross bred performance in pigs. Jounal of Animal Breeding and Genetics 133: 443-51.

Hidalgo, A. M. 2015. Exploiting genomic information on purebred and crossbred pigs. PhD Thesis, Swedish University of Agricultural Sciences, Uppsala, Sweden.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.

Irgang, R., J. A. Fávero, and B. W. Kennedy. 1994. Genetic parameters for litter size of different parities in Duroc, Landrace, and large white sows. Journal of Animal Science 72: 2237-46.

Johnson, R. K., M. K. Nielsen, and D. S. Casey. 1999. Responses in ovulation rate, embryonal survival, and litter traits in swine to 14 generations of selection to increase litter size. Journal of Animal Science 77: 541-57.

Jones, D. F. 1917. Dominance of linked factors as a means of accounting for heterosis. Proceedings of the National Academy of Sciences 3: 310-12.

Keller, L. F., and D. M. Waller. 2002. Inbreeding effects in wild populations. Trends in Ecology & Evolution 17: 230-41.

Kennedy, B. 1991. CR Henderson: The unfinished legacy. Journal of dairy science 74: 4067-81.

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

Legarra, A. 2016. Comparing estimates of genetic variance across different relationship models. Theoretical population biology 107: 26-30.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. Journal of dairy science 92: 4656-63.

Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livestock Science 166: 54-65.

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics 200: 455-68.

Leutenegger, A. L., A. Labalme, E. Génin, A. Toutain, E. Steichen, F. Clerget-Darpoux, and P. Edery. 2006. Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to Taybi-Linder syndrome. The American journal of human genetics 79: 62-66.

Li, L., K. Lu, Z. Chen, T. Mu, Z. Hu, and X. Li. 2008. Dominance, overdominance and epistasis condition the heterosis in two heterotic rice hybrids. Genetics 180: 1725-42.

Lillehammer, M., and A. K. Sonesson. 2011. Genomic selection for maternal traits in pigs. Journal of Animal Science 89: 3908-16.

Lo, L., R. Fernando, and M. Grossman. 1997. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. Journal of Animal Science 75: 2877-84.

Lopes, M. 2016. Genomic selection for improved crossbred performance. PhD Thesis, Wageningen University, Wageningen, the Netherlands.

Lopes, M., J. Bastiaansen, L. Janss, E. Knol, and H. Bovenhuis. 2015. Genomic prediction of growth in pigs based on a model including additive and dominance effects. Journal of Animal Breeding and Genetics 133: 180-86.

Lund, M. S., M. Puonti, L. Rydhmer, and J. Jensen. 2002. Relationship between litter size and perinatal and pre-weaning survival in pigs. Animal Science 74: 217-22.

Luo, L. J., Z. K. Li, H. W. Mei, Q. Y. Shu, R. Tabien, D. B. Zhong, and A. H. Paterson. 2001. Overdominant epistatic loci are the primary genetic basis of inbreeding depression and heterosis in rice. II. Grain yield components. Genetics 158: 1755-71.

Lutaaya, E., I. Misztal, J. W. Mabry, T. Short, H. H. Timm, and R. Holzbauer. 2001. Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. Journal of Animal Science 79: 3002-07.

Ma, P. 2013. Methods and strategies to impute missing genotypes for improving genomic prediction. PhD thesis, Aarhus University, Tjele, Denmark.

Ma, P., R. F. Brøndum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. Journal of dairy science 96: 4666-77.

Marchini, J., and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nature Reviews Genetics 11: 499-511.

Mavrogenis, A. 1996. Environmental and genetic factors influencing milk and growth traits of Awassi sheep in Cyprus. Heterosis and maternal effects. Small Ruminant Research 20: 59-65.

Meuwissen, T., B. Hayes, and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-29.

Meuwissen, T., and Z. Luo. 1992. Computing inbreeding coefficients in large populations. Genetics Selection Evolution 24: 4.

Misztal, I. 1997. Estimation of variance components with large-scale dominance models. Journal of Dairy Science 80: 965-74.

Misztal, I., L. Varona, M. Culbertson, J. K. Bertrand, J. Mabry, T. J. Lawlor, and N. Gengler. 1998. Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. Biotechnologie, agronomie, société et environnement 2: 227-33.

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

Mäntysaari, E. A., M. Koivula, I. Strandén, J. Pösö, and G. P. Aamand. 2011. Estimation of GEBVs using deregressed individual cow breeding values. Interbull Bulletin 44.

Nielsen, B., G. Su, M. S. Lund, and P. Madsen. 2013. Selection for increased number of piglets at d 5 after farrowing has increased litter size and reduced piglet mortality. Journal of Animal Science 91: 2575-82.

Rauw, W., E. Kanis, E. Noordhuizen-Stassen, and F. Grommers. 1998. Undesirable side effects of selection for high production efficiency in farm animals: a review. Livestock Production Science 56: 15-33.

Rothschild, M., J. Bidanel, and A. Ruvinsky. 1998. Biology and genetics of reproduction. The genetics of the pig 2: 313-43.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2011. FImpute-An efficient imputation algorithm for dairy cattle populations. Journal of Dairy Science 94: 421.

Scheet, P., and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics 78: 629-44.

Shikano, T., and N. Taniguchi. 2002. Using microsatellite and RAPD markers to estimate the amount of heterosis in various strain combinations in the guppy (Poecilia reticulata) as a fish model. Aquaculture 204: 271-81.

Simeone, R., I. Misztal, I. Aguilar, and Z. Vitezica. 2012. Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. Journal of Animal Breeding and Genetics 129: 3-10.

Silió, L., M. C. Rodríguez, A. Fernández, C. Barragán, R. Benítez, C. Óvilo, and A. I. Fernández. 2013. Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. Journal of Animal Breeding and Genetics 130: 349-60.

Solberg, T., A. Sonesson, and J. Woolliams. 2008. Genomic selection using different marker types and densities. Journal of Animal Science 86: 2447-54.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund. 2012a. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLoS ONE 7: e45293.

Su, G., M. S. Lund, and D. Sorensen. 2007. Selection for litter size at day five to improve litter size at weaning and piglet survival rate. Journal of Animal Science 85: 1385-92.

Su, G., P. Madsen, U. S. Nielsen, E. A. Mäntysaari, G. P. Aamand, O. F. Christensen, and M. S. Lund. 2012b. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. Journal of Dairy Science 95: 909-17.

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

Sørensen, P. 2003. DENMARK'S Country Report on Farm Animal Genetic Resources. Danish Institute of Agricultural Sciences, Ministry of Food, Agriculture and Fisheries in Denmark.

Ter Braak, C. J., M. P. Boer, L. R. Totir, C. R. Winkler, O. S. Smith, M. C. Bink. 2010. Identity-by-descent matrix decomposition using latent ancestral allele models. Genetics 185: 1045-57.

Tier, B. 1990. Computing inbreeding coefficients quickly. Genetics Selection Evolution 22: 419-30.

Toro, M. Á., L. A. García-Cortés, and A. Legarra. 2011. A note on the rationale for estimating genealogical coancestry from molecular markers. Genetics Selection Evolution 43: 27.

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

Tusell, L., H. Gilbert, J. Riquet, M. J. Mercat, A. Legarra, and C. Larzul. 2016. Pedigree and genomic evaluation of pigs using a terminal-cross model. Genetics Selection Evolution 48: 32.

Tusell, L., P. Pérez-Rodríguez, S. Forni, X.-L. Wu, and D. Gianola. 2013. Genome-enabled methods for predicting litter size in pigs: a comparison. Animal 7: 1739-49.

Van Arendonk, J., P. Bijma, H. Bovenhuis, R. Crooijmans, and T. Van der Lende. 2010. Animal breeding and genetics. Lecture notes ABG-20306, Wageningen University, the Netherlands.

VanRaden, P. 1992. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. Journal of Dairy Science 75: 3136-44.

VanRaden, P. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414-23.

VanRaden, P., K. Olson, G. Wiggans, J. Cole, and M. Tooker. 2011. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science 94: 5673-82.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92: 16-24.

Ventura, R. V., D. Lu, F. S. Schenkel, Z. Wang, C. Li, and S. P. Miller. 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. Journal of Animal Science 92: 1433-44.

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

Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genetics Research 93: 357-66.

Vitezica, Z. G., L. Varona, J. M. Elsen, I. Misztal, W. Herring, and A. Legarra. 2016. Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genetics Selection Evolution 48: 6.

Wei, M. 1992. Combined crossbred and purebred selection in animal breeding. PhD Thesis, Wageningen University and Research Centre, The Netherlands.

Wei, M., and H. A. M. Steen. 1991. Comparison of reciprocal recurrent selection with pure-line selection systems in animal breeding (a review). Anim Breed Abstr 59: 281-98.

Wei, M., and J. van der Werf. 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim Prod 59: 401-13.

Weigel, K. A., G. de Los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of Dairy Science 93: 5423-35.

Wientjes, Y. C., R. F. Veerkamp, and M. P. Calus. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193: 621-31.

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

Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. Journal of Dairy Science 93: 5487-94.

CHAPTER 2: PAPER I.

Imputation of genotypes in Danish purebred and two-way crossbred pigs using low-density panels

Tao Xiang[1,2*], Peipei Ma[1], Tage Ostersen[3], Andres Legarra[2], Ole Fredslund Christensen[1]

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark

[2]INRA, UR1388 GenPhySE, CS-52627, F-31326 Castanet-Tolosan, France

[3]Pig Research Centre, Danish Agricultural and Food Council, DK-1609 Copenhagen, Denmark

[*]Corresponding author

This paper was published in *Genetics Selection Evolution* (2015) 47:54. DOI 10.1186/s12711-015-0134-4

Abstract

Background

Genotype imputation is commonly used as an initial step in genomic selection since the accuracy of genomic selection does not decline if accurately imputed genotypes are used instead of actual genotypes but for a lower cost. Performance of imputation has rarely been investigated in crossbred animals and, in particular, in pigs. The extent and pattern of linkage disequilibrium differ in crossbred versus purebred animals, which may impact the performance of imputation. In this study, first we compared different scenarios of imputation from 5K to 8K single nucleotide polymorphisms (SNPs) in genotyped Danish Landrace and Yorkshire and crossbred Landrace-Yorkshire datasets and, second, we compared imputation from 8K to 60K SNPs in genotyped purebred and simulated crossbred datasets. All imputations were done using software Beagle version 3.3.2. Then, we investigated the reasons that could explain the differences observed.

Results

Genotype imputation performs as well in crossbred animals as in purebred animals when both parental breeds are included in the reference population. When the size of the reference population is very large, it is not necessary to use a reference population that combines the two breeds to impute the genotypes of purebred animals because a within-breed reference population can provide a very high level of imputation accuracy (correct rate ≥ 0.99, correlation ≥ 0.95). However, to ensure that similar imputation accuracies are obtained for crossbred animals, a reference population that combines both parental purebred animals is required. Imputation accuracies are higher when a larger proportion of haplotypes are shared between the reference population and the validation (imputed) populations.

Conclusions

The results from both real data and pedigree-based simulated data demonstrate that genotype imputation from low-density panels to medium-density panels is highly accurate in both purebred and crossbred pigs. In crossbred pigs, combining the parental purebred animals in the reference population is necessary to obtain high imputation accuracy.

Background

Implementation of genomic selection (GS) [1] in breeding programs requires dense molecular marker genotypes since increasing marker density increases the probability that a marker is in strong linkage disequilibrium (LD) with a quantitative trait locus (QTL) [2]. However, the high costs of genotyping are a key constraint to efficient implementation of GS [3]. To partly overcome this problem, it has become current practice to genotype candidates for selection using low-density single nucleotide polymorphism (SNP) chips (up to 10 000 SNPs), while a limited number of individuals chosen as reference animals are genotyped with a high-density chip (50 000 SNPs or more). Imputation is then carried out from low density to high density [4,5]. Studies on US Jersey cattle have confirmed that the accuracy of GS does not decline when using imputed genotypes if the low-density panel includes more than 3000 evenly distributed SNPs [4]. Furthermore, missing genotypes that are not called by some of the standard genotyping methods must be imputed prior to inclusion in models for GS [6]. Overall, genotype imputation is generally considered as an initial step for GS.

Genomic selection has been successfully applied for purebred populations [7,8], but it is also possible to select purebred animals for crossbred performance by combining information from crossbred animals with genomic information from purebred animals [9]. Crossbreeding is very common in pigs bred for meat production because of the increased performance of crossbred compared to purebred individuals [10]. Due to the difficulty and high cost of collecting phenotypic and pedigree data on crossbred animals [11] and genotyping costs, data on both purebred and crossbred animals are rarely available. Performances of GS in crossbred and purebred pigs may differ because of dominance effects in combination with different allele frequencies in the two pure breeds, and because the extent of LD between SNPs and QTL may differ between crossbred and purebred populations. Thus, the effects of SNPs may be breed-specific [9].

Algorithms for genotype imputation (such as that implemented in Beagle [12]) depend crucially on LD patterns across markers [13], which may be breed-specific. Therefore, the performance of genotype imputation might differ between crossbreds and purebreds. Since genotypes are rarely available for crossbred individuals in livestock, most studies that have investigated the critical factors that affect the performance of imputation have been based on purebred plant [14] and livestock populations [15-18]. Recently, an analysis of imputation from 6K to 50K SNP chip genotypes in crossbred taurine beef cattle was reported [19], but, to our knowledge, this has not been evaluated in crossbred pigs.

In this study, different scenarios of imputation from lower density (5K) to higher density (8K) SNP chips were compared using two Danish pig breeds, Landrace and Yorkshire, and a two-way crossbred Landrace-Yorkshire population. Differences in imputation accuracies between purebred and crossbred animals were

investigated to set up an optimal strategy for imputation from a low-density (8K) to a medium-density (60K) SNP chip in crossbred pigs and results were validated using a simulated dataset of crossbred medium-density (60K) genotypes. Previous studies indicated that the relationship between imputed and reference individuals is one of the major factors that affects performance of imputation [3,6,20]; Hayes et al. [3] reported that it could account for up to 64% of the variation in accuracy of imputation in sheep. Thus, to better understand the results in the current study, we tried to quantify relationships between animals within and across datasets, using genomic relationships and indexes of haplotype similarities.

Methods

Animals and genotypes

All data were provided by the Danish Pig Research Centre. The numbers of genotyped purebred Danish Landrace (LL), Danish Yorkshire (YY) and two-way crossbred Danish Landrace-Yorkshire pigs were 9328, 9393 and 5639, respectively. Crossbred animals that had a Landrace sire and a Yorkshire dam were referred to as 'Landrace_Yorkshire', while those that had a Yorkshire sire and a Landrace dam were referred to as 'Yorkshire_Landrace'. Crossbred animals consisted of 4432 Landrace_Yorkshire (LY) and 1207 Yorkshire_Landrace (YL) pigs. Purebred and crossbred animals were born between 1998 and 2013, and between 2009 and 2012, respectively. All crossbred pigs were results of matings between the two pure breeds. Pedigrees of both purebred and crossbred pigs were available and all crossbred animals could be traced back to their purebred ancestors. Among the 5639 crossbred pigs, 4956 had genotyped sires (n = 1580) but only nine pigs had genotyped dams (n = 4). In addition, 1441 maternal grandsires of the crossbreds were genotyped. Crossbred animals were divided into two subsets: those that had a genotyped sire (4956) and those that did not (683).

Both pure breeds were genotyped with the Illumina PorcineSNP60 Genotyping BeadChip [21]. Two different versions of the 60K SNP chip (Illumina PorcineSNP60 v1 and PorcineSNP60 v2) were used to genotype purebred animals, i.e. about 50% animals with each version. About 2% of the SNPs worked in one version but not in the other version and vice versa. The two different chip versions should be taken into account when applying a quality filter on SNPs. Previous unpublished analyses (Tage Ostersen, Danish Pig Research Centre, personal communication) on purebred pigs showed that when applying a quality filter on SNPs, varying the minimum call rate for individuals from 70 to 90% did not affect the accuracy of genomic predictions significantly. This combined with the fact that very few animals had a call rate between 80% and 90%, we chose to set the minimum call rate of individuals to 80%. SNP quality controls were applied for the dataset that consisted of both pure breeds combined as follows: SNPs with a call rate less than 90% were

removed; SNPs with a minor allele frequency lower than 0.01 across both purebred populations were removed; SNPs that showed a strong deviation from Hardy Weinberg equilibrium within breeds ($p < 10^{-7}$) were also excluded. After filtering, a common set of 42 483 SNPs was retained for the two purebred populations (these are referred to as 60K). Crossbred individuals were genotyped with a 8.5K GGP-Porcine Low Density Illumina Bead SNP chip [22] and very few animals had a call rate between 80% and 90%. Using the same quality controls for the crossbred animals as for purebred animals (except for Hardy-Weinberg equilibrium, which does not hold for crossbred animals), 7940 markers were retained, which represents a subset of the 42 483 SNPs retained for the purebred animals. SNPs were mapped to pig chromosomes using the pig genome build 10.2 [23].

Imputation scenarios

To mimic an imputation strategy similar to what is routinely applied in real genetic evaluations, 5162 LL and 5130 YY pigs that were born in 2012 and 2013 were used as validation animals. The remaining 4166 LL and 4263 YY pigs that were born before 2012 were used as reference animals for imputation. All 5639 crossbred pigs were treated as validation animals. Based on pedigree information, the parents of the crossbred animals were all born before 2012. Thus, if the parental genotypes of the crossbred individuals were known, they were included in the reference population.

To compare the performance of imputation between purebred and crossbred animals, first imputation from 5K to 8K was evaluated, which was applied to the common set of 7940 SNPs. SNPs were sorted by map position and then, one of every three SNPs was masked (i.e. 2647 SNPs were masked) and the remaining SNPs were retained to represent the lower density panel (5K). To ensure consistency of imputation results, this was repeated three times by shifting the masked SNPs by one position each time. For the purebred populations, imputations were first done by using one of the pure breeds as reference population, which consisted of individuals that were either from their own breed (within-breed scenario) or the other pure breed (external-breed scenario), i.e., we imputed Landrace animals using Yorkshire animals as the reference population and vice versa. Then, each breed was imputed by a combined Landrace and Yorkshire population (combined-breed scenario). Finally, for the crossbred population, imputation was done by using either a single purebred reference population (one of the two pure breeds) or a combined Landrace and Yorkshire population (4166 LL + 4263 YY). In order to eliminate the effect of population size of the reference panel, its size was fixed to 8429 animals for all scenarios of imputation of crossbred animals. Thus, when only one purebred reference population was used, it had to also contain animals that were born after 2011 in order to constitute such a large population of genotyped single purebred animals.

A second strategy of imputation from 8K to 60K was implemented in purebred animals by using a combined reference population. In the validation dataset, SNPs that were not present on the low-density chip were masked and subsequently imputed. However, results of imputation from 5K to 8K for both purebred and crossbred animals, and those of imputation from 8K to 60K for purebred animals could not completely describe how imputation worked from 8K to 60K for crossbred animals. Therefore, the quality of imputation from 8K to 60K for crossbred animals was validated using simulated data from the 60K SNP chip for crossbred animals. Genotypes of crossbred animals were simulated according to the genotypes of their ancestors based on frequencies of recombination according to Haldane's mapping function [38]. Additional file 1 [see Additional file 1] describes in more detail the steps used to simulate the 60K genotypes for 5639 crossbred animals. All imputations were done using the software Beagle version 3.3.2 [12].

Evaluation of imputation accuracies

Accuracies of imputation for each strategy are presented by mean correct rates and mean correlation coefficients between imputed genotypes and real genotypes. Mean correct rates were calculated per SNP (across individuals) as the proportion of correctly imputed genotypes, and then averaged over all imputed SNPs (for details, see [24]). Correlation coefficients were calculated per SNP across all imputed individuals and then averaged over SNPs, following [25].

Genomic relationships across breeds

Genomic relationships among individuals were estimated based on 8K real genotypes using VanRaden's method [26] as $\mathbf{G} = \frac{\mathbf{ZZ'}}{2\sum p(1-p)}$, where $\mathbf{Z}$ is a matrix of genotypes coded as {-1, 0, 1}, and p was set to 0.5, so that a unique reference point was chosen and results could be compared within and across breeds. Compared to pedigree-based relationships, all estimated genomic relationships will be biased upwards, but bias will be the same across breeds and subgroups of animals. The genomic relationships are thereby comparable both across and within breeds, which is the objective of our study. For each individual in the validation population, the average genomic relationship to individuals in the reference population was computed by averaging coefficients from the appropriate section of the genomic relationship matrix. Furthermore, for each crossbred individual in the validation population, the average of the top10 relationships between this individual and individuals in the reference population [27] was also computed. To visualize the distribution of relationships, density curves of genomic relationships were drawn. In addition, as suggested by [28], a principal components analysis (PCA) of the matrix of genomic relationships was conducted for a preliminary analysis of the genotypes, since PCA can help to investigate ethnic background of individuals [29].

Proportion of shared haplotypes between reference and validation populations

Following imputation by Beagle, 8K phased genotypes were available for all animals in the reference and validation populations. It was assumed that a haplotype consisted of a specific number of consecutive SNP alleles in the same phase. Lengths of haplotypes were set to 10, 20, 30, 50 and 100 SNPs. If a haplotype in the validation population could exactly match at least one haplotype at the same position in the reference population, this haplotype was considered to be shared between the reference and validation populations. The number of shared haplotypes was counted and then divided by the total number of haplotypes in the validation population, and this was referred to as the proportion of shared haplotypes (PSH). In addition, the number of unique haplotypes (NUH) in the reference populations was counted to represent the number of different patterns for a specific haplotype length across all individuals in the reference population. Values for PSH and NUH were averaged over non-overlapping windows of a specific size.

Results

Imputation strategy '5K to 8K'

Performance of purebred imputation

Figure 1 shows imputation accuracies from 5K to 8K across the 18 autosomes for the purebred Landrace and Yorkshire pigs when using a within-breed reference population. On the whole, accuracies did not vary much between chromosomes. Correct rates were larger than or equal to 0.99, except for chromosomes 3, 10, 12 and 18 for both breeds. No differences in mean correct rate were observed between the two purebreds. Correlation coefficients between imputed and true genotypes ranged from 0.90 (chromosome 10) to 0.97 (chromosome 13) for the Yorkshire breed and from 0.93 (chromosome 3) to 0.98 (chromosome 16) for the Landrace breed. Slight differences in mean correlation coefficients (0.012) were observed between the two breeds. Overall, the Landrace breed performed slightly better than the Yorkshire breed, especially in terms of the correlation coefficients. Variations of correlation coefficients were generally consistent with those of correct rates across the whole genome.

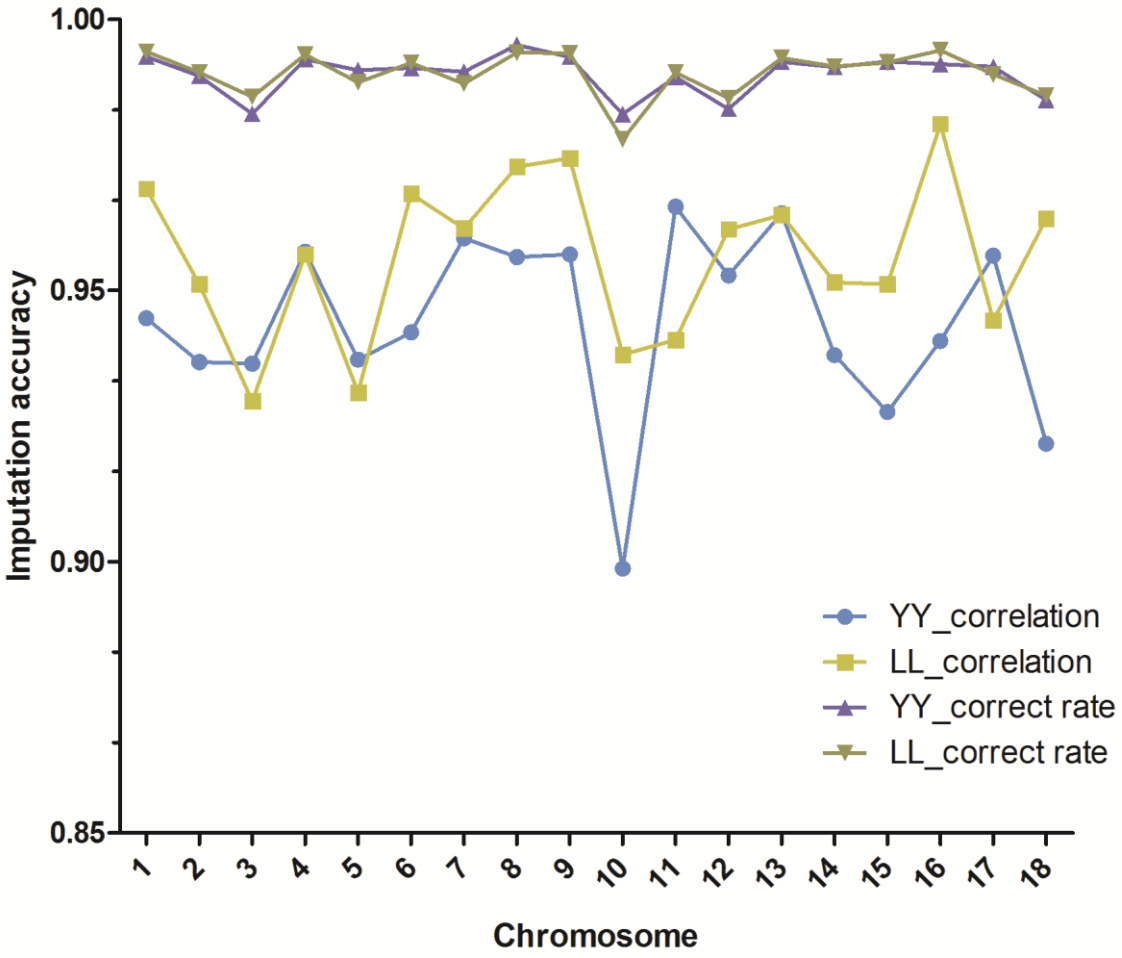

Figure 1 Variation in imputation accuracy for the scenario from 5 k to 8 k across different chromosomes using within-breed reference populations. Within-breed reference means Landrace pigs were imputed using a reference population that consisted of Landrace pigs only and Yorkshire pigs were imputed using a reference population that consisted of Yorkshire pigs only.

Comparison of imputation accuracies that were obtained in the different imputation scenarios from 5K to 8K for purebred animals is in Figure 2. Correct rates for purebred animals were identical for the within-breed and combined-breed scenarios for both breeds, but correlation coefficients increased slightly (around 0.01) in the combined-breed scenario. However, in the external-breed scenario, both correct rates and correlation coefficients decreased sharply for both breeds compared with the within-breed scenario. Landrace animals had marked lower imputation accuracies than Yorkshire animals in the external-breed scenario, whereas imputation accuracies were similar between the two breeds in the within-breed and combined-breed scenarios, both in terms of correct rates and correlation coefficients.

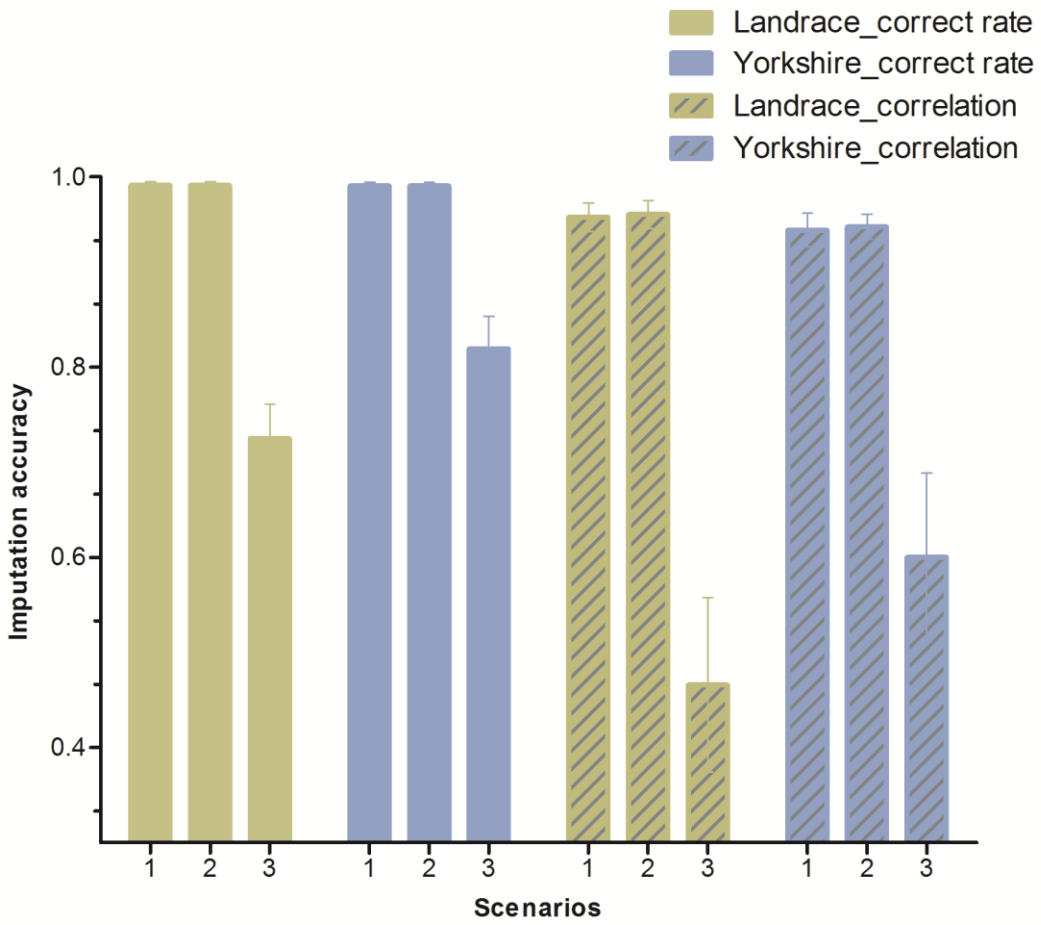

Figure 2 Comparison of imputation accuracies obtained by different imputation scenarios in Landrace and Yorkshire breeds. 1 indicates that the reference population consisted of either 4166 LL or 4263 YY, depending on the respective breed (within-breed scenario); 2 indicates that the reference population consisted of 8429 combined LL and YY (combined-breed scenario) and 3 indicates that the reference population consisted of animals that belonged to another purebred breed (external-breed scenario), which means that Landrace animals were imputed using a reference population that contained Yorkshire pigs only and Yorkshire animals were imputed using a reference population that contained Landrace pigs only. Error bars are standard deviations.

Performance of imputation for crossbred animals and comparison with that of purebred animals

Table 1 summarizes the performance of imputation from 5K to 8K for purebred and crossbred animals when the size of the reference populations was fixed to 8249. When a combined reference population was used, imputation was better for purebred animals than for crossbred animals in terms of correct rate, although the improvement was very small (around 0.006). However, in terms of correlation coefficient, imputation accuracy was slightly greater for crossbred animals than for Yorkshire pigs, but slightly lower for crossbred animals than for Landrace pigs. However, if the reference population used for imputation of crossbred animals was replaced by a pure breed population, both correct rate and correlation coefficient decreased dramatically by about 0.10 and 0.25, respectively. Imputation of crossbred animals using a reference

population that included only Yorkshire pigs resulted in a larger decline in accuracies than using a reference population that included Landrace pigs only. Table 2 presents imputation accuracies (correlation coefficients) for the subsets of crossbreds with a genotyped sire and those with a non-genotyped sire. Regardless of the reference population used, the differences were small, although the subset of crossbreds with a genotyped sire always had slightly higher accuracies than the subset of crossbreds with a non-genotyped sire.

Table 1 Accuracy of imputation from 5K to 8K for Landrace (LL), Yorkshire (YY) and crossbred animals

Imputed	Reference	Correct rate	Correlation
LL	LL+YY	0.9910	0.9606
YY	LL+YY	0.9907	0.9477
Crossbred	LL+YY	0.9849	0.9566
Crossbred	LL	0.9034	0.7595
Crossbred	YY	0.8667	0.6871

Table 2 Imputation accuracy (correlation coefficients) from 5K to 8K for crossbred animals with genotyped and non-genotyped sires

Reference	Sire non-genotyped	Sire genotyped
LL+YY	0.9529	0.9576
LL	0.7596	0.7603
YY	0.6883	0.6911

The first row indicates the components of the reference population whether it consists of a purebred Landrace (LL), Yorkshire (YY) or a combined population (LL + YY). There are 4956 crossbred animals with genotyped sires and 683 with non-genotyped sires in each subset, respectively.

Genomic relationships across breeds

The two main principal components on the matrix of genomic relationships of each individual across Landrace, Yorkshire and crossbred Landrace-Yorkshire animals are in Figure 3. The first two components explained 22.8 and 0.9% of variability across individuals, respectively. The first principal component (x-axis) separated the three populations, whereas the second component (y-axis) could not distinguish between breeds. There was hardly any connection between the two clouds of points representing the Landrace and Yorkshire breeds, whereas the cloud of points representing the crossbred Landrace-Yorkshire population was generally in between. Connections between Landrace and crossbred pigs seemed to be slightly tighter than those between Yorkshire and crossbred pigs, since there are many more points distributed in the interval between Landrace and crossbred pigs than between Yorkshire and crossbred pigs. Overall, connections between crossbred and purebred animals were not strong.

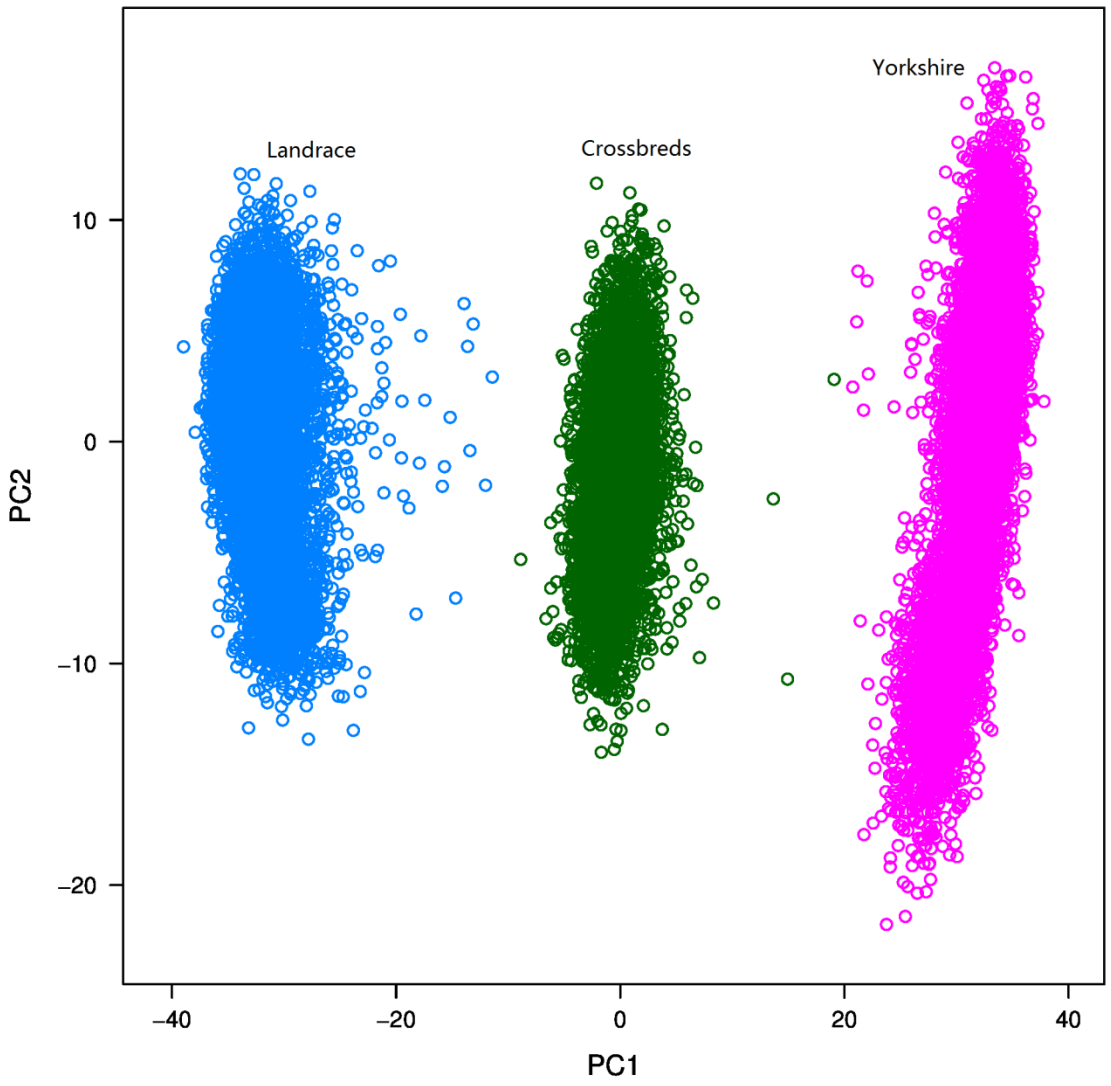

Figure 3 Principal components analysis on the matrix of genomic relationships within breeds. The first two main principal components are presented on the x-axis and y-axis, respectively. The proportions of variability across individuals explained by PC1 and PC2 were 22.92 % and 0.88 %, respectively.

Table 3 provides averaged genomic relationships between individuals in the reference and validation populations that correspond to the different imputation scenarios evaluated. The results in Table 3 show that the mean relationship within breeds was always the largest for all scenarios. When a breed was imputed using a reference population that comprised individuals of the other pure breed (external-breed scenario), the mean relationship decreased to approximately one fifth of that obtained with the within-breed scenario. When a combined reference population was implemented to impute purebred animals, logically, mean relationships were intermediate to the values found with the within-breed and external-breed scenarios. In addition, regardless of which reference population was used to impute crossbred animals, mean relationships

were similar. Distributions of genomic relationships between reference and validation populations obtained with different scenarios of imputation are represented by density curves in Figure 4. In general, for the Landrace and Yorkshire purebred pigs, the distributions of relationships were similar regardless of which reference population was used (as shown in Figures 4a, 4b and 4c). For the crossbred animals, density curves were highly consistent whether the reference population consisted of animals from one breed or from different populations (Figure 4d). The density curves of the top10 mean genomic relationships between crossbred animals and animals from the three different reference populations are in Figure 5. Landrace pigs had closer top 10 mean genomic relationships with crossbred animals than Yorkshire pigs, and by construction, animals of the combined-breed population had higher top10 mean genomic relationships with crossbred animals than either of the populations that consisted of a pure breed.

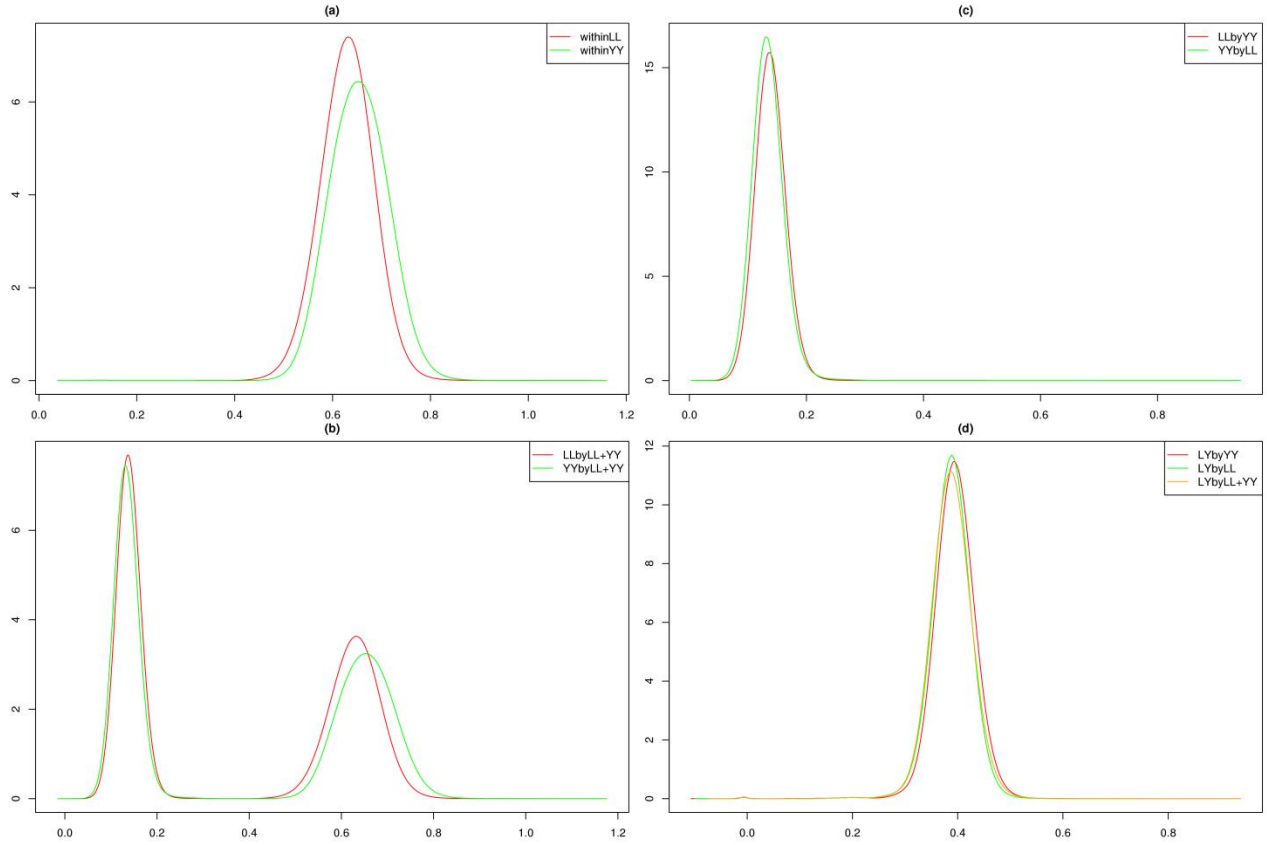

Figure 4 Density curves of genomic relationships between reference and validation populations for different imputation scenarios.(**a**) within-breed scenario for purebred Landrace and Yorkshire; (**b**) imputation of purebreds by using a combined Landrace and Yorkshire population; (**c**) the external-breed scenario for purebred Landrace (LLbyYY) and Yorkshire (YYbyLL) and (**d**) imputation of crossbreds by using either one purebred reference population (LYbyYY and LYbyLL) or a combined Landrace and Yorkshire population (LYbyLL + YY). All scenarios were under the imputation strategy of '5 K to 8 K'.

Table 3 Average genomic relationship between reference and validation populations

Validation \ Reference	LL	YY	LL + YY
LL	0.6398	0.1388	0.3874
YY	0.1343	0.6442	0.3932
Crossbred	0.3869	0.3943	0.3875

The first row indicates the components of the reference populations, whether it consists of a purebred breed Landrace (LL), Yorkshire (YY) or a combined population (LL + YY).

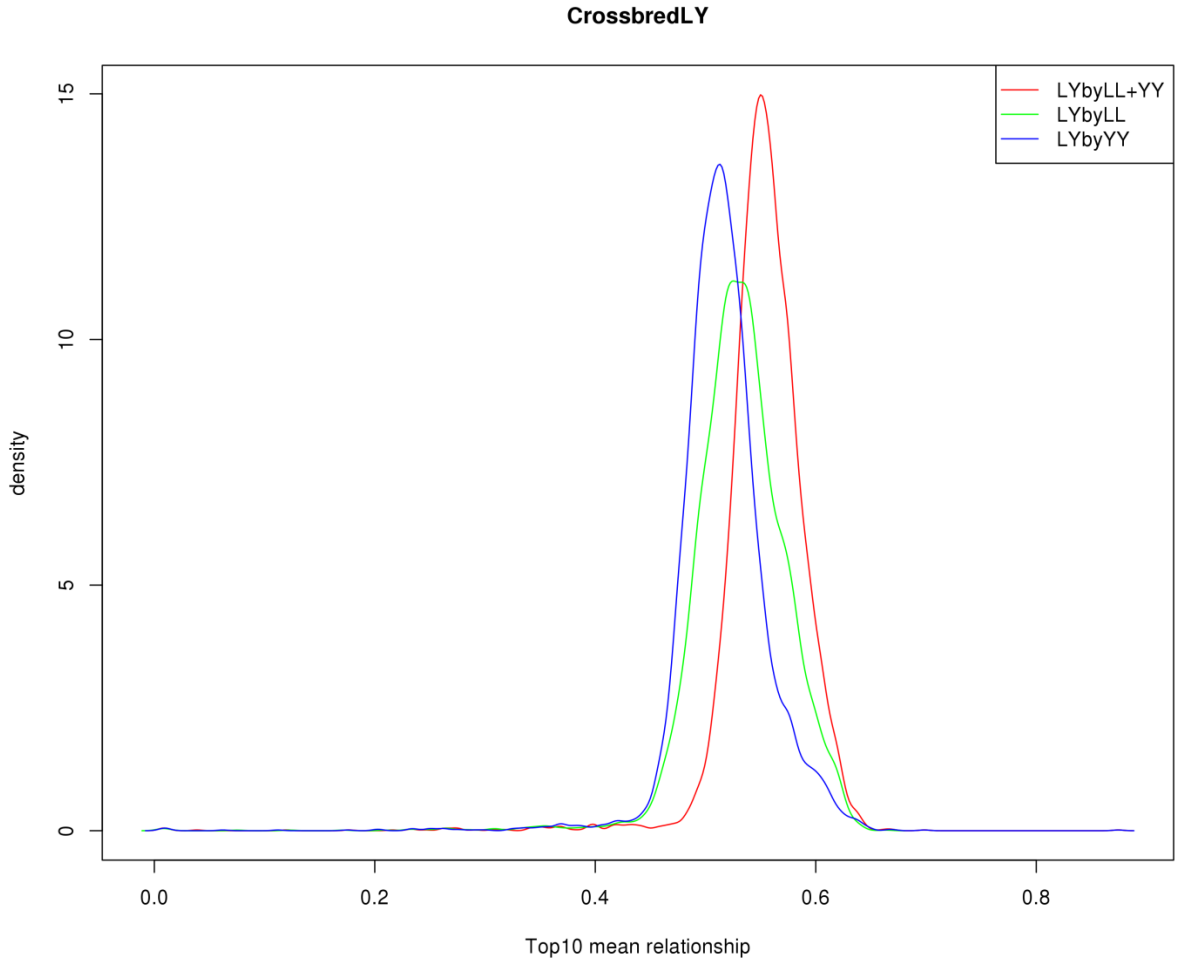

Figure 5 Density curves of the top10 mean genomic relationships between crossbred animals and three different reference populations. The reference population consisted either of a single purebred reference (LYbyLL and LYbyYY) or a combined Landrace and Yorkshire population (LYbyLL + YY).

Proportions of shared haplotypes (PSH)

Proportions of haplotypes that were shared between reference and validation populations for different imputation scenarios are in Table 4. The results show that PSH decreased as the length of haplotypes increased. For purebred animals, PSH was always very similar between Landrace and Yorkshire breeds when a within-breed or a combined population was used as reference population, regardless of the length of

the haplotypes. However, PSH decreased dramatically when the reference population consisted of only of the other breeds (external-breed). Differences in PSH existed between Landrace and Yorkshire breeds in different scenarios: for the within-breed scenario, LL had slightly higher PSH than YY when haplotypes were longer than 30 markers, but slightly lower PSH for shorter haplotypes; for the external-breed scenario, PSH was consistently lower for LL than for YY. Among the scenarios for imputation of crossbred animals, PSH was highest when a combined population was used as reference population. PSH declined when the reference population was changed from a combined population to a pure breed population. In particular, PSH was lowest when the reference population consisted of only the Yorkshire breed. The number of unique haplotypes (NUH) that existed in the reference population for different imputation scenarios is in Table 5, which shows that if only one breed was used as a reference population, Landrace animals always provided more haplotypes than Yorkshire animals. Furthermore, if the reference population consisted of a combined population, it always had a much larger NUH than if it consisted of only one breed. However, the NUH in the combined population was not equal to the sum of the NUH in each breed and was in fact smaller than this sum. In other words, some haplotypes were shared by the two breeds.

Table 4 Proportions of shared haplotypes between the reference and validation populations for different imputation scenarios

Validation	Reference	10*	20*	30*	50*	100*
LL	LL	0.9965	0.9814	0.9549	0.8838	0.6417
LL	YY	0.5043	0.1836	0.0877	0.0463	0.0141
LL	LL+YY	0.9972	0.9832	0.9556	0.8847	0.6606
YY	YY	0.9967	0.9817	0.9545	0.8825	0.6295
YY	LL	0.6806	0.3419	0.2232	0.1267	0.0364
YY	LL+YY	0.9971	0.9829	0.9589	0.8843	0.6579
Crossbred	LL	0.8579	0.6758	0.5947	0.5016	0.3280
Crossbred	YY	0.8108	0.6132	0.5125	0.4004	0.2765
Crossbred	LL+YY	0.9902	0.9606	0.9135	0.8092	0.5357

*Number of consecutive SNP alleles assumed for each haplotype. LL stands for Landrace; YY stands for Yorkshire. All the scenarios were under the imputation strategy of '5K to 8K'.

Table 5 Numbers of unique haplotypes that existed in the reference populations for different imputation scenarios

Validation	Reference	Size of reference	10*	20*	30*	50*	100*
Purebreds	LL	4166	63	223	441	956	2297
Purebreds	YY	4263	58	216	445	966	2298
Purebreds	LL+YY	8429	109	432	880	1916	4585
Crossbreds	LL	8429	79	314	669	1579	4170
Crossbreds	YY	8429	74	300	665	1571	4101
Crossbreds	LL+YY	8429	109	432	880	1916	4585

*Number of consecutive SNP alleles assumed for each haplotype. LL stands for Landrace; YY stands for Yorkshire. All the scenarios were under the imputation strategy of '5K to 8K'. Numbers in the table are averages over non-overlapping windows of a specific size.

Imputation strategy '8K to 60K'

Figure 6 shows the comparison between imputation accuracies from 8K to 60K across breeds. The 60K datasets comprised real genotypes for purebred animals and simulated genotypes for crossbred animals. According to Figure 6, in terms of correct rate, performance of imputation for crossbred animals was almost as good as that for purebred animals. Figure 6 also shows that crossbred animals performed even better than purebred animals in terms of correlation coefficients. Comparison of the results with the corresponding imputation scenarios in strategy '5K to 8K' (first three lines in Table 1) clearly indicates that both correct rates and correlation coefficients are larger for the '8K to 60K' strategy. For instance, accuracies of imputation from 8K to 60K for Landrace and Yorkshire pigs were about 0.005 and 0.015 larger than those from 5K to 8K for the correct rate and correlation coefficient, respectively. Before performing imputation from 8K to 60K in the simulated crossbred datasets, first we investigated the imputation from 5K to 8K in both the simulated and the real genotyped crossbred datasets. Results (not shown) showed that the performance of the simulated crossbred dataset was very close to that of the real crossbred dataset (0.004 greater correct rates).

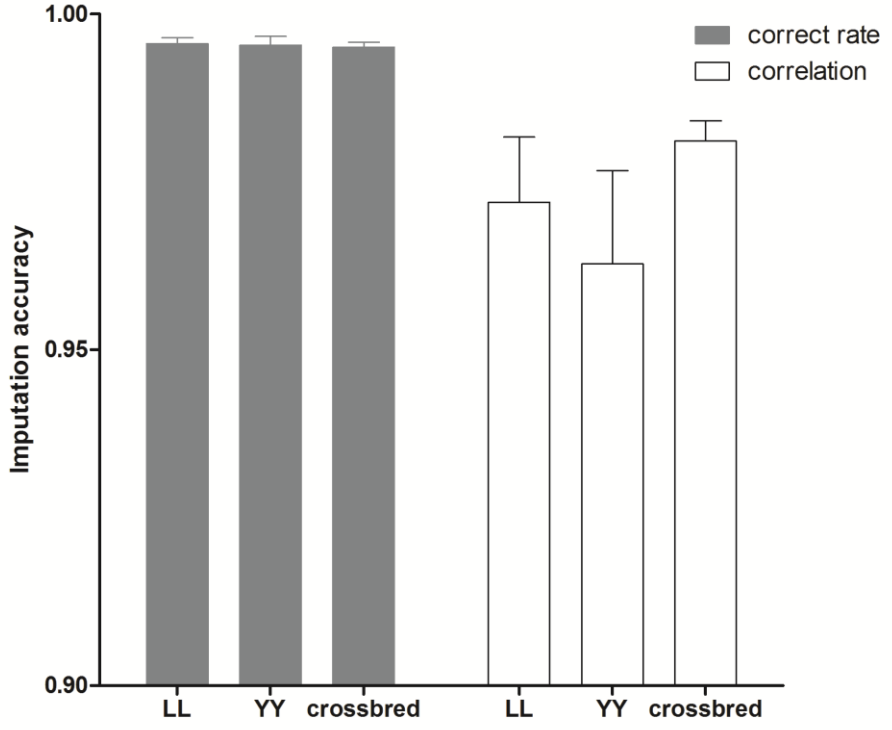

Figure 6 Comparison of imputation accuracies from 8 K to 60 K across breeds. Real genotypes were used for purebred Landrace and Yorkshire animals but simulated genotypes were used for crossbreds. Error bars are standard deviations.

Discussion

Our aim was to verify the performance of imputation in Danish purebred and crossbred pigs using different scenarios. First, we studied imputation from 5K to 8K in genotyped purebred and crossbred datasets; the performance of imputation for each autosome of the purebred animals was evaluated only in the within-breed scenario; then imputations in purebred and crossbred animals were compared in within-breed, external-breed and combined-breed scenarios. Second, imputation from 8K to 60K was evaluated using genotyped purebred and simulated crossbred data. Overall, across all imputation scenarios, correct rates and correlation coefficients were consistent with each other, i.e. higher correct rates were associated with higher correlation coefficients.

The performance of imputation for purebred animals was high and consistent across the whole genome, which indicated that the strategy performed well for all pig autosomes. Among the 18 pig autosomes, imputation was, however, slightly worse on chromosomes 3, 10, 12 and 18, which is consistent with the results of a study on the average LD on pig autosomes using a similar dataset [30]. Among the pig autosomes, autosomes 10 and 12 had a relatively low average LD, which tends to decrease the length of shared haplotypes and therefore decreases imputation accuracy, since Beagle relies crucially on local LD structure [12]. Moreover, specific SNPs on a chromosome with an extremely low minor allele frequency (MAF) reduce the average correlation coefficient for the chromosome. For instance, three SNPs on chromosome 10 had an extremely low MAF (0.000097, 0.00039 and 0.00029, respectively) in the Yorkshire dataset. Correct rates for these three SNPs were 0.994, 0.997 and 0.998, but correlations coefficients were -0.0017, 0.00045 and -0.000027, respectively. When these three SNPs were removed, the correlation coefficient for chromosome 10 increased from 0.90 to 0.93. However, in the Landrace dataset, these SNPs had a MAF of 0.497, 0.185 and 0.499, respectively, and therefore they were retained in the analysis.

Based on Figure 2, we concluded that pooling two purebred populations did not improve imputation accuracy compared to using a purebred reference population within a breed. This is in agreement with some previous studies in ruminants, which showed that combining reference populations from different breeds did not improve within-breed imputation [3,20]. A possible explanation is that haplotypes on which imputation relies are less conserved across pig breeds compared to within breeds and those that were conserved were already present in the within-breed reference population. The sharp decrease in imputation accuracies when an external breed was used as reference population also supports that haplotypes are less conserved across breeds. However, several other studies [31,32] showed that multi-breed reference populations enhance imputation accuracies compared to a single-breed reference population, but it should be noted that, in these studies, the within-breed reference population was small and imputation was done from high-density

genotyping data to sequence data, which was not the case in our study. Therefore, to impute genotypes in purebred pigs, the reference population should include at least some individuals from the breed itself or a closely related population.

Based on Table 1, imputation in crossbred animals with a reference population that combined the two purebred populations performed almost as well as imputation in purebred animals, especially in terms of correlation coefficients. One possible explanation for crossbred animals having slightly greater correlation coefficients but lower correct rates compared to purebred animals may be due to the quality control criterion used (MAF > 0.01) across both purebred populations. The distribution of MAF of the masked SNPs in the imputation strategy '5K to 8K' for Landrace (LL), Yorkshire (YY) and crossbred animals is in Figure 7. This figure shows that some SNPs had a MAF equal to 0 within a breed but not in crossbred animals. Crossbred animals tended to have higher MAF and SNPs with a very low MAF were more likely to occur for purebred animals, which decreases the correlation and increases the correct rate [6]. Imputation accuracies of crossbred animals significantly decreased when the reference population consisted of animals from only one breed. A previous study [3] suggested that imputation accuracies are expected to improve if sires and other ancestors were in the reference data, because relatives share common and longer stretches of haplotypes than distantly related animals [33]. In this study, up to 88% of the sires of crossbred animals were present in the combined purebred reference population. Haplotypes of crossbred animals can be accurately identified and imputed based on the haplotypes of their relatives. Logically, crossbred animals that were imputed using a single breed reference population had much lower imputation accuracy. One explanation is that some haplotypes of the breed that is not in the reference population are not "detected" by the imputation software which, therefore, tries to impute them based on the other breed, which has a different LD pattern. In other words, by removing one breed from the reference population, all information from one parent and its ancestors is removed. This effect is visualized in Figure 3, which shows that there were no connections between the two purebred populations for the first principal component (x-axis), and both breeds appeared to have almost equally weak connections with crossbred animals. Thus, both contributing pure breeds should be included in the reference population when imputing crossbreds to avoid inaccurately estimated haplotype blocks due to breed composition. In general, when imputing crossbred animals, it is desirable to include as many individuals of their purebred parental breeds in the reference population as possible.

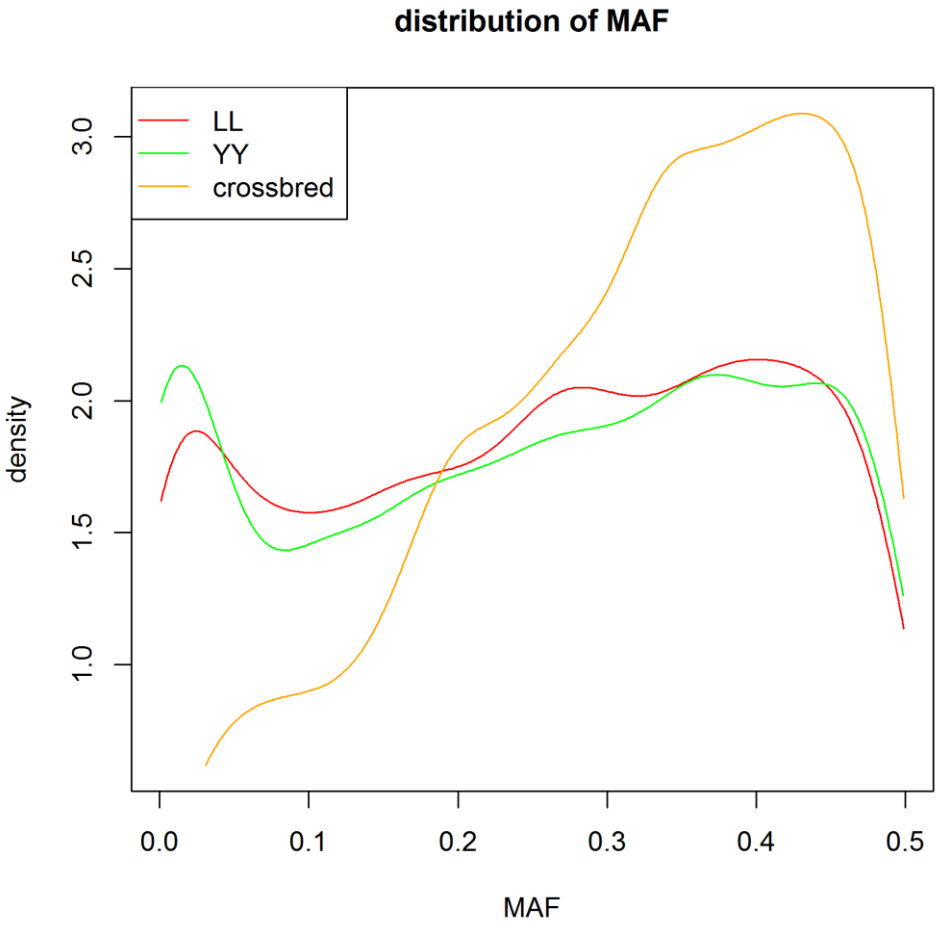

Figure 7 Density curves of minor allele frequency of the 2647 masked SNPs in the 8 K SNP chips. LL and YY represent Landrace and Yorkshire breeds, respectively. Allele frequency was calculated for each SNP across all individuals within the validation population.

Interestingly, Figure 2 and Table 1 show that Landrace pigs had higher imputation accuracies than Yorkshire pigs when a reference population that consisted of a within-breed or a combined population was used, whereas Landrace pigs performed less well than Yorkshire pigs when the reference population consisted of an external breed. Among the factors that can affect imputation accuracies and were put forward by Iwata and Jannink [14], (genomic) relationships between the validation and reference populations constitute a major factor. In this study, the two pure breeds had similar family structures, which resulted in the distribution of genomic relationships between validation and reference populations being similar for the two breeds. As shown in Figures 4a, 4b and 4c, there was no obvious difference in the density curves of relationship coefficients for Landrace and Yorkshire animals across different imputation scenarios. Thus, average genomic relationships between the validation and reference populations were similar for Landrace and Yorkshire pigs, as shown in the first two rows of Table 3. However, based on Table 3, it was not obvious that higher genomic relationships between the validation and reference populations would lead to higher

imputation accuracies, as was proposed in many other studies, such as [3,24]. Similarly, imputation accuracies for crossbred animals were also higher when imputation was done using a reference population of Landrace pigs only compared to Yorkshire pigs only, although the average genomic relationship between the crossbred validation population and the Landrace reference population was smaller than that between the crossbred validation population and the Yorkshire reference population, as shown in the last row of Table 3. All of these unexpected results indicate that the average genomic relationship is not sufficient to completely characterize the performance of imputation.

A possible explanation why imputation accuracies for crossbred animals were higher when imputation was done using a reference population of Landrace pigs only compared to Yorkshire pigs only is that close relationships play a much greater role in imputation accuracies than distant relationships [34]. According to Figure 5, the density curves of the top10 mean genomic relationships suggested that crossbreds had a closer relatedness with Landrace pigs than with Yorkshire pigs. One fact is that the number of Landrace_Yorkshire crossbreds (4432) in the crossbred dataset was much larger than the number of Yorkshire_Landrace (1207) and most of the purebred sires were genotyped and included in the reference population. This fact may lead to improved performance of imputation of crossbred animals, which is consistent with the result that subsets with genotyped sires had slightly higher imputation accuracies than subsets with non-genotyped sires (Table 2). However, a closer examination of the results in Table 2 shows that the subset of non-genotyped sires resulted in a higher accuracy when imputation used a reference population that consisted of Landrace pigs only compared to Yorkshire pigs only and that it also resulted in a higher accuracy than the subset of genotyped sires when imputation used a reference population that consisted of Yorkshire pigs only. Thus, we conclude that having a genotyped sire is not the main cause of the differences in imputation accuracies for crossbred animals when imputation used a reference population that consisted of Landrace pigs only compared to Yorkshire pigs only. Another possible interpretation of why imputation accuracies for crossbred animals were higher when imputation used a reference population that consisted of Landrace pigs only compared to Yorkshire pigs only is that the Landrace breed contains Yorkshire haplotypes. The present Danish Landrace population is based on the old Danish Landrace breed, with some known imports from other European Landrace breeds in the 1970s. It is also known that imported Yorkshire animals were crossed with the original Danish Landrace stock in the 1890s, but it was later attempted to weed out these Yorkshire crosses again [39]. Thus, it is possible that the current Danish Landrace breed contains some Yorkshire haplotypes, but not vice versa. Finally, one remarkable difference between this study and other studies is that the size of the reference populations was much larger (10 to 20 times) in our study. A large number of reference animals can provide a large number of haplotype blocks and increase the possibility that specific haplotypes in the validation population match those in the reference population. When the reference population is very large, even a small proportion of close relationships can provide many shared haplotypes between reference and validation populations and thereby improve imputation accuracies.

The proportion of shared haplotypes can explain differences in performance of imputation among scenarios across breeds. A higher PSH indicates that a larger proportion of the haplotypes in the validation population, which need to be imputed, can be matched to corresponding haplotypes in the reference population and thereby be more accurately imputed. In general, our results agree with this hypothesis, as shown in Table 4. This could be one reason why imputation of a purebred or crossbred population by using a reference population that consists of Landrace animals only, always performed better than by using a Yorkshire reference population, although all other important factors (such as relationships, LD and MAF) were very similar in the two pure breeds. The fact that LL had slightly smaller PSH than YY, when the haplotypes were short (haplotype consisted of < 30 markers), but larger PSH when the haplotypes were long, indicates different patterns of sharing: long haplotypes are from recent ancestors and short haplotypes are from old ancestors, and there were more genotyped Landrace sires than genotyped Yorkshire sires. Table 5 quantitatively shows that although the combined-breed scenario provides more diverse haplotypes in the reference population than the single-breed scenario, these non-conserved haplotypes would not contribute to improve imputation of purebred animals. Clearly, the corresponding PSH in Table 4 did not increase as the reference population was changed from a within-breed to a combined population. Likewise, the simultaneous increase in PSH and NUH illustrates quantitatively the importance of using a reference population that consists of a combined population for the imputation of crossbred animals.

The higher accuracies of imputation obtained from 8K to 60K than from 5K to 8K for purebred animals confirmed previous studies [6], which showed that increasing the number of SNPs in low-density chips can improve the performance of imputation, because with denser SNPs local LD across markers becomes stronger. Therefore, it can be inferred that the performance of imputation for crossbred animals would also be marginally improved in the 8K to 60K scenario. Accuracies of imputation from 8K to 60K for purebred animals and simulated 60K crossbreds were promising. To check that the simulation gave realistic results, the performance of imputation from 5K to 8K with a simulated crossbred dataset was compared with the performance of imputation from 5K to 8K with the real crossbred dataset (results not shown). The performance of imputation with the simulated 8K dataset was slightly better than with the real 8K dataset. The slight increase in accuracy was due to the simulation using haplotypes phased by Beagle. Thus, Beagle performed imputation based on data that had been generated under its own underlying model. Our results show that the improvement is negligible. Therefore, results from the simulated crossbred dataset can be trusted. It should be noted that there was an upper limit to the accuracy of phasing if the SNPs were sufficiently dense to be in high LD [12]. From an economic point of view, 8K markers in a low-density panel seem sufficiently dense for imputation to medium-density (60K) panels.

In pig breeding, imputation for purebred animals has also been done from very low densities (384 SNPs) to 60K densities [35-37]. Consequently, we also evaluated the imputation accuracy from very low density (425

SNPs, 1% of total SNPs retained) to 8K in a crossbred dataset with a reference population that combined animals from both pure breeds. However, the accuracies were very low, around 0.7 and 0.5 for correct rates and correlation coefficients, respectively, which seems inadequate to implement genomic evaluation for crossbred performance in pigs.

Our goal was to compare the imputation performance between purebred and crossbred animals. We used the Beagle software. Although many other software programs have been developed for imputation, their comparison was beyond the scope of our study. All the imputation scenarios were executed on a Linux server with an Intel(R) Xeon(R) E5450@3.00 GHz CPU. The system is configured to allow computation with a maximum of four cores and a total of 32 GB RAM. Running time for imputing chromosome 1 of purebred animals in the within-breed and external-breed scenarios and strategy "5K to 8K" was 4 h ± 10 min, while the running time for imputing chromosome 1 of purebred animals in the combined-breed scenarios was around 6.5 h. The running time for imputing chromosome 1 of crossbred animals was about 6.5 h ± 15 min when different reference populations were used. For strategy "8K to 60K", only the combined-breed scenario was implemented in purebred and crossbred animals and the running time for imputing chromosome 1 of crossbred animals was 67 h ± 30 min.

Conclusions

Using the software Beagle, imputation performs very well and consistently across the whole genome and, as well, in crossbreds as in purebred animals, when the reference population combines animals from both parental breeds. For purebred animals, a reference population of within-breed animals ensures a good performance of imputation, especially when the size of the reference population is large. A combined reference population does not increase imputation accuracy for purebred animals compared to a within-breed reference population. A reference population that consists of an external breed only results in very poor imputation accuracy. For crossbred animals, a highly accurate imputed 60K crossbred dataset can be achieved from 8K by using a reference population that combines both parental breeds. The best method for imputation of crossbred animals is to include all purebred parental breeds in the reference population. Relationships can account for differences in imputation accuracy, but its effect will be limited by the size of the reference population. The proportion of shared haplotypes between the reference and validation populations gives an appropriate interpretation for the performance of imputation in both purebred and crossbred pigs.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TX performed data analysis and wrote the manuscript. OFC and AL coordinated the project, conceived the study, made substantial contribution for the results interpretation and revised the manuscript. PM improved the manuscript and added valuable comments during the study. TO provided with the data and added valuable comments. All authors read and approved the manuscript.

Acknowledgements

The work was funded through the Green Development and Demonstration Programme (grant no. 34009-12-0540) by the Danish Ministry of Food, Agriculture and Fisheries, the Pig Research Centre and Aarhus University. The first author benefited from a joint grant from the European Commission and Aarhus University, within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG". AL thanks financing from INRA SelGen metaprogram projects X-Gen and SelDir. Useful comments from two anonymous reviewers are acknowledged.

References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819-29.

2. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387-406.

3. Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW, Van der Werf JH. Accuracy of genotype imputation in sheep breeds. Anim Genet. 2012;43:72-80.

4. Weigel KA, de Los Campos G, Vazquez AI, Rosa GJ, Gianola D, Van Tassell CP. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J Dairy Sci. 2010;93:5423-35.

5. Habier D, Fernando RL, Dekkers JCM. Genomic selection using low-density marker panels. Genetics. 2009;182:343-53.

6. Hickey JM, Crossa J, Babu R, de los Campos G. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 2012;52:654-63.

7. Loberg A, Dürr J. Interbull survey on the use of genomic information. Interbull Bull. 2009;39:3-14.

8. Fulton JE. Genomic selection for poultry breeding. Anim Front. 2012;2:30-6.

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

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

11. Dekkers JC. Marker-assisted selection for commercial crossbred performance. J Anim Sci. 2007;85:2104-14.

12. Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet. 2008;124:439-50.

13. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009;84:210-23.

14. Iwata H, Jannink JL. Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: accuracy evaluation in barley breeding lines. Crop Sci. 2010;50:1269-78.

15. Zhang Z, Druet T. Marker imputation with low-density marker panels in Dutch Holstein cattle. J Dairy Sci. 2010;93:5487-94.

16. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol. 2012;44:9.

17. Duarte JLG, Bates RO, Ernst CW, Raney NE, Cantet RJC, Steibel JP. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genet. 2013;14:38.

18. Badke YM, Bates RO, Ernst CW, Fix J, Steibel JP. Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3 (Bethesda). 2014;4:623-31.

19. Ventura RV, Lu D, Schenkel FS, Wang Z, Li C, Miller SP. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J Anim Sci. 2014;92:1433-44.

20. Hozé C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, et al. High-density marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol. 2013;45:33.

21. Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One. 2009;4:e6524.

22. GeneSeek Company. GGP-for Porcine LD (GeneSeek Genomic Profiler for Porcine Low Density). 2012, http://www.neogen.com/Genomics/pdf/Slicks/GGP_PorcineFlyer.pdf.

23. Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF et al. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. 2012;491:393-8.

24. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, et al. Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet. 2009;84:235-50.

25. Calus MP, Bouwman AC, Hickey JM, Veerkamp RF, Mulder HA. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: A review of livestock applications. Animal. 2014,8:1743-53.

26. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414-23.

27. Daetwyler HD, Calus MP, Pong-Wong R, de los Campos G, Hickey JM. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics. 2013;193:347-65.

28. Legarra A, Baloche G, Barillet F, Astruc JM, Soulas C, Aguerre X, et al. Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J Dairy Sci. 2014;97:3200-12.

29. McVean G. A genealogical interpretation of principal components analysis. PLoS Genet. 2009;5:e1000686.

30. Wang L, Sørensen P, Janss L, Ostersen T, Edwards D. Genome-wide and local pattern of linkage disequilibrium and persistence of phase for 3 Danish pig breeds. BMC Genet. 2013;14:115.

31. Brøndum RF, Guldbrandtsen B, Sahana G, Lund MS, Su G. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics. 2014;15:728.

32. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda). 2011;1:457-70.

33. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499-511.

34. Pszczola M, Strabel T, Mulder HA, Calus MP. Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci. 2012;95:389-400.

35. Hickey JM, Kranis A. Extending long-range phasing and haplotype library imputation methods to impute genotypes on sex chromosomes. Genet Sel Evol 2013;45:10.

36. Huang YJ, Hickey JM, Cleveland MA, Maltecca C. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet Sel Evol. 2012;44:25.

37. Wellmann R, Preuss S, Tholen E, Heinkel J, Wimmers K, Bennewitz J. Genomic selection using low density marker panels with application to a sire line in pigs. Genet Sel Evol. 2013;45:28.

38. Haldane JBS. The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet. 1919;8:299-309.

39. King JWB. Pig breeds of the world:Their distributions and adaptation. In: Maijala K, editors. Genetic resources of pig, sheep and goat. Elsevier Science Publishers; 1991. p. 52-53.

Additional file: Pedigree-based simulation

Description: The process to simulate a medium density (60 K) crossbred chip is described in which simulation uses real genotypes of purebred ancestors to simulate genotypes of their crossbred offspring.

In this additional file, we present the pedigree-based simulation process of creating a medium density (60K) crossbred chip data. Simulation used real genotypes of purebred ancestors to simulate genotypes of their crossbred offspring, as follows.

Landrace and Yorkshire can either be sires or dams of those crossbred pigs. To mimic the real data structure, the crossbred pigs were split into three main categories (see Table 6 and Table 7).

Category one was the situation when the parental genotypes of crossbreds were known, so that the crossbred progenies could be formed directly through a simulation algorithm. The simulation algorithm is an imitated meiosis process on the basis of a premise that SNPs were sorted by their physical positions on each chromosome already: according to Haldane's map function [38], the number of crossovers between two phases within one chromosome follow a Poisson distribution, with parameter equal to the genetic distance of the chromosome in Morgan; the genetic distance was assumed such that markers separated by 10 kb of DNA had an expected rate of chromosomal crossovers of 0.01 per generation; it was assumed that the places where crossovers happened follow a uniform distribution along each chromosome; a 'gamete' of one individual would then be obtained through the meiosis process; each progeny would consist of such two dropped gametes, one from each parent. One crucial premise of this algorithm is awareness of the two phases of the parents. Software Beagle was used to phase genotypes in purebreds before we implemented the simulation meiosis process.

Category two was that sires of crossbreds were genotyped, but dams of crossbreds were not. Most crossbred individuals in the real dataset belonged to this category. In order to generate the genotypes of those dams, we simulated their genotypes from one generation earlier (maternal grandparents of crossbreds). If one of the grandparents' genotypes were unknown, it would then be traced back to the previous generation (categories II-2 and II-3 in Table 6), and simulated in the same manner. This procedure would be processed until four generations back from crossbreds (category II-3 in Table 6). However, genotypes of some individuals in F_{-4} generation were still lacking (not included in reference panels). To keep the structure of dataset as analogous as a real dataset, we artificially created the genotypes of those individuals as follows: each chromosome for one individual was randomly sampled from that chromosome of another individual that was not included in its reference panel. These created individuals are shown within square brackets in Table 6 and Table 7.

Category three was a situation where neither sires nor dams of crossbreds were genotyped (Table 7). Simulation was done on those crossbreds' paternal families as well as maternal families. For all the three main categories, progenies were generated based on the original mating order that was recorded in pedigree information, if the true genotypes of those purebreds were known. When the artificially created purebred individuals were involved in simulation process, we used random mating to generate progenies. Finally, 5,639 crossbreds, which consisted of 4,432 LY and 1,207 YL, were generated by the simulation algorithm.

For example, one crossbred YL (F_0) from Category II-3 in Table 6. Its sire (F_{-1}_sire) was genotyped, but not the dam (F_{-1}_dam). One earlier generation (F_{-2}) was traced back according to the pedigree information, which means we looked at the maternal grandfather (F_{-2}_Mgf) and maternal grandmother (F_{-2}_Mgm) of the crossbred YL. Animal F_{-2}_Mgf was genotyped, but not F_{-2}_Mgm. Another earlier generation (F_{-3}) was needed to be traced back. Likewise, the great grandfather (F_{-3}_Ggf) of crossbred YL was genotyped, but not the great grandmother (F_{-3}_Ggm). Thus, one more previous generation (F_{-4}) had to be traced backward. Only the father of great grandmother (F_{-4}_Fa_ggm) was genotyped, but not the mother of great grandmother (F_{-4}_Mo_ggm). The tracing was stopped at the F_{-4} generation, thus, we had to artificially create the genotypes of F_{-4}_Mo_ggm. For the created F_{-4} female Landrace, the first chromosome would be randomly selected from another Landrace individual's chromosome one; then chromosome two was sampled from chromosome two of a third individual and so on until chromosome eighteen had been sampled. Then, based on the simulated meiosis algorithm, F_{-4} generation would produce the genotypes of F_{-3}_Ggm; F_{-3}_Ggm mated with F_{-3}_Ggf and they reproduced the F_{-2}_Mgm; alike, F_{-2}_Mgm mated with F_{-2}_Mgf and they reproduced the F_{-1}_dam; finally, F_{-1}_dam mated with F_{-1}_sire and the crossbred YL was obtained.

Table 6 Simulation process for crossbreds for which at least one parent are genotyped.
Category one represents the situation that parental genotypes for crossbreds are known; category two stands for the situation that genotypes of sires for crossbreds are known, but not the dams. Mating pattern is the way of simulating progenies, either based on mating order that recorded in pedigree information or random mating. LL is Landrace; YY is Yorkshire; LY is crossbred Landrace-Yorkshire and YL is crossbred Yorkshire-Landrace. F_0_crossbreds is the crossbred generation; F_{-1}_sires and F_{-1}_dams represent the parental generation of crossbreds; F_{-2}_Mgf and F_{-2}_Mgm are maternal grandfather and maternal grandmother for crossbreds respectively; F_{-3}_Ggf and F_{-3}_Ggm are great grandfather and great grandmother for crossbreds, representing sires and dams of the F_{-2}_Mgm respectively; F_{-4}_Fa_ggm and F_{-4}_Mo_ggm are the parents of F_{-3}_Ggm. Numbers within round brackets ( ) stand for genotypes of that number of animals are unknown. Their genotypes were simulated, basing on their corresponding parental genotypes; numbers within square brackets [ ] represent genotypes of that number of animals were made according to the way that described in the main text; numbers without any brackets mean genotypes of that number of animals are known. Short dash (-) means this information was not used.

Category	F_0_Crossbreds	F_{-1}_Sires	F_{-1}_Dams	F_{-2}_Mgf	F_{-2}_Mgm	F_{-3}_Ggf	F_{-3}_Ggm	F_{-4}_Fa_ggm	F_{-4}_Mo_ggm	Mating pattern
I	9 YL	5 YY	4 LL	-	-	-	-	-	-	in pedigree
II-1	92 LY	58 LL	(53 YY)	38 YY	36 YY	-	-	-	-	in pedigree
	286 YL	182 YY	(157 LL)	65 LL	47 YY	-	-	-	-	in pedigree
II-2	946 LY	293 LL	(486 YY)	170 YY	(216 YY)	203 YY	95 YY	-	-	in pedigree
	343 YL	160 YY	(213 LL)	117 LL	(137 LL)	146 LL	67 YY	-	-	in pedigree
II-3	2751 LY	549 LL	(1441 YY)	322 YY	(612 YY)	203 YY	(309 YY)	126 YY	187 YY + [19 YY]	at random
	529 YL	333 YY	(339 LL)	245 LL	(234 LL)	146 LL	(164 LL)	85 LL	38 LL + [83 LL]	at random

Table 7 Simulation process for crossbreds for which genotypes of parents are unknown

Category III-1 is simulation process for crossbred Landrace-Yorkshire and category III-2 is simulation process for crossbred Yorkshire-Landrace. Mating pattern is the way of simulating progenies, either based on mating order that recorded in pedigree information or random mating. LL is Landrace; YY is Yorkshire; LY is crossbred Landrace-Yorkshire and YL is crossbred Yorkshire-Landrace. F_0_crossbreds is the crossbred generation; F_{-1}_sires and F_{-1}_dams represent the parental generation of crossbreds; F_{-2}_Gf and F_{-2}_Gm are paternal or maternal grandfather and paternal or maternal grandmother for crossbreds respectively, decided by corresponding process; F_{-3}_Ggf and F_{-3}_Ggm are great grandfather and great grandmother for crossbreds, representing sires and dams of the F_{-2}_Gm respectively; F_{-4}_Fa_ggm and F_{-4}_Mo_ggm are the parents of F_{-3}_Ggm. Numbers within round brackets ( ) stand for genotypes of that number of animals are unknown. Their genotypes were simulated, basing on their corresponding parental genotypes; numbers within square brackets [ ] represent genotypes of that number of animals were made according to the way that described in the main text; numbers without any brackets mean genotypes of that number of animals are known. Short dash (-) mean the information was not used.

Category	F_0_crossbreds	F_{-1}_Sires	F_{-1}_Dams	F_{-2}_Gf	F_{-2}_Gm	F_{-3}_Ggf	F_{-3}_Ggm	F_{-4}_Fa_ggf	F_{-4}_Mo_ggf	Mating pattern
III-1	643 LY	(125 LL)	-	48 LL	(32 LL)	19 LL	22 LL	-	-	in pedigree
					(64 LL)	44 LL	(59 LL)	50 LL	27 LL + [28 LL]	at random
		-	(404 YY)	152 YY	(58 YY)	29 YY	20 YY	-	-	in pedigree
					(218 YY)	99 YY	(145 YY)	103 YY	37 YY + [68 YY]	at random
III-2	40 YL	(29 YY)	-	25 YY	(11 YY)	9 YY	10 YY	-	-	in pedigree
					(13 YY)	16 YY	(12 YY)	11 YY	6 YY + [6 YY]	at random
		-	(32 LL)	29 LL	(7 LL)	6 LL	7 LL	-	-	in pedigree
					(21 LL)	20 LL	(21 LL)	20 LL	11 LL + [10 LL]	at random

CHAPTER 3: PAPER II.

Application of single-step genomic evaluation for crossbred performance in pig

Tao Xiang[1,2*], Bjarne Nielsen[3], Guosheng Su[1], Andres Legarra[2], Ole Fredslund Christensen[1]

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark

[2]INRA, UR1388 GenPhySE, CS-52627, F-31326 Castanet-Tolosan, France

[3] SEGES, Pig Research Centre, DK-1609 Copenhagen, Denmark

[*]Corresponding author

This paper was published in *Journal of Animal Science* (2016) 94(3):936-48. DOI: 10.2527/jas2015-9930

ABSTRACT

Crossbreeding is predominant and intensively used in commercial meat production systems, especially in poultry and swine. Genomic evaluation has been successfully applied for breeding within purebreds, but also offers opportunities of selecting purebreds for crossbred performance by combining information from purebreds with information from crossbreds. However, it generally requires that all relevant animals are genotyped, which is costly and presently does not seem to be feasible in practice. Recently, a novel single-step BLUP method for genomic evaluation of both purebred and crossbred performance has been developed, which can incorporate marker genotypes into a traditional animal model. This new method has not been validated in real datasets. In this study, we applied this single-step method to analyze data for the maternal trait of total number of piglets born in Danish Landrace, Yorkshire and two-way crossbred pigs in different scenarios. The genetic correlation between purebred and crossbred performances was investigated firstly, and then the impact of (crossbred) genomic information on prediction reliability for crossbred performance was explored. The results confirm the existence of a moderate genetic correlation, and it was seen that the standard errors on the estimates were reduced when including genomic information. Models with marker information, especially crossbred genomic information, improve model-based reliabilities for crossbred performance of purebred boars, and also improve the predictive ability for crossbred animals and to some extent reduce the bias of prediction. We conclude that the new single-step BLUP method is a good tool in the genetic evaluation for crossbred performance in purebred animals.

Key words: single-step method, crossbred performance, genomic evaluation, reliability, genetic correlation, pig

INTRODUCTION

Crossbreeding is predominant and intensively used in meat production systems (Wei, 1992), especially in swine and chicken. In two-way crossbreeding schemes, selection of purebreds for their crossbred performance is the ultimate goal (Wei, 1992; Bijma and Bastiaansen, 2014). Since there exist genetic differences between breeds and genotype-by-environment interaction effects, additive genetic effects estimated based on purebred performance cannot be used to predict the crossbred performance perfectly (Lo et al., 1997). Ideally, combined purebred and crossbred information is required to implement the genetic evaluation for crossbred performance (Wei and van der Werf, 1994). However, due to the difficulty and high cost of collection of data from crossbred animals (Dekkers, 2007), it is not common to have access to crossbred data.

Genomic selection has been successfully applied in purebreds based on data from purebred animals (Loberg and Dürr, 2009; Fulton, 2012), but it also offers opportunities of selecting purebreds for crossbred performance by using combined information from purebreds and crossbreds (Ibáñez-Escriche et al., 2009; Zeng et al., 2013) or by using purebred data only (Esfandyari et al., 2015). However, it generally requires that all relevant animals are genotyped. Recently, a novel single-step BLUP method (Christensen et al., 2014) for genomic evaluation of both purebred and crossbred performance in a two-way crossbreeding system was developed, which is an extension of single-step BLUP method (Legarra et al., 2009; Christensen and Lund, 2010) from purebred performance to combined purebred and crossbred performances.

The aim of this study is to implement the new single-step BLUP method by using both purebred and crossbred data of total number of piglets born (TNB) in different scenarios, estimating the genetic correlation between purebred and crossbred performance and then explore the impact of (crossbred) genomic information on prediction reliability for crossbred performance.

MATERIALS AND METHODS

Data

For this study, all datasets were provided by Danish Pig Research Centre. Three populations were analyzed simultaneously: Danish Landrace (LL), Danish Yorkshire (YY) and two-way crossbred Danish Landrace-Yorkshire. Crossbred animals that had Landrace sire and Yorkshire dam were termed ‘Landrace_Yorkshire (LY)’, while ‘Yorkshire_Landrace (YL)’ represented crossbreds with Yorkshire sires and Landrace dams. The TNB data in this study comprised the records of the first parity in all the three populations. Totally, TNB was recorded in 293,339 LL, 180,112 YY and 10,974 crossbred animals. This dataset is termed “full population” throughout the whole paper.

Among the crossbreds, 7,407 were LY and 3,567 were YL. All of the purebred animals had first farrowing dates between 2003 and 2013, while the crossbred animals first farrowed between 2010 and 2013. The pedigree for both purebred and crossbred animals was available and all the crossbreds were traced back to their purebred ancestors until 1994 by the DMU Trace program (Madsen, 2012). Consequently, 332,929 LL, 210,554 YY and 10,974 crossbreds were in the pedigree. Among those animals, 7,723 LL and 7,785 YY were genotyped with Illumina PorcineSNP60 Genotyping BeadChip (Ramos et al., 2009). Two thirds of purebred genotyped animals were boars. For the crossbreds, 5,203 animals (4,077 LY and 1,126 YL) were genotyped with a 8.5K GGP-Porcine Low Density Illumina Bead SNP Chip (GeneSeek, 2012). SNP quality controls were applied on the same dataset in a previous study (Xiang et al., 2015), where more details can be found. Finally, 41,009 SNPs and 7,916 SNPs in autosome chromosomes were accessible in purebreds and crossbreds, respectively. Imputation was implemented in crossbreds from 7,916 SNPs to 41,009 SNPs with software Beagle (Browning, 2008), which outputs phased SNPs for both reference and imputed population, by using a joint reference panel of the two pure breeds (Xiang et al., 2015). As a result, phased 41,009 genotyped SNPs were available for the genotyped animals in both purebreds and crossbreds for the current study.

Single-step BLUP model for purebred and crossbred performances

The new single-step BLUP method of evaluating both purebred and crossbred performance was developed by Christensen et al. (2014). The model reformulates the "full" Wei and van der Werf (1994) A1 model and incorporates genomic information by using two breed-specific combined relationship matrices, which extend the marker-based relationship matrices to the non-genotyped animals.

The Wei and van der Werf model is a trivariate model,

$$\boldsymbol{y}_L = \boldsymbol{X}_L\boldsymbol{\beta}_L + \boldsymbol{Z}_L\boldsymbol{a}_L + \boldsymbol{e}_L,$$

$$\boldsymbol{y}_Y = \boldsymbol{X}_Y\boldsymbol{\beta}_Y + \boldsymbol{Z}_Y\boldsymbol{a}_Y + \boldsymbol{e}_Y,$$

$$\boldsymbol{y}_{LY} = \boldsymbol{X}_{LY}\boldsymbol{\beta}_{LY} + \boldsymbol{Z}_{LY}\boldsymbol{c}_{LY} + \boldsymbol{e}_{LY},$$

where $\boldsymbol{y}_L$, $\boldsymbol{y}_Y$ and $\boldsymbol{y}_{LY}$ contain phenotypes for purebred LL, purebred YY and F1 crossbred animals, respectively; $\boldsymbol{X}_L\boldsymbol{\beta}_L$, $\boldsymbol{X}_Y\boldsymbol{\beta}_Y$ and $\boldsymbol{X}_{LY}\boldsymbol{\beta}_{LY}$ represent fixed effects; $\boldsymbol{e}_L$, $\boldsymbol{e}_Y$ and $\boldsymbol{e}_{LY}$ were overall random residual effects, assumed to be independently normally distributed with mean 0 and variance $\mathbf{I}\sigma^2_{e_L}$, $\mathbf{I}\sigma^2_{e_Y}$ and $\mathbf{I}\sigma^2_{e_{LY}}$, respectively; $\boldsymbol{a}_L$ and $\boldsymbol{a}_Y$ contain breeding values for breed LL and breed YY for their purebred performance (mating within each own breed), $\boldsymbol{c}_{LY}$ stands for the additive genetic effects of F1 crossbred animals, and $\boldsymbol{Z}_L$, $\boldsymbol{Z}_Y$ and $\boldsymbol{Z}_{LY}$ are the respective incidence matrices. Note that the $\boldsymbol{c}_{LY}$ animal additive genetic effects are

actually formed as the sum of two additive gametic effects, one from LL and another from YY. In other words, a crossbred diploid genome decomposes into two purebred haploid genomes.

The Christensen et al. (2014) method, first, assumes that effects of markers across the different origins (Yorkshire and Landrace, in this case) are unrelated. Under this assumption, the additive effect of the genome of an F1 crossbred animal can be split into the sum of two additive gametic effects, one gamete from each breed, where the two gametic effects are uncorrelated by assumption of the model. Therefore, separate matrices of pedigree-based or genomic-based relationships can be set up within each breed, and then be combined according to purebred theory for the single-step (Legarra et al., 2009; Christensen and Lund, 2010). The analysis proceeds by estimating solutions to two different breed-specific random effects. The key to disentangle the breeds of origin for the genetic effect of the F1 individuals is the ability to construct pedigree-based partial relationship matrices (García-Cortés and Toro, 2006) or separate (by origin) genomic matrices, which in turn requires ascertainment of breed origin of the marker genotypes. More specifically, there are three steps:

Step 1). Reformulate the Wei and van der Werf model by splitting additive genetic effects for crossbred animals (LY) into breed of origin specific genetic effects, i.e, split the additive genetic value of the i-th F1 crossbred in two additive genetic values, one from each origin (LL or YY): $c_{LYi} = c^{L}_{LYi} + c^{Y}_{LYi}$. It has to be understood that neither of these is a breeding value *strictu sensu*, instead, they are additive effects in the statistical sense as "regression of value on gene dosage" as explained by Falconer et al. (1985), who clarifies the various definitions of average effect of genes in absence of random mating. Note that the new single-step model (Christensen et al. 2014) is not the animal model used by Lo et al. (1997) and Lutaaya et al. (2001). Actually, the new single-step model is a reformulation of the full model from Wei and van der Werf (1994, equation A1), whereas Lo et al. (1997) and Lutaaya et al. (2001) refer to the reduced animal model from Wei and van der Werf (1994, equation A2). In presence of pedigree information only, the full and the reduced animal model are equivalent, but in presence of crossbred genomic information this is no longer the case. In the papers of Lo et al. (1997) and Lutaaya et al. (2001), the additive genetic value of the i-th F1 crossbred is $u_{LYi} = (u^{L}_{LYp(i,L)} + \Phi_{Li}) + (u^{Y}_{LYp(i,Y)} + \Phi_{Yi})$. Here $u^{L}_{LYp(i,L)}$and $u^{Y}_{LYp(i,Y)}$ are half the additive genetic values of the purebred parents $p(i,L)$ and $p(i,Y)$, which are common to all the offspring of the same sire or dam, and Φ_{Li} and Φ_{Yi} are the respective Mendelian samplings, which are different for each offspring. In the reduced animal model, both Mendelian sampling terms are included in the residual effect of the crossbred animals, and only $u^{L}_{LYp(i,L)}$ and $u^{Y}_{LYp(i,Y)}$ are estimated. This is for two reasons: first, with pedigree information only, this term cannot be estimated; second, setting up matrices of additive relationships (and their inverse) for crossbred animals at the animal model is not straightforward (Lo et al. 1993; García-Cortés and Toro, 2006). Therefore, in the works of Lo et al. (1997) and Lutaaya et al. (2001), the additive genetic value of the i-th F1 crossbred u_{LYi} is replaced by $u^{L}_{LYp(i,L)} + u^{Y}_{LYp(i,Y)}$. With genomic relationships and in the

model of Christensen et al. (2014), these Mendelian sampling terms are embedded into a genomic relationship matrix (relationships across animals for purebreds and gametes for crossbreds) and they are no longer uncorrelated. Thus, the absorption of this term into the residual error term is not suitable. In the current study, $c_{LYi}^{L} = u_{LYp(i,L)}^{L} + \Phi_{LYi}^{L}$ and $c_{LYi}^{Y} = u_{LYp(i,Y)}^{Y} + \Phi_{LYi}^{Y}$. Additive genetic value of the i-th F1 crossbred c_{LYi} is not identical to $u_{LYp(i,L)}^{L} + u_{LYp(i,Y)}^{Y}$ in Lo et al. (1997) and Lutaaya et al. (2001). Thus, our model (which is a gametic model at the level of crossbreds) is not a single-step model equivalent of Lo et al. (1997) and Lutaaya et al. (2001), which, at the level of crossbreds, are reduced animal models.

Step 2). Construct breed-specific partial relationship matrices for each breed of origin genetic effects. Considering pedigree relationships, the variance and covariance between additive genetic purebred ($\boldsymbol{a}$) and crossbred ($\boldsymbol{c}$) effects of breed LL is described as

$$Var\begin{bmatrix}\boldsymbol{a}\\ \boldsymbol{c}\end{bmatrix} = \begin{bmatrix}\sigma_{a_L}^2 & \sigma_{a_L,c_L}\\ \sigma_{c_L,a_L} & \sigma_{c_L}^2\end{bmatrix} \otimes \mathbf{H}^{(L)}$$

This is a two-trait representation. For better understanding, the genetic effects can be split into animal effects belonging to purebred animals ($\boldsymbol{a}_L$, $\boldsymbol{c}_L$) and gametic effects belonging to crossbred animals ($\boldsymbol{a}_{LY}^{(L)}$, $\boldsymbol{c}_{LY}^{(L)}$):

$$Var\begin{bmatrix}\boldsymbol{a}_L\\ \boldsymbol{a}_{LY}^{(L)}\\ \boldsymbol{c}_L\\ \boldsymbol{c}_{LY}^{(L)}\end{bmatrix} = \begin{bmatrix}\sigma_{a_L}^2 & \sigma_{a_L,c_L}\\ \sigma_{c_L,a_L} & \sigma_{c_L}^2\end{bmatrix} \otimes \mathbf{H}^{(L)} = \begin{bmatrix}\sigma_{a_L}^2 & \sigma_{a_L,c_L}\\ \sigma_{c_L,a_L} & \sigma_{c_L}^2\end{bmatrix} \otimes \begin{bmatrix}\mathbf{H}_{\mathrm{L,L}} & \mathbf{H}_{\mathrm{L,LY}}^{(\mathrm{L})}\\ \mathbf{H}_{\mathrm{LY,L}}^{(\mathrm{L})} & \mathbf{H}_{\mathrm{LY,LY}}^{(\mathrm{L})}\end{bmatrix},$$

where matrix $\mathbf{H}^{(\mathrm{L})}$ is a matrix of partial relationships which contains four blocks, one for within purebred animals ($\mathbf{H}_{\mathrm{L,L}}$), two for purebred with crossbred animals ($\mathbf{H}_{\mathrm{L,LY}}^{(\mathrm{L})}$) and vice versa ($\mathbf{H}_{\mathrm{LY,L}}^{(\mathrm{L})}$), and one for within crossbred animals $\mathbf{H}_{\mathrm{LY,LY}}^{(\mathrm{L})}$. If there are nL pure Landrace animals and nLY crossbred animals the size of $\mathbf{H}^{(L)}$ is $(nL + nLY) \times (nL + nLY)$. The nL purebred animals have additive effects, which are breeding values, $\boldsymbol{a}_L$ (when mated within breed) and $\mathbf{c}_{\mathrm{L}}$ (when mated to the other breed). The nLY purebred gametes of crossbred animals have additive effects $\mathbf{c}_{\mathrm{LY}}^{(\mathrm{L})}$ (within the cross itself). The covariance structure includes, for ease of representation, $\boldsymbol{a}_{LY}^{(L)}$, which are effects of crossbred gametes in purebred performance; these effects are merely conceptual but they simplify the representation and computation. The covariance structure for breed YY is similar:

$$Var\begin{bmatrix}\boldsymbol{a}_Y\\ \boldsymbol{a}_{LY}^{(Y)}\\ \boldsymbol{c}_Y\\ \boldsymbol{c}_{LY}^{(Y)}\end{bmatrix} = \begin{bmatrix}\sigma_{a_Y}^2 & \sigma_{a_Y,c_Y}\\ \sigma_{c_Y,a_Y} & \sigma_{c_Y}^2\end{bmatrix} \otimes \mathbf{H}^{(Y)} = \begin{bmatrix}\sigma_{a_Y}^2 & \sigma_{a_Y,c_Y}\\ \sigma_{c_Y,a_Y} & \sigma_{c_Y}^2\end{bmatrix} \otimes \begin{bmatrix}\mathbf{H}_{\mathrm{Y,Y}} & \mathbf{H}_{\mathrm{Y,LY}}^{(\mathrm{Y})}\\ \mathbf{H}_{\mathrm{LY,Y}}^{(\mathrm{Y})} & \mathbf{H}_{\mathrm{LY,LY}}^{(\mathrm{Y})}\end{bmatrix}$$

with size of $\mathbf{H}^{(Y)}$ equal to $(nY + nLY) \times (nY + nLY)$, and both structures are assumed independent, i.e., there is no covariance between LL effects and YY effects. As in Wei and Van der Werf (1994), there are six genetic (co)variance components, three for each breed.

Matrix $\mathbf{H}^{(\mathrm{L})}$ can be constructed based on available information (pedigree, markers) as follows. The pedigree-based and marker-based breed LL partial relationship matrices are $\mathbf{A}^{(\mathrm{L})} = \begin{bmatrix} \mathbf{A}_{\mathrm{L,L}} & \mathbf{A}^{(\mathrm{L})}_{\mathrm{L,LY}} \\ \mathbf{A}^{(\mathrm{L})}_{\mathrm{LY,L}} & \mathbf{A}^{(\mathrm{L})}_{\mathrm{LY,LY}} \end{bmatrix}$ and $\mathbf{G}^{(\mathrm{L})} = \begin{bmatrix} \mathbf{G}_{\mathrm{L,L}} & \mathbf{G}^{(\mathrm{L})}_{\mathrm{L,LY}} \\ \mathbf{G}^{(\mathrm{L})}_{\mathrm{LY,L}} & \mathbf{G}^{(\mathrm{L})}_{\mathrm{LY,LY}} \end{bmatrix}$, respectively, where the partition divides purebred animals from purebred gametes in crossbred animal. Because of the split into breed-specific gametes, the pedigree-based partial relationship matrices $\mathbf{A}^{(\mathrm{L})}$ and $\mathbf{A}^{(\mathrm{Y})}$ must be computed as in García-Cortés and Toro (2006).

Construction of the breed-specific marker-based relationship matrices assumes that the breed of origin of phased alleles in crossbred animals is known. In other words, it is known which phased allele in a crossbred animal LY is from breed LL and which one is from breed YY. Then, the marker-based partial relationship matrix contains cross-products of centered genotypes:

$$\mathbf{G}_{\mathrm{L,L}} = \left(\mathbf{m}^{\mathrm{L}} - 2\mathbf{p}^{\mathrm{L}}\mathbf{1}'\right)\left(\mathbf{m}^{\mathrm{L}} - 2\mathbf{p}^{\mathrm{L}}\mathbf{1}'\right)'$$

$$\mathbf{G}^{(\mathrm{L})}_{\mathrm{L,LY}} = \left(\mathbf{m}^{\mathrm{L}} - 2\mathbf{p}^{\mathrm{L}}\mathbf{1}'\right)\left(\mathbf{q}^{\mathrm{LY}} - \mathbf{p}^{\mathrm{L}}\mathbf{1}'\right)'$$

$$\mathbf{G}^{(\mathrm{L})}_{\mathrm{LY,LY}} = \left(\mathbf{q}^{\mathrm{LY}} - \mathbf{p}^{\mathrm{L}}\mathbf{1}'\right)\left(\mathbf{q}^{\mathrm{LY}} - \mathbf{p}^{\mathrm{L}}\mathbf{1}'\right)'$$

where $\mathbf{m}^{\mathrm{L}}$ and $\mathbf{q}^{\mathrm{LY}}$ contain breed-specific allele contents of the second allele for purebred LL (coded as 0, 1, 2) and crossbred animals (coded as 0, 1), respectively; vector $\mathbf{p}^{\mathrm{L}}$ are breed LL specific allele frequencies based on marker genotypes for purebred and crossbred animals.

Later, matrix $\mathbf{G}^{(\mathrm{L})}$ is adjusted to be compatible with $\mathbf{A}^{(\mathrm{L})}$: $\mathbf{G}^{(\mathrm{L})}_{\mathrm{a}} = \mathbf{G}^{(\mathrm{L})}\beta + \mathbf{K}\alpha$, where $\mathbf{K} = \begin{bmatrix} \mathbf{J} & \mathbf{J}/2 \\ \mathbf{J}/2 & \mathbf{J}/4 \end{bmatrix}$, and $\mathbf{J}$ denotes a matrix of ones partitioned as $\mathbf{G}^{(\mathrm{L})}$. Scalars α and β are estimated through solving the two following equations:

$$\overline{\mathrm{A}}^{(L)}_{22} = \overline{\mathrm{G}}^{(L)}\beta + \overline{\mathrm{K}}\alpha,$$

$$\overline{\mathrm{dA}}^{(L)}_{22} = \overline{\mathrm{dG}}^{(L)}\beta + \overline{\mathrm{dK}}\alpha,$$

e.g., equating the averages of the full matrices and equating the averages of the diagonals of pedigree and genomic relationships for genotyped individuals (Christensen et al., 2012). Matrix $\mathbf{A}_{22}^{(L)}$ contains pedigree relationships for genotyped LL individuals. Procedure is identical for breed YY.

Step 3). Combine the pedigree-based and adjusted marker-based partial relationship matrices to a combined partial relationship matrix $\mathbf{H}^{(\mathrm{L})}$, which is similar to **H** matrix used in single-step method for purebred animals (Legarra et al., 2009; Christensen and Lund, 2010). The inverse of $\mathbf{H}^{(\mathrm{L})}$ is

$$\left(\mathbf{H}^{(\mathrm{L})}\right)^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \left(\mathbf{G}_{\omega}^{(\mathrm{L})}\right)^{-1} - \left(\mathbf{A}_{22}^{(\mathrm{L})}\right)^{-1} \end{bmatrix} + \left(\mathbf{A}^{(\mathrm{L})}\right)^{-1},$$

where [1]$\mathbf{G}_{\omega}^{(\mathrm{L})} = (1-\omega)\mathbf{G}_{\mathrm{a}}^{(\mathrm{L})} + \omega\mathbf{A}_{22}^{(\mathrm{L})}$. Parameter ω is the relative weight on the residual polygenic effect. Many other studies have investigated the weighting factors between the pedigree-based and marker-based relationship matrices (Christensen and Lund, 2010; Christensen et al., 2012; Gao et al. 2012; Su et al., 2012; Guo et al., 2015) and commonly they put forward that the weighting factors should be determined by the specific trait and the dataset analyzed. We investigated weighting factors from 0.1 to 0.5. Preliminary analysis (results not shown) for different weighing factors showed that $\omega = 0.4$ was appropriate, in terms of balance between predictive abilities and biases for crossbred animals. Procedure is identical for breed YY. The sparse inverse partial relationship matrices $\left(\mathbf{H}^{(\mathrm{L})}\right)^{-1}$and $\left(\mathbf{H}^{(\mathrm{Y})}\right)^{-1}$are used as input to solve the mixed model equations of the model.

Step 4) Thus, the complete representation of the final model for genetic evaluation is:

$$\boldsymbol{y}_L = \boldsymbol{X}_L\boldsymbol{\beta}_L + \boldsymbol{Z}_L\boldsymbol{a}_L + \boldsymbol{e}_L,$$

$$\boldsymbol{y}_Y = \boldsymbol{X}_Y\boldsymbol{\beta}_Y + \boldsymbol{Z}_Y\boldsymbol{a}_Y + \boldsymbol{e}_Y,$$

$$\boldsymbol{y}_{LY} = \boldsymbol{X}_{LY}\boldsymbol{\beta}_{LY} + \boldsymbol{c}_{LY}^{(L)} + \boldsymbol{c}_{LY}^{(Y)} + \boldsymbol{e}_{LY},$$

$$Var\begin{bmatrix} \boldsymbol{a}_L \\ \boldsymbol{a}_{LY}^{(L)} \\ \boldsymbol{c}_L \\ \boldsymbol{c}_{LY}^{(L)} \end{bmatrix} = \begin{bmatrix} \sigma_{a_L}^2 & \sigma_{a_L,c_L} \\ \sigma_{c_L,a_L} & \sigma_{c_L}^2 \end{bmatrix} \otimes \mathbf{H}^{(\mathrm{L})}$$

$$Var\begin{bmatrix} \boldsymbol{a}_Y \\ \boldsymbol{a}_{LY}^{(Y)} \\ \boldsymbol{c}_Y \\ \boldsymbol{c}_{LY}^{(Y)} \end{bmatrix} = \begin{bmatrix} \sigma_{a_Y}^2 & \sigma_{a_Y,c_Y} \\ \sigma_{c_Y,a_Y} & \sigma_{c_Y}^2 \end{bmatrix} \otimes \mathbf{H}^{(\mathrm{Y})}$$

[1] There is a typographical error in the page 939 (right column) of the online published paper.

$$Var(\boldsymbol{e}_L) = \boldsymbol{I}\sigma_L^2;\ Var(\boldsymbol{e}_Y) = \boldsymbol{I}\sigma_Y^2;\ Var(\boldsymbol{e}_{LY}) = \boldsymbol{I}\sigma_{LY}^2$$

This is a three *observed* trait model (performance in LL, YY and F1) but with two genetic effects (LL and YY), each with two *genetic* traits: purebred and crossbred performance. Estimation of genetic parameters by REML and BLUP predictions were done using the DMU software (Madsen and Jensen, 2013).

Crossbred allele tracing

Software Beagle, which was used to impute and phase genotypes in crossbred animals, does not give breed allele origins as an output. Thus, to infer the allele origins in crossbred animals, we proceeded as follows. The allele tracing was processed separately on each chromosome per individual.

Among the 5,203 genotyped crossbred animals, sires of 4,520 crossbreds were genotyped, while neither parent of the other 683 crossbreds was genotyped. When the sire was genotyped, total differences between the two sets of phased imputed alleles of a crossed animal and two sets of phased alleles of its corresponding purebred sire were compared. Comparisons between crossbred and purebred phased alleles were made on each SNP along the chromosome. For a specific comparison, if a crossbred allele was different from the corresponding purebred allele, that SNP was counted as one difference. Along the chromosome, if the sum of differences between one set of crossbred phased alleles and one set of specific purebred phased alleles was lowest among the four comparisons, then this set of specific crossbred phased alleles was considered as originating from the breed of the sire. Logically, the other set of crossbred phased alleles was assigned to the other breed.

When neither parent was genotyped, one of the two sets of phased imputed crossbred alleles was studied segment-by-segment. Each crossbred phased chromosome was split into several small segments, which consisted of 50 consecutive SNP markers. These were compared with the corresponding collection of segments from phased chromosomes of two purebred reference populations LL and YY, which were used for imputing crossbred genotypes. Each small segment in the crossbred animals should exactly match at least one segment in the reference panel since each crossbred segment was imputed by the purebred reference population. Copies of that specific segment being detected in the reference population of LL and YY were counted separately and were divided by total number of segments in the same position in the reference panel of LL and YY to get proportions of matched segment. If the proportion was higher in one breed, the crossbred segment was considered to originate from this breed. Throughout all the segments within a crossbred phased chromosome, if the vast majority of segments were considered as originating from one specific breed, then the crossbred phased chromosome was assigned to that breed. Consequently, 5,203 crossbred phased alleles were traced to either breed LL or YY.

Statistical model

For Landrace and Yorkshire, the statistical model was as follows:

$$y_{ijklmn} = \mu + hys_i + month_j + hybrid_k + b_1 \times age_{ijklmn} + b_2 \times age^2_{ijklmn} + a_m + sb_n + e_{ijklmn},$$

where the dependent variable y_{ijklmn} represented TNB in the first parity in breed LL or YY; μ was the general mean; hys_i, $month_j$ and $hybrid_k$ represented fixed effects of herd-year-season, month at farrowing and hybrid indicator of service sire (same or different breed as sow); age_{ijklmn} and age^2_{ijklmn} were covariates for the age of farrowing and its squared value, with regression coefficient b_1 and b_2, respectively; a_m was the random additive genetic effect of sow; sb_n was a random service sire effect; e_{ijklmn} was the random residual effect. Random effects were assumed to be independently normally distributed, $\boldsymbol{a}$ ~ N($\mathbf{0}$, $\mathbf{H}^{(\mathrm{L})}\sigma^2_{a_L}$) or $\boldsymbol{a}$ ~ N($\mathbf{0}$, $\mathbf{H}^{(\mathrm{Y})}\sigma^2_{a_Y}$), depending on which pure breed; $\boldsymbol{sb}$ ~ N($\mathbf{0}$, $\mathbf{I}\sigma^2_{\mathrm{sb}}$) and $\boldsymbol{e}$ ~ N($\mathbf{0}$, $\mathbf{I}\sigma^2_{\mathrm{e}}$), in which $\mathbf{H}^{(\mathrm{L})}$ and $\mathbf{H}^{(\mathrm{Y})}$ were defined previously; $\mathbf{I}$ was the identify matrix; $\sigma^2_{a_L}$ and $\sigma^2_{a_Y}$ were additive genetic variances for breed LL and YY for purebred performances, respectively; σ^2_{sb} and σ^2_{e} were variance of service boar effect and variance of residual effect.

Model for crossbred records was:

$$y_{ijlm} = \mu + hys_i + month_j + b_1 \times age_{ijlm} + b_2 \times age^2_{ijlm} + c^{(L)}_m + c^{(Y)}_m + e_{ijlm},$$

where dependent variable y_{ijlm} represented TNB in the first parity in crossbred animals; μ, hys_i, $month_j$, age_{ijlm} and e_{ijlm} represented same effects as in model for purebred records; $c^{(L)}_m$ and $c^{(Y)}_m$ were breed LL and YY origin additive genetic effect, respectively. The two additive genetic effects were assumed to be independently normally distributed, $\boldsymbol{c}^{(L)}$~ N($\mathbf{0}$, $\mathbf{H}^{(\mathrm{L})}\sigma^2_{c_L}$) and $\boldsymbol{c}^{(Y)}$~ N($\mathbf{0}$, $\mathbf{H}^{(\mathrm{Y})}\sigma^2_{c_Y}$), in which $\mathbf{H}^{(\mathrm{L})}$ and $\mathbf{H}^{(\mathrm{Y})}$ were breed LL or YY specific partial additive genetic relationships; $\sigma^2_{c_L}$ and $\sigma^2_{c_Y}$ were additive genetic variances for crossbred performances of breed LL and YY, respectively.

Scenarios

Variance components, heritabilities and genetic correlations between purebred and crossbred performances (r_{pc}) were first investigated in the full population. Heritability for purebred performance was defined as the ratio of additive genetic variances for purebred performance (σ^2_a) to phenotypic variances ($\sigma^2_p = \sigma^2_a + \sigma^2_{sb} + \sigma^2_e$), whereas heritability for crossbred animals was defined as the ratio of total additive genetic variance of crossbred performance for two breed-specific gametes ($0.5(\sigma^2_{c_L} + \sigma^2_{c_y})$) to phenotypic variances ($0.5\left(\sigma^2_{c_L} + \sigma^2_{c_y}\right) + \sigma^2_{e_{LY}}$). To explore the effect of different genotyping strategies on genetic evaluation for crossbred

performance, the breed-specific partial relationship matrices were constructed based on three different scenarios (SC) : (1) Nogen_SC, pedigree information only, which represented the traditional BLUP method; (2) Genpure_SC, pedigree information and purebred genotypes (7,723 LL and 7,785 YY), representing genotyping purebreds only; (3) Genall_SC, pedigree information, all purebred and crossbred genotypes (7,723 LL, 7,785 YY and 5,203 crossbreds). The purposes of studying Genall_SC were to check the necessity of including crossbred genomic information, which is normally not available, and to study the improvement of genomic prediction of purebred animals for crossbred performance. Information on each scenario is shown in Table 1. To make the results comparable across all studies, specific relationship matrices for breed LL and YY were calculated using allelic frequencies estimated from "old" purebred population (born before January 1, 2011), which were 2210 LL and 2161 YY, respectively. For each scenario, the variance components for purebred and crossbred performances were estimated and the genetic correlation between them was obtained.

Table 1 Scenarios for model-based reliability

Scenario	Genotypes	Phenotypes
Nogen_SC	No genotypes	Full data: 293,339 LL, 180,112 YY, and 10,974 crossbred animals
Genpure_SC	7,723 LL, 7,785 YY	
Genall_SC	7,723 LL, 7,785 YY, 5,203 crossbreds	

Secondly, model-based reliabilities of crossbred performance for purebred boars were calculated in the mentioned three different scenarios. According to pedigree, 7,407 LY and 3,567 YL were offspring of 765 LL and 465 YY sires, respectively. These sires were divided into two subgroups of genotyped and non-genotyped animals and mean model-based reliabilities were computed in each subgroup. Mean model-based reliability was calculated as (Mrode, 2005): $r^2 = \sum_{i=1}^{n}(1 - \mathrm{SEP_i^2}/\sigma_c^2)/n$, where $\mathrm{SEP_i}$ was the standard error of prediction for animal i; σ_c^2 was the variance of additive genetic effect for crossbred performance and n was the number of purebred boars that were studied. In addition, the proportion of animals that have higher model-based reliabilities in one scenario compared to another scenario in each subgroup was also investigated.

Finally, the predictive ability for crossbred animals in the validation population (4,195 crossbreds) was investigated in different scenarios. The farrowing date of January 1, 2012 was used as the cut-off date to divide recorded sows in the full population into training and validation populations. For purebred genotyped boars, only birth dates were accessible, not days of farrowing. Thus, for genotyped animals, the birth date of January 1, 2011 was instead used as the cut-off date. As a result, 240,543 LL, 139,868 YY and 6,779 crossbreds were contained in training population, with genotyped 2,210 LL, 2,161 YY and 2,357 crossbreds being included as well. The validation population for crossbred performance included 4,195 crossbreds,

among which 2,846 were genotyped. Phenotypes of crossbred animals in the validation population were corrected for fixed and random effects other than additive genetic effect ($\boldsymbol{Y}_c = \mathbf{c}_{\mathrm{m}}^{(\mathrm{L})} + \mathbf{c}_{\mathrm{m}}^{(\mathrm{Y})} + \mathbf{e}$). $\boldsymbol{Y}_c$ were obtained by using full population data, with partial relationship matrices constructed in Genall_SC .

Breed-specific partial relationship matrices were constructed based on scenarios, concerning genotypes of animals in the training population: Nogen_T is the scenario where relationship matrices **H** contained only pedigree information (i.e. $\mathbf{H}^{(\mathrm{L})} = \mathbf{A}^{(\mathrm{L})}$ and $\mathbf{H}^{(\mathrm{Y})} = \mathbf{A}^{(\mathrm{Y})}$); Genpure_T is the scenario where relationship matrices **H** contained pedigree information and purebred genotypes of 2,210 LL and 2,161 YY; Genpc_T is the scenario where relationship matrices **H** contained pedigree information and genotypes of the 2,210 LL, 2,161 YY and 2,357 crossbreds that were involved in the training dataset; Genall_T is the scenario where relationship matrices **H** comprised all information in Genpc_T, plus extra genomic information (but not the phenotypic information) of the 2,846 crossbreds in the validation population. Detailed information on each scenario is shown in Table 2. Variance components were estimated based on phenotypes from the training population in each scenario, being only slightly different from those based on phenotypes from the full population. The predictive ability of crossbreds was measured by validation correlations $cor(\hat{\boldsymbol{c}}, \boldsymbol{Y}_c)$ in each scenario, where $\hat{\boldsymbol{c}}$ were the estimated additive genetic effects for crossbreds ($\hat{\boldsymbol{c}} = \mathbf{c}_{\mathrm{m}}^{(\mathrm{L})} + \mathbf{c}_{\mathrm{m}}^{(\mathrm{Y})}$) in the validation population from different scenarios; For Genall_T, the validation population was divided into two subgroups of genotyped and non-genotyped animals and the validation correlations were made in the subgroup as well as in the whole validation population. Hotelling-Williams t-test at confidence level 5% was applied to evaluate the significance for the differences of validation correlations in each scenario. Moreover, in order to detect the possible inflation or deflation of predictions, the regression coefficients of $\boldsymbol{Y}_c$ on $\hat{\boldsymbol{c}}$ were explored to check whether they were close to one. In addition, to measure uncertainty associated with results, bootstrap sampling (Mäntysaari and Koivula, 2012; Cuyabano et al., 2015) was used in the test population to estimate means and standard errors of correlations. Results were similar to the Hotelling-Williams test above and are not shown.

Table 2 Scenarios for predictive ability

Scenario	Genotypes	Phenotypes
Nogen_T	No genotypes	Training: 240,543 LL, 139,868 YY, 6,779 crossbred animals
Genpure_T	2,210 LL, 2,161 YY	
Genpc_T	2,210 LL, 2,161 YY, 2,357 crossbreds	Validation: 52,796 LL, 40,244 YY, 4,195 crossbred animals
Genall_T	2,210 LL, 2,161 YY, 5,203 crossbreds	

To check the possible impact of different genotyping scenarios on the ranking and selection of purebred animals for their crossbred performance, Spearman's rank correlations (Spearman, 1904) between breeding

values of purebred sires (765 LL and 465 YY) for crossbred performance were calculated across different scenarios. In addition, the breeding values for crossbred performance were ranked from highest to lowest in different scenarios, and then the consistency of the purebred boars in the top 5% highest breeding values was checked across different scenarios. Furthermore, to investigate re-rankings in a situation closer to the way selection for crossbred performance could be implemented in practice for such a sow-trait, the Spearman's rank correlation and the top 5% studies were also made on the "young" sows that were included in the validation population (52,796 LL and 40,244 YY), i.e. purebred animals without own records. Among these "young" sows, 1,103 LL and 1,085 YY were genotyped. These two studies were processed on the genotyped and non-genotyped "young" sows, separately.

The new single-step BLUP method for crossbreds is complex, and therefore we tried a simpler single-trait single-step BLUP method (Legarra et al., 2009; Christensen and Lund, 2010). This method assumed that all animals belonged to a single population, using a single relationship matrix, where the compatibility adjustment of $\mathbf{G}$ to $\mathbf{A_{22}}$ was done as in Christensen et al. (2012). Predictive abilities for crossbred animals in the validation population were also measured as $cor(\hat{\boldsymbol{c}}, \boldsymbol{Y}_c)$.

RESULTS

Variance components, heritabilities and genetic correlations

Estimates of variance components and genetic correlations between purebred and crossbred performances for Landrace and Yorkshire in each scenario are shown in Table 3 together with calculated heritabilities. For each scenario, both pure breeds showed higher additive genetic variances for purebred performance (σ_a^2) than for crossbred performance (σ_c^2). Residual variances for purebred animals (σ_e^2) were larger than those for crossbred animals ($\sigma_{e_{LY}}^2$). For all scenarios, the estimated heritabilities for purebred performance (h^2) were always 0.11 and 0.09 for Landrace and Yorkshire, respectively. Heritabilities for crossbred animals (h_{LY}^2) were around 0.09 in the different scenarios. The estimated genetic correlation between purebred and crossbred ranged from 0.70 in Nogen_SC to 0.78 in Genall_SC for the Landrace breed and ranged from 0.57 in Nogen_SC to 0.68 in Genall_SC for the Yorkshire breed. Standard errors were generally large, but kept decreasing from around 0.12 (Nogen_SC) to 0.1 (Genall_SC) for both breeds. Slight differences of the estimated genetic correlation were observed between the two breeds. The Landrace breed showed slightly higher genetic correlation between purebred and crossbred performance than that for the Yorkshire breed.

Table 3 Variance components[1], heritabilities for purebred performance[2], genetic correlation between purebred and crossbred performance for Landrace and Yorkshire[3] and heritabilities for crossbred animals[4]

Scenario	Breed	σ_a^2	$\sigma_{a,c}$	σ_c^2	σ_{sb}^2	σ_e^2	r_{pc}(s. e)	h^2	$\sigma_{e_{LY}}^2$	h_{LY}^2
Nogen_SC	Landrace	1.63	0.62	0.48	0.83	12.07	0.70 (0.12)	0.11	8.36	0.08
	Yorkshire	1.23	0.61	0.92	0.73	11.47	0.57 (0.13)	0.09		
Genpure_SC	Landrace	1.65	0.78	0.68	0.88	12.16	0.73 (0.11)	0.11	8.40	0.09
	Yorkshire	1.21	0.64	0.96	0.72	11.67	0.59 (0.12)	0.09		
Genall_SC	Landrace	1.65	0.89	0.79	0.88	12.16	0.79 (0.09)	0.11	8.33	0.10
	Yorkshire	1.23	0.75	0.99	0.72	11.67	0.68 (0.10)	0.09		

[1] σ_a^2=additive genetic variance for purebred performance; $\sigma_{a,c}$=genetic covariance between purebred and crossbred performance; σ_c^2=additive genetic variance for crossbred performance; σ_{sb}^2=variance of service-boar effect; σ_e^2=residual variance for purebred performance; $\sigma_{e_{LY}}^2$=residual variance for crossbred animals.

[2] h^2=heritability for purebred performance ($= \sigma_a^2/(\sigma_a^2 + \sigma_{sb}^2 + \sigma_e^2)$).

[3] r_{pc}=genetic correlation between purebred and crossbred performance.

[4] h_{LY}^2=heritability for crossbred animals ($= 0.5(\sigma_{c_L}^2 + \sigma_{c_y}^2)/(0.5(\sigma_{c_L}^2 + \sigma_{c_y}^2) + \sigma_{e_{LY}}^2)$).

Model-based reliability

Table 4 compares the mean model-based reliabilities for purebred sires for their crossbred performance in different scenarios across all boars and for genotyped and non-genotyped subgroups. The genotyped subgroup always had higher model-based reliabilities than the non-genotyped group, and for the group of all boars, model-based reliabilities were in-between those of the subgroups of genotyped and non-genotyped animals in each scenario. Model-based reliabilities increased from about 0.28 to 0.39 for the Landrace breed and from about 0.22 to 0.37 for the Yorkshire breed from Nogen_SC to Genall_SC. From Nogen_SC to Genall_SC, model-based reliabilities kept increasing in all the three groups. Overall, methods with marker information (Genpure_SC and Genall_SC) presented higher model-based reliabilities than the pedigree-based scenario (Nogen_SC). In addition, proportions of purebred boars that have larger model-based reliabilities between pairwise scenarios were also studied. Result shows that 100% of LL and YY boars had larger model-based reliabilities in the Genall_SC compared to the Nogen_SC and Genpure_SC (results not shown). Concerning the single-trait, single-step BLUP model, model-based reliabilities for purebred LL and YY in Genall_SC were 0.70 ± 0.12 and 0.69 ± 0.12, respectively. Although these values are much higher than results shown in Table 4, they cannot be compared directly, because they represent the reliability of animals drawn from a breed that would be a mixture of YY and LL, which is not the case. In fact this single trait model has lower predictive abilities than Christensen's model as will be shown next.

Table 4 Mean model-based reliabilities of purebred boars for their crossbred performance

	All[1]			Genotyped[2]			Non-genotyped[3]		
	Nogen_SC	Genpure_SC	Genall_SC	Nogen_SC	Genpure_SC	Genall_SC	Nogen_SC	Genpure_SC	Genall_SC
LL	0.303	0.332	0.385	0.307	0.341	0.391	0.280	0.279	0.346
YY	0.262	0.284	0.365	0.264	0.288	0.369	0.218	0.223	0.301

[1]All = all the sires of crossbred animals, consisting of 765 Landrace and 465 Yorkshire.
[2]Genotyped = genotyped sires of crossbred animals, consisting of 656 Landrace and 443 Yorkshire.
[3]Non-genotyped = Non-genotyped sires of crossbred animals, consisting of 109 Landrace and 22 Yorkshire.

Predictive abilities

Predictive abilities for crossbred pigs in the validation group for different scenarios are shown in Table 5. The Pearson correlation between the corrected phenotypes and the estimated breeding values ($cor(\hat{\boldsymbol{c}}, \boldsymbol{Y}_c)$) range from 0.084 in Nogen_T to 0.120 in Genall_T, as shown in the second row of Table 5. No statistically significant differences between Genpure_T and Nogen_T were found, but Genpc_T and Genall_T were statistically significantly more accurate than those two scenarios. For the Genall_T, the subgroup of 2,846 genotyped crossbred pigs reveals larger correlation coefficients than that in the subgroup of non-genotyped pigs. Furthermore, the subgroup of non-genotyped pigs in Genall_T shows larger correlation coefficients than those in other scenarios.

Regression coefficients of corrected phenotypes on the estimated breeding values are shown in Table 5. In general, regression coefficients were a little bit larger than one for all the scenarios. Regression coefficients for scenarios with marker information (Genpure_T, Genpc_T and Genall_T) were closer to one than that for pedigree based scenario (Nogen_T). Among scenarios with marker information, in terms of unbiasedness, there was no clear trend showing which scenario performed better, but none was clearly biased. For the Genall_T, the subgroup of genotyped animals had less bias than the subgroup of non-genotyped animals.

Table 5 Predictive abilities for crossbred animals in the validation population in different scenarios

	Nogen_T	Genpure_T	Genpc_T	Genall_T		
				All	genotyped	non-genotyped
$cor(\hat{\boldsymbol{c}}, \boldsymbol{Y}_c)$[1]	0.084[a]	0.088[a]	0.097[b]	0.120[c]	0.126	0.106
Regression coefficients[2]	1.179	1.049	1.081	1.067	1.048	1.105
Single-trait $cor(\hat{\boldsymbol{c}}, \boldsymbol{Y}_c)$[3]	0.079	0.084	0.088	0.106	0.109	0.103
Single-trait regression coefficients[4]	0.588	0.644	0.644	0.698	0.875	0.647

[1] $cor(\hat{\boldsymbol{c}}, \boldsymbol{Y}_c)$ is correlation coefficients between corrected phenotypes and estimated breeding values; different superscripts of small letters among scenarios indicate significant differences ($p<0.05$) by Hotelling-Williams t-test

[2]Regression coefficients of corrected phenotypes on estimated breeding values

[3]Single-trait $cor(\hat{\boldsymbol{c}}, \boldsymbol{Y}_c)$ is correlation coefficients between corrected phenotypes and estimated breeding values based on the single trait single-step BLUP method

[4]Single-trait regression coefficients is regression coefficients of corrected phenotypes on estimated breeding values based on the single trait single-step BLUP method

Single-trait single-step BLUP predictive abilities

Predictive abilities by a single-trait single-step BLUP method for crossbred animals in the validation population were shown in last two rows in Table 5. They increase from 0.079 in Nogen_T to 0.106 in Genall_T. It can also be seen that the predictive abilities calculated based on the single-trait model show similar trends as those calculated from the three-trait model, but are smaller than in each corresponding scenario. Regression coefficients increase slightly from 0.59 in Nogen_T to 0.70 in Genall_T, but are further from 1 when compared with regression coefficients calculated based on the three-trait model. For Genall_T, the genotyped subgroup also had higher predictive abilities than that in non-genotyped subgroup.

Re-ranking of purebred animals across scenarios

The Spearman's rank correlations between estimated crossbred breeding values of purebred boars (765 LL and 465 YY) in pairwise scenarios were shown in Table 6. For both breeds, it can be seen that the pairwise correlations are always smaller than 1. In terms of the "top 5%" study, from 60% to 82% of purebred boars (either LL or YY) were shared from one scenario to another in the top 5% highest breeding values (Results not shown). Similar results were observed in "young" purebred sows (Results not shown).

Table 6 Spearman's rank correlations between crossbred breeding values for 765 Landrace boars (above the diagonal) and 465 Yorkshire boars (below the diagonal) of crossbred animals in pairwise scenarios

	Nogen_SC	Genpure_SC	Genall_SC
Nogen_SC	1.00	0.92	0.90
Genpure_SC	0.93	1.00	0.98
Genall_SC	0.87	0.95	1.00

DISCUSSION

This study implemented the single-step BLUP method of Christensen et al. (2014) by using both purebred and crossbred data from Danish Landrace and Yorkshire in several scenarios with regard to different amounts of genomic information. Results indicated that the model was applicable. The genetic correlation between purebred and crossbred performance for TNB was successfully estimated. Methods with marker information were powerful for genetic evaluation for crossbred performance with regard to the predictive ability and unbiasedness. In addition, this study demonstrated that, in order to implement genetic evaluation for crossbred performance, crossbred genomic information is useful in addition to purebred genotypes.

In the model, a key assumption was that breed origins of phased marker genotypes for crossbred animals were known. In this study, crossbred 60K genotypes were imputed from 8K crossbred panel. Although Xiang et al (2015) concluded that the imputation accuracies would be larger than 99% in terms of allele correct rates and 95% in terms of correlation coefficients between imputed genotypes and true genotypes, the uncertainty of crossbred genotypes cannot be totally eliminated. The algorithm of tracing alleles in the current study were considered as working efficiently, since the differences between two purebred reference panels were considerably large in several sampled chromosomes. However, errors of tracing alleles still probably appeared if the similarity of two phased crossbred segments were high. All in all, a hidden risk of using incorrect alleles may still exist when building the breed-specific partial relationship matrix. This needs further research.

The additive genetic variances for purebred performance (σ_a^2) were larger than those for crossbred performance (σ_c^2) implying that the phenotypes of purebred animals could be more diverse than for the crossbred animals, which was in line with the phenotypic variances for purebred animals (15.12 and 14.14 for LL and YY, respectively) were larger than those for crossbred animals (9.49). The heritabilities for crossbred animals (h_{LY}^2) were not dramatically different from heritabilities for purebred performance (h^2), which was opposite to results in Wei and van der Werf (1995), and is due to the fact that in the current study, variances of environmental effects for crossbreds ($\sigma_{e_{LY}}^2$) were only two-thirds of those for purebreds (σ_e^2), which could be a consequence of heterosis and phenotypic plasticity (better fitness) to the multiple herds for crossbreds than for purebreds (Misztal and Løvendahl, 2012) or alternatively be due to fact that only three different herds were used for crossbreds. Crossbreeding capitalizes on heterosis effects and complementarity between breeds and results in an increased performance of crossbreds compared to purebreds (Dekkers, 2007).

When selection is based on purebred performance, the genetic correlation between purebred and crossbred performance (r_{pc}) is a key genetic parameter in crossbreeding schemes (Bell, 1982; Bijma and Bastiaansen,

2014). The genetic correlations between purebred and crossbred performance for TNB were around 0.75 and 0.63 for Landrace and Yorkshire, which confirmed the existence of a moderate correlation. The r_{pc} is smaller than one, which is due to different environments for purebreds and crossbreds (Lutaaya et al. 2001) and the presence of dominant gene action combined with different allele frequencies in the two breeds (Lo et al., 1997; Christensen et al. 2014). This result was in line with Wong et al. (1971), which reported that the r_{pc} for litter size was 0.74. However, Wei et al. (1992) reviewed some other studies that reported low or even negative genetic correlations between purebred and crossbred performance for litter size. A change of r_{pc} over time was reported to be caused by long-term purebred selection (Pirchner and VonKrosigk, 1973) and thus, it needs to be estimated regularly. The standard errors on the estimated genetic correlations were generally large in the current study, which implies that the sample size was not large enough, especially for crossbreds. Taking the standard errors into account, the estimated correlations in different scenarios were not very different. Nevertheless, the slight decrease of standard errors with an increased amount of genomic information indicated that genotypes, especially crossbred genotypes, would reduce the uncertainty of r_{pc}. The decreasing standard errors demonstrated the better performance of the new single-step model incorporating crossbred marker information compared to the pedigree-based selection of purebred animals for crossbred performance. Bijma and Bastiaansen (2014) showed that when using pedigree relationships, the standard error of r_{pc} was determined by number of sire families and reliabilities of EBVs and suggested that the standard error should not exceed 0.05. In the current study, the TNB was a low heritable trait (around 0.1) and only 1,018 sires of the 1,230 sires of 10,974 crossbred animals were genotyped, which was also low. Thus, large standard errors were expected. Results in the current study showed that r_{pc} for Landrace was slightly larger than that for Yorkshire, although standard errors were large. Genetic correlations (r_{pc}) also consistently increase with number of genotypes used. One possible explanation could be that there are still some discordances between the definition of base populations in genomic and pedigree relationships. Concerning heritabilities for purebred performance, our estimates confirmed the results of Guo et al. (2015), which estimated heritability 0.11 and 0.09 for TNB in Landrace and Yorkshire, respectively.

The model-based reliabilities for purebred boars for their crossbred performance were generally low in the current study. The magnitude of these reliabilities is a direct function of the prediction error variances, which in this case are mostly determined by the numbers of offspring per boar (Dufrasne et al., 2011). In the current study, the numbers of crossbred offspring for each boar ranged from 1 to 11, with average 5, which were low and led to high uncertainty of prediction. According to Table 4, model-based reliabilities tended to increase as the amount of genomic information increased for both two breeds. The scenarios with marker information presented larger model-based reliabilities than the pedigree-based scenario, which may be due to the additional marker information. Reliabilities for the subgroup of genotyped animals were larger than that for the subgroup of non-genotyped animals in each scenario, but non-genotyped animals also benefitted from genomic information of genotyped animals, as the reliabilities for non-genotyped subgroup kept increasing

from Nogen_SC to Genall_SC. These results are in line with Lourenco et al. (2015). Reliabilities for non-genotyped animals in Genall_SC were even larger than those in Nogen_SC and Genpure_SC for genotyped animals, implying the benefit of genotyping crossbred animals. In addition, 100% of purebred boars had larger model-based reliabilities in the Genall_SC than that in the other two scenarios which also evidenced that the model incorporating crossbred marker information was useful for genetic evaluation for crossbred performance in purebred boars. We concluded that crossbred genomic information plays a role in improving reliabilities for crossbred performance in purebred boars. Nevertheless, it has been reported that the model-based reliabilities overestimated the true reliabilities (VanRaden et al., 2009), because the markers may overfit the dataset (Su et al., 2012). Thus, further investigation on true reliabilities is needed, potentially by a simulation study.

Correlation coefficients between corrected phenotypes and estimated breeding values for TNB in crossbred animals were lower than results for daily gain and feed conversion ratio in Christensen et al. (2012). This may be related to the fact that the heritability was higher for the traits of daily gain and feed conversion ratio than for the TNB in current study. Moreover, the additive genetic effects for crossbred animals required estimating two breed of origin genetic effects $c^{(L)}$ and $c^{(Y)}$, which may led to more uncertainty for crossbred animals than studies for purebred animals in Christensen et al. (2012). The $cor(\hat{\boldsymbol{c}}, \boldsymbol{Y}_c)$ in different scenarios confirmed that the methods with marker information would enhance the predictive ability. The crossbred genomic information was useful to improve the prediction, since scenarios with only purebred genotypes did not show significant improvement compared with the pedigree-based scenario, but improved significantly when crossbred genomic information was also involved. Results showed that genotyped animals had larger $cor(\hat{c}, Y_c)$ than non-genotyped animals, which was opposite to studies by Guo et al. (2015). This could be because in current study, the validation group consisted of crossbred animals among which the genotyped subset was a random sample, without biases for prediction (Su et al., 2012), whereas in Guo et al. (2015) the validation group consisted of purebred animals among which the genotyped subset was a preselected group. Preselection reduces accuracies of estimated breeding values (Bijma, 2012; Lourenco et al., 2015). The non-genotyped subgroup of crossbred animals in Genall_T had larger accuracies than those in other scenarios, indicating that non-genotyped validated animals benefited from crossbred genotyped animals in the validation population. Thus, we suggest to genotype crossbred animals as well as purebred animals when implementing genomic selection for crossbred animals.

Regression coefficients of corrected phenotypes on EBVs did not show a clear preference for a specific scenario, but coefficients in all scenarios with marker information were closer to one than in the pedigree-based scenario. All the regression coefficients were larger than 1, suggesting the underestimation (deflation) of variation of the estimated genomic breeding values (Gao et al., 2012).

Both the values of Spearman's rank correlations lower than one and the "top 5%" study indicated that rankings of purebred animals' breeding values for crossbred performance were not consistent across different scenarios. The selected purebred candidates for crossbred performance will be different with the availability of (crossbred) genomic information.

In terms of the predictive abilities and bias, the single-trait model was less robust than the three-trait model, although easier to implement. With crossbred genomic information, the three-trait model showed up to 13% higher predictive abilities than the single-trait model, which seems an interesting gain for this low heritable trait.

CONCLUSION

The new single-step model works well for genetic evaluation for crossbred performance in pigs. A moderate, positive genetic correlation between purebred and crossbred performance (r_{pc} ranged from 0.57 to 0.78) for TNB in purebred Landrace and Yorkshire is confirmed. Crossbred genomic information reduces the standard error on the estimate of this genetic correlation. Models with marker information, especially crossbred genomic information, improve model-based reliabilities for crossbred performance of purebred boars, and also improve the predictive ability for validated crossbred animals and somehow reduce the bias of prediction. The single-step model that considered the three populations as a single one resulted in lower predictive abilities. The model is a good tool in the genetic evaluation for crossbred performance in purebred animals.

LITERATURE CITED

Bell, A. E. 1982. Selection for heterosis - results with laboratory and domestic animals. In: 2nd World Congress on Genetics applied to Livestock Production, Madrid, Spain, October 4-8, 1982. 6: 206-227.

Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. J Anim Breed Genet. 129: 345-358.

Bijma, P., and J. W. Bastiaansen. 2014. Standard error of the genetic correlation: how much data do we need to estimate a purebred-crossbred genetic correlation? Genet Sel Evol. 46: 79.

Browning, S. R. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet. 124: 439-450.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 42: 2.

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

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

Cuyabano, B. C., G. Su, G. J. Rosa, M. S. Lund, and D. Gianola. 2015. Bootstrap study of genome-enabled prediction reliabilities using haplotype blocks across Nordic Red cattle breeds. J. Dairy Sci. 98: 7351-7363.

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

Dufrasne, M., M. Rustin, V. Jaspart, J. Wavreile, and N. Gengler. 2011. Using test station and on-farm data for the genetic evaluation of Pietrain boars used on Landrace sows for growth performance. J. Anim. Sci. 89: 3872-3880.

Esfandyari, H., A. Sorensen, and P. Bijma. 2015. Maximizing crossbred performance through purebred genomic selection. Genet Sel Evol. 47: 16.

Falconer, D. S. 1985. A note on Fisher's 'average effect' and 'average excess'. Genet Res. 46: 337-347.

Fulton, J. 2012. Genomic selection for poultry breeding. Animal Frontiers. 2: 30-36.

Gao, H., O. F. Christensen, P. Madsen, U. S. Nielsen, Y.Zhang, M. S. Lund, and G. Su. 2012. Comparison on genomic predictions using three GBLUP methods and two single-step blending methods in the Nordic Holstein population. Genet Sel Evol. 44: 8.

García-Cortés, L. A., and M. Á. Toro. 2006. Multibreed analysis by splitting the breeding values. Genet Sel Evol. 38: 601-615.

GeneSeek. 2012. GGP-for Porcine LD (GeneSeek Genomic Profiler for Porcine Low Density). http://www.neogen.com/Genomics/pdf/Slicks/GGP_PorcineFlyer.pdf.

Guo, X., O. F. Christensen, T. Ostersen, Y. Wang, M. S. Lund, and G. Su. 2015. Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. J. Anim Sci. 93: 530-512.

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

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

Lo, L. L., R. L. Fernando, and M. Grossman. 1993. Covariance between relatives in multibreed populations - additive-model. Theor Appl Genet. 87: 423-430.

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

Loberg, A., and J. W. Dürr. 2009. Interbull survey on the use of genomic information. Interbull Bulletin. 39: 3-14.

Lourenco, D. A. L., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. Genet Sel Evol. 47: 56.

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

Madsen, P. 2012. DMU Trace, A program to trace the pedigree for a subset of animals from a large pedigree file. Version 2., Center for Quantitative Genetics and Genomics. Dept. of Molecular Biology and Genetics. Aarhus University, Denmark.

Madsen, P., and J. Jensen. 2013. A user's guide to DMU. Version 6, release 5.2. Center for Quantitative Genetics and Genomics. Dept. of Molecular Biology and Genetics. Aarhus University, Tjele, Denmark.

Mäntysaari, E. A., and M. Koivula. 2012. GEBV Validation Test Revisited. Interbull Bulletin. 45: 11-16.

Misztal, I., and P. Løvendahl. 2012. Environmental Physiology of Livestock. John Wiley & Sons, West Sussex, UK.

Mrode, R. A. 2005. Linear Models for the prediction of animal breeding values. 2nd ed. CAB Int. Publ., Midlothian, UK.

Pirchner, F., and C. VonKrosigk. 1973. Genetic parameters of cross-and purebred poultry. Brit Poultry Sci. 14: 193-202.

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

Spearman, C. 1904. The proof and measurement of association between two things. Amer. J. Psychol. 15: 72-101.

Su, G., P. Madsen, U. S. Nielsen, E. A. Mäntysaari, G. P. Aamand, O. F. Christensen, and M. S. Lund. 2012. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. J Dairy Sci. 95: 909-917.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16-24.

Wei, M. 1992. Combined crossbred and purebred selection in animal breeding. PhD Thesis, Wageningen University and Research Centre, Wageningen, the Netherlands.

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

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

Wong, W. C., W. J. Boylan, and W. E. Rempel. 1971. Purebred versus crossbred performance as a basis of selection in swine. J. Anim Sci. 32: 605-610.

Xiang, T., P. Ma, T. Ostersen, A. Legarra, and O. F. Christensen. 2015. Imputation of genotypes in Danish purebred and two-way crossbred pigs using low-density panels. Genet Sel Evol. 47: 54.

Zeng, J., A. Toosi, R. Fernando, J. Dekkers, and D. Garrick. 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol. 45: 11.

CHAPTER 4: PAPER III.

Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders

Tao Xiang[1,2*], Ole Fredslund Christensen[1], Andres Legarra[2]

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark

[2]INRA, UR1388 GenPhySE, CS-52627, F-31326 Castanet-Tolosan, France

[*]Corresponding author

This paper was published in *Journal of Animal Science* (2017) 95(4):1472-80. DOI: 10.2527/jas 2016.1155

ABSTRACT

A single-step genomic BLUP method (ssGBLUP) has been successfully developed and applied for purebred and crossbred performance in pigs. However, it requires phasing the genotypes and inferring breed origin of alleles in crossbred animals, which is somehow inconvenient. Recently, a new concept of metafounders that considers relationship within and across base populations was developed. With this concept of metafounders, regular methods to build and invert the pedigree relationships matrix can be used with only minor modifications and moreover, genomic relationships and pedigree-based relationships are automatically compatible in the ssGBLUP. In this study, data of total number of piglets born in Danish Landrace, Yorkshire, and two-way crossbred pigs and models for purebred and crossbred performance was revisited by use of ssGBLUP with two metafounders. Genetic variances and genetic correlations between purebred and crossbred performances were first re-estimated. Then, model-based reliabilities of purebred boars for their crossbred performance and predictive abilities for crossbred animals were compared in different scenarios. Results in this study were compared to those in a previous study with identical data but with models that required known breed origin of crossbred genotypes. Results show that relationships for base individuals within Landrace and within Yorkshire are similar, and that the ancestor populations for Landrace and Yorkshire are related. In terms of model-based reliabilities and predictive abilities, ssGBLUP with metafounders performs at least as well as the single-step method requiring phasing, at a lower complexity.

Key words: crossbred performance, genomic evaluation, metafounders, pig, single-step method.

INTRODUCTION

Single-step genomic BLUP (ssGBLUP) (Legarra et al., 2009; Christensen and Lund, 2010) has been successfully used in genomic evaluation to handle the situation where only a fraction of animals are genotyped. Christensen (2012) summarized two issues with ssGBLUP. First, in theory allelic frequencies in the base population of the pedigree should be used in the genomic relationship matrix (VanRaden, 2008), but these frequencies are rarely available. Second, how to make genomic and pedigree-based relationship matrices compatible.

Compatibility of genomic and pedigree relationship across populations is difficult as pedigree implies unrelatedness of base populations, whereas markers "show" relatedness across base populations. In a series of papers, Christensen (2012), Legarra et al. (2015) and Christensen et al. (2015) present a solution: genomic relationships should be constructed using 0.5 allelic frequencies, and pedigree relationships should refer to these allelic frequencies in the base populations; i.e. base populations must be assumed related and inbred, using the concept of metafounders. Metafounders are thus a generalization of unknown parent groups, in the spirit claimed by Kennedy (1991). Matrix $\boldsymbol{\Gamma}$ describes relationships within and across metafounders, that is, base populations. With metafounders added to the pedigree, regular methods to construct a sparse inverse of the new pedigree relationship matrix $\mathbf{A}(\boldsymbol{\Gamma})$ exist (Legarra et al. 2015).

How genomic evaluation with metafounders performs in real data and for several populations and crosses is unknown yet. The aim in this study is to revisit the dataset and models for purebred and crossbred performances in Xiang et al. (2016), and investigate the effect of ssGBLUP with metafounders. In that study, the additive genetic effects of crossbreds were split into two breed-specific additive gametic effects and a ssGBLUP method with two breed-specific partial relationship matrices was investigated. On the contrary, in the current study we employ one pedigree relationship matrix containing two metafounders and all animals across the three populations.

MATERIALS AND METHODS

Data

The data set analysed is the same as in Xiang et al. (2016), and is only briefly presented here; further details can be found in Xiang et al. (2016). The trait of total number of piglets born (TNB) in the first parity was recorded in three pig populations: 293,339 Danish Landrace (L), 180,112 Danish Yorkshire (Y) and 10,974 two-way F1 crosses (LY) between these two pure populations. Among the F1, 7,407 had Landrace sire and

Yorkshire dam and 3,567 had Yorkshire sire and Landrace dam. The F1 were daughters of 765 L and 465 Y boars, respectively. Totally, 7,723 L and 7,785 Y were genotyped with the PorcineSNP60 BeadChip and 5,203 LY were genotyped with a 8.5K GGP-Porcine Low Density Illumina Bead SNP Chip. After SNP quality controls and imputation as described in Xiang et al. (2015), 41,009 phased genotyped SNP were available for the genotyped animals in both purebreds and crossbreds. A complete pedigree was traced from crossbred offspring to their purebred parents and backwards until the year 1994 by the DMU Trace program (Madsen, 2012). Consequently, 332,929 L, 210,554 Y, and 10,974 LY were in the pedigree. Two metafounders were added corresponding to Landrace and Yorkshire populations, respectively.

Within and across breed relationships using metafounders

In this research, we fit additive models to the data as in Wei and Van der Werf (1994) and Lo et al (1997), where non-additive effects at the individual level ignored, but modelled at the mean population level and using the three-trait (two purebreds and crossbred) parametrization with genetic correlations. Legarra et al. (2015) defined an additive relationship matrix $\mathbf{A}(\boldsymbol{\Gamma})$ across all animals with the relationships among the base populations (metafounders) that is determined by a positive definite matrix $\boldsymbol{\Gamma}$. The matrix $\boldsymbol{\Gamma}$ consists of elements representing relatedness for base animals within or across breeds. In this study, there are two purebred populations, and thus two metafounders, and the matrix $\boldsymbol{\Gamma} = \begin{bmatrix} \gamma_L & \gamma_{L,Y} \\ \gamma_{L,Y} & \gamma_Y \end{bmatrix}$, where γ_L and γ_Y are relatedness for base animals within L and Y, respectively; $\gamma_{L,Y}$ is the relatedness across base populations L and Y. The $\mathbf{A}(\boldsymbol{\Gamma})$ matrix is defined as follows. First, for the metafounders L and Y self-relationships were set to $a_{11} = \gamma_L$ and $a_{22} = \gamma_Y$ and the relationships $a_{21} = a_{12} = \gamma_{L,Y}$ (in other words, the upper left corner of $\mathbf{A}(\boldsymbol{\Gamma})$ is set to $\boldsymbol{\Gamma}$). The remaining elements in the $\mathbf{A}(\boldsymbol{\Gamma})$ matrix are defined by the usual tabular rules; see Legarra et al. (2015):

$$a_{ii} = 1 + 0.5a_{sd}$$

$$a_{ij} = 0.5(a_{sj} + a_{dj}),$$

where the s and d are sire and dam of animal i. Then, following the recursive process, the whole $\mathbf{A}(\boldsymbol{\Gamma})$ matrix is defined. Explicit construction of $\mathbf{A}(\boldsymbol{\Gamma})$ is not needed, because its inverse is constructed directly.

In principle, the matrix $\boldsymbol{\Gamma}$ should be determined by observed phenotypes and marker genotypes, but in practice, it can be estimated by observed marker genotypes only (Christensen, 2012). Christensen (2012) used maximum likelihood to estimate $\boldsymbol{\Gamma}$, while Legarra et al. (2015) suggested using the method of moments based on summary statistics (Legarra et al., 2015). García-Baccino et al. (2017) compared several ways of estimating $\boldsymbol{\Gamma}$ in a simulation study with a single metafounder and showed that generalized least squares (GLS) and maximum likelihood obtained the most accurate $\boldsymbol{\Gamma}$. Therefore, in this study, GLS was used to estimate $\boldsymbol{\Gamma}$.

According to Christensen (2012) and García-Baccino et al. (2017), the relatedness within breed $\gamma_{LL} = 8\sigma^2_{p_{LL}}$ and $\gamma_Y = 8\sigma^2_{p_{YY}}$, where $\sigma^2_{p_{LL}}$ and $\sigma^2_{p_{YY}}$ are the variances of true (but usually unobserved) allelic frequencies in the base population L and Y, respectively; the relatedness across L and Y is $\gamma_{L,Y} = 8cov(p_L, p_Y) = 8\sigma_{p_L,p_Y}$, where σ_{p_L,p_Y} is the covariance between allelic frequencies across all loci of individuals in Landrace and Yorkshire populations. An explanation is as follows. Cockerham (1969) observed that $\theta = \sigma^2_p/p(1-p)$ was "the coancestry of one population with itself" where σ^2_p refers to the variance of *existing* allelic frequencies in a population whereas p refers to the (*assumed*) allelic frequencies in the meta-population from which this population is conceptually drawn. In other words, θ is the covariance of the numeric values of two alleles drawn at random from such population. In the metafounder method, p is assumed to be 0.5 for all loci, by which we impose that the meta-population has $p = 0.5$. Thus, and because relationship is twice the coancestry, substituting $p = 0.5$ in the above expression yields $\gamma = 8\sigma^2_p$. The base allelic frequencies that constitute σ^2_p are estimated by GLS (equivalently, BLUP), as described in the next paragraphs.

The procedure for estimating allele frequencies is as follows. Define regular additive genetic relationship matrices $\boldsymbol{A}_{LL}$ and $\boldsymbol{A}_{YY}$ according to the pedigrees of Landrace and Yorkshire, respectively. Gene content at one marker is the number of copies of a particular reference allele (e.g coded as [0,1,2] for genotypes *AA*, *AB* and *BB*) (Falconer and Mackay, 1996). The gene content can be seen as a quantitative trait with heritability of 1 and all variation is strictly additive genetic (Forneris et al., 2015). The mean of gene content in the base population is $\mu_i = 2p_i$, where p_i is the allelic frequency at the base population, whereas its variance is $2p_iq_i$ with $q_i = 1 - p_i$. The covariance of gene contents between two individuals is a function of coancestry (Cockerham, 1969) and it equals $A_{ij}(2p_iq_i)$, where A_{ij} is the additive relationship between two individuals. Thus, a linear model for gene content can be written as:

$$\boldsymbol{m}_{ij} = \mathbf{1}\mu_i + \boldsymbol{W}\boldsymbol{u}_{ij} + \boldsymbol{e}_{ij},$$

where $\boldsymbol{m}_{ij}$ is a vector with genotypes in the form [0,1,2] for locus i across all j animals; overall mean $\mu_i = 2p_i$ is the mean of gene content for each locus i; random effect $\boldsymbol{u}_{ij}$ is deviation of each individual from this mean, following a multivariate normal distribution, $\boldsymbol{u}_{ij} \sim N(\mathbf{0}, \boldsymbol{A}2p_iq_i)$ and $\boldsymbol{A}$ is the regular additive genetic relationship matrix. In this study, the above mentioned $\boldsymbol{A}_{LL}$ and $\boldsymbol{A}_{YY}$ were used for Landrace and Yorkshire populations, respectively. $\boldsymbol{W}$ is an incidence matrix relating individuals to genotypes; $\boldsymbol{e}_{ij}$ is an error term, with $\sigma^2_e = 0.001$ so that the heritability is almost 1. This model has been independently proposed by McPeek et al. (2004) and Gengler et al. (2007).

For each locus i, two separate BLUPs (one for Landrace and one for Yorkshire, with respective matrices $\boldsymbol{A}^{-1}_{LL}$ or $\boldsymbol{A}^{-1}_{YY}$) were used to estimate, for each locus, the respective means μ_{iL} and μ_{iY} using BLUPF90 (Misztal et

al., 2002). Then the allelic frequency $\hat{p}_i$ was calculated as half of the $\hat{\mu}_i$. Across all the loci, the empirical variance $\sigma_p^2 = var(p_i)$ can be obtained, and finally $\gamma_L = 8\sigma_{p_{LL}}^2$, $\gamma_Y = 8\sigma_{p_{YY}}^2$ and $\gamma_{L,Y} = 8\sigma_{p_L,p_Y}$ are estimated.

ssGBLUP with metafounders

According to Legarra et al. (2015) and Christensen et al. (2015), with metafounders, the inverse of the relationship matrix that combined the pedigree and marker information $\mathbf{H}(\boldsymbol{\Gamma})^{-1}$ is:

$$\mathbf{H}(\boldsymbol{\Gamma})^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}(\boldsymbol{\Gamma})_{22}^{-1} \end{bmatrix} + \mathbf{A}(\boldsymbol{\Gamma})^{-1},$$

where $\mathbf{G}$ is a matrix including genomic relationships for genotyped individuals; $\mathbf{A}(\boldsymbol{\Gamma})_{22}$ is a pedigree-based relationship matrix across genotyped individuals. The matrix $\mathbf{G}$ is as in Christensen et al. (2015):

$$\mathbf{G} = \left(\mathbf{m} - \mathbf{1}\mathbf{1}'\right)\left(\mathbf{m} - \mathbf{1}\mathbf{1}'\right)' /s,$$

where $\mathbf{m}$ is the allele contents matrix with entries 0, 1, 2 and s is a scaling parameter. García-Baccino et al. (2017) showed that s is equal to half of the number of markers. In this study, the number of markers finally used were 41,009 SNP markers so that $s = 20504.5$. To accommodate that not all genetic variance is captured by marker genotypes, $\mathbf{G}$ is replaced by $\mathbf{G}_\omega = \mathbf{G}(1-\omega) + \omega\mathbf{A}(\boldsymbol{\Gamma})$, where $\omega = 0.4$ is the proportion of genetic variance in the genotyped individuals not attributable to markers. We used the value of 0.4 as this was the optimal value in Xiang et al (2016) for prediction, but it could alternatively be computed using variance component estimation (Christensen and Lund, 2010). Matrix $\mathbf{A}(\boldsymbol{\Gamma})^{-1}$ is constructed using inversion of $\boldsymbol{\Gamma}$ followed by Henderson's rules (1976). Matrix $\mathbf{A}(\boldsymbol{\Gamma})_{22}^{-1}$ is the inverse of matrix $\mathbf{A}(\boldsymbol{\Gamma})_{22}$, which is constructed following the algorithm by Colleau (2002).

Statistical model

A three-trait model, as in (Xiang et al., 2016), was used here. The model was:

$$\boldsymbol{y}_L = \boldsymbol{X}_L\boldsymbol{\beta}_L + \boldsymbol{Z}_L\boldsymbol{a}_L + \boldsymbol{W}_L\boldsymbol{s}\boldsymbol{b}_L + \boldsymbol{e}_L,$$

$$\boldsymbol{y}_Y = \boldsymbol{X}_Y\boldsymbol{\beta}_Y + \boldsymbol{Z}_Y\boldsymbol{a}_Y + \boldsymbol{W}_Y\boldsymbol{s}\boldsymbol{b}_Y + \boldsymbol{e}_Y,$$

$$\boldsymbol{y}_{LY} = \boldsymbol{X}_{LY}\boldsymbol{\beta}_{LY} + \boldsymbol{Z}_{LY}\boldsymbol{c}_{LY} + \boldsymbol{e}_{LY},$$

where $\boldsymbol{y}_L$, $\boldsymbol{y}_Y$ and $\boldsymbol{y}_{LY}$ contain records of TNB in the first parity for L, Y and LY, respectively; $\boldsymbol{X}_L\boldsymbol{\beta}_L$, $\boldsymbol{X}_Y\boldsymbol{\beta}_Y$ and $\boldsymbol{X}_{LY}\boldsymbol{\beta}_{LY}$ contain fixed effects, including herd–year–season, month at farrowing, hybrid indicator of service sire (same or different breed as the sow) and covariates for the age of farrowing and its squared value;

$\boldsymbol{a}_L$ and $\boldsymbol{a}_Y$ represent purebred breeding values for L and Y, respectively; $\boldsymbol{c}_{LY}$ is the additive genetic effects of crossbred LY animals; $\boldsymbol{sb}_L$ and $\boldsymbol{sb}_Y$ are random effects of service boar for L and Y, respectively; $\boldsymbol{Z}$ and $\boldsymbol{W}$ are the respective incidence matrices; $\boldsymbol{e}_L$, $\boldsymbol{e}_Y$ and $\boldsymbol{e}_{LY}$ are overall residual effects. Assumptions were that random effects were independently normally distributed: $\boldsymbol{sb} \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2_{sb})$ and $\boldsymbol{e} \sim N(\mathbf{0}, \boldsymbol{I}\sigma^2_e)$. The genetic variance and covariance structure for the additive genetic effects was:

$$var\begin{bmatrix} \boldsymbol{a}_L \\ * \\ \boldsymbol{c}_L \\ * \\ \boldsymbol{a}_Y \\ \boldsymbol{c}_Y \\ * \\ * \\ \boldsymbol{c}_{LY} \end{bmatrix} = \mathbf{H}(\boldsymbol{\Gamma}) \otimes \mathbf{G}_0,$$

where $\boldsymbol{a}$ stands for breeding values for purebred performance (when mated within breed); $\boldsymbol{c}_L$ and $\boldsymbol{c}_Y$ stand for breeding values for crossbred performance (when mated to the other breed); $\boldsymbol{c}_{LY}$ is additive genetic effects for crossbred animals; $*$ denotes artificial random vectors and $\mathbf{G}_0$ is a 3×3 genetic variance components:

$$\mathbf{G}_0 = \begin{bmatrix} \sigma^2_{A_L} & \sigma_{A_L A_Y} & \sigma_{A_L A_{LY}} \\ \sigma_{A_Y A_L} & \sigma^2_{A_Y} & \sigma_{A_Y A_{LY}} \\ \sigma_{A_L A_{LY}} & \sigma_{A_Y A_{LY}} & \sigma^2_{A_{LY}} \end{bmatrix},$$

where elements in the diagonal are additive genetic variances within each population, $\sigma_{A_L A_Y}$ is the genetic covariance between breeding values for purebred L and Y, and $\sigma_{A_L A_{LY}}$ and $\sigma_{A_Y A_{LY}}$ are genetic covariances between breeding values for purebred and crossbreed performances for L and Y, respectively. The three-trait model accounts for potential genotype by environment or genotype by genotype (dominance, epistasis) interactions (Wei and Van der Werf, 1994; Christensen et al. 2014).

$\mathbf{H}(\boldsymbol{\Gamma})$ is an additive genetic relationship matrix across three populations, which has 9 blocks:

$$\mathbf{H}(\boldsymbol{\Gamma}) = \begin{bmatrix} \boldsymbol{H}^{\gamma}_{L,L} & \boldsymbol{H}^{\gamma}_{L,Y} & \boldsymbol{H}^{\gamma}_{L,LY} \\ \boldsymbol{H}^{\gamma}_{Y,L} & \boldsymbol{H}^{\gamma}_{Y,Y} & \boldsymbol{H}^{\gamma}_{Y,LY} \\ \boldsymbol{H}^{\gamma}_{LY,L} & \boldsymbol{H}^{\gamma}_{LY.Y} & \boldsymbol{H}^{\gamma}_{LY,LY} \end{bmatrix},$$

i.e. three blocks for within populations, $\boldsymbol{H}^{\gamma}_{L,L}$, $\boldsymbol{H}^{\gamma}_{Y,Y}$ and $\boldsymbol{H}^{\gamma}_{LY,LY}$, two blocks for between purebreds, $\boldsymbol{H}^{\gamma}_{L,Y}$ and $\boldsymbol{H}^{\gamma}_{Y,L}$, and the remaining four blocks for between purebred and crossbred animals, $\boldsymbol{H}^{\gamma}_{L,LY}$, $\boldsymbol{H}^{\gamma}_{Y,LY}$, $\boldsymbol{H}^{\gamma}_{LY,L}$ and

$\boldsymbol{H}^{\gamma}_{LY,Y}$. In this study, different genomic information was used to construct this additive genetic relationship matrix, which will be described more detailed in the following section.

The differences between the model here and the model in (Xiang et al., 2016) are as follows. First, in the previous study, a breed-of-origin model was used, where breeds of origin on genotypes were traced and the rules of García-Cortés and Toro (2006) were applied to split the genetic effects of crossbred animals into two independent breed-of-origin terms. Therefore, two breed-specific partial relationship matrices were used. The assumption behind this is that base individuals within and across different breeds are totally unrelated. In this study, to the contrary, it is assumed that base individuals (and marker effects) are related within and across different breeds. Hence, pedigree relationships are specified across all the animals in the pedigree, and genomic relationships are also specified across the three populations. Consequently only one combined relationship matrix is specified in this study. Second, genetic parameters are different in the two studies. The parameters in this study are not corresponding to the usual genetic variances where the individuals in the base population are unrelated (Legarra et al. 2015). Besides, genetic correlation between purebred L and Y effects ($\sigma_{A_L A_Y}$) did not exist in the previous study. Finally, in this study the genetic variance in the crossbreds has one component ($\sigma^2_{A_{LY}}$) whereas in the previous study it had two, one coming from each breed.

Scenarios

Scenarios used in this study were the same as those in (Xiang et al., 2016). Variance and covariance parameters, heritabilities and genetic correlations between purebred and crossbred performances (r_{PC}) were first investigated in the three scenarios: *Nogen*, where the relationship matrix $\mathbf{H}(\boldsymbol{\Gamma})$ was replaced by $\mathbf{A}(\boldsymbol{\Gamma})$, which was constructed based on pedigree information; *Genpure*, where pedigree information in combination with purebred genomic information (7,723 L and 7,785 Y) was used to construct the $\mathbf{H}(\boldsymbol{\Gamma})$; *Genall*, where $\mathbf{H}(\boldsymbol{\Gamma})$ was constructed based on all purebred and crossbred genomic information (7,723 L, 7,785 Y and 5,203 LY) and pedigree information. Initial convergence of REML was very slow. Thus, initial guesses of variance components were estimated by Gibbs sampling using GIBBS1F90 (Misztal et al., 2002). A total of 300,000 iterations of the sampler were made, with the first 5,000 iterations discarded as burn-in samples and every 25th sample included in the posterior analysis. The posterior means were used as starting values for REMLF90 (Misztal et al., 2002), which converged after a few (≤ 3) iterations. Legarra et al. (2015) and Christensen et al. (2015) pointed out that the estimated genetic parameters (unscaled) in the method with metafounders were not directly comparable with the usual genetic parameters where the animals in base populations were assumed to be unrelated. To compare the estimated genetic parameters with those in (Xiang et al., 2016), genetic parameters need to be multiplied by $(1-\gamma_b/2)$, corresponding to the (co)variances among the unrelated breed b animals (scaled). More specifically, the scaled genetic variances of purebred performances (σ^2_a) were $\sigma^2_{A_L}(1-\gamma_L/2)$ within L and $\sigma^2_{A_Y}(1-\gamma_Y/2)$ within Y; the scaled

genetic variances of crossbred performances (σ_c^2) were $\sigma_{A_{LY}}^2(1-\gamma_L/2)$ for L and $\sigma_{A_{LY}}^2(1-\gamma_Y/2)$ for Y; the scaled genetic covariance between purebred and crossbred performances ($\sigma_{a,c}$) were $\sigma_{A_L A_{LY}}(1-\gamma_L/2)$ and $\sigma_{A_Y A_{LY}}(1-\gamma_Y/2)$ for L and Y, respectively. Heritabilities for purebred performance (h_{LL}^2 for L and h_{YY}^2 for Y) were defined as the ratio of the scaled additive genetic variances for purebred performance ($\sigma_a^2 = \sigma_{A_L}^2\left(1-\frac{\gamma_L}{2}\right)$ for L and $\sigma_a^2 = \sigma_{A_Y}^2\left(1-\frac{\gamma_Y}{2}\right)$ for Y) to phenotypic variances ($\sigma_{P_L}^2 = \sigma_{A_L}^2\left(1-\frac{\gamma_L}{2}\right) + \sigma_{sb_L}^2 + \sigma_{e_L}^2$ for L and $\sigma_{P_Y}^2 = \sigma_{A_Y}^2\left(1-\frac{\gamma_Y}{2}\right) + \sigma_{sb_Y}^2 + \sigma_{e_Y}^2$ for Y). The genetic correlations between purebred and crossbred performances (r_{pc_L} for L and r_{pc_Y} for Y) were calculated as $r_{pc_L} = \frac{\sigma_{A_L A_{LY}}}{\sqrt{\sigma_{A_L}^2 \sigma_{A_{LY}}^2}}$ and $r_{pc_Y} = \frac{\sigma_{A_Y A_{LY}}}{\sqrt{\sigma_{A_Y}^2 \sigma_{A_{LY}}^2}}$.

Model-based reliabilities of crossbred performance for purebred boars (765 L and 465 Y) were calculated in the mentioned three different scenarios. These boars were divided into genotyped (656 L and 443 Y) and non-genotyped (109 L and 22 Y) subgroups and mean model-based reliabilities were computed in each subgroup. Mean model-based reliability was calculated as (Van Vleck, 1993): $r^2 = \sum_{i=1}^{n}(1-\frac{SEP_i^2}{Hii*\sigma_{A_{LY}}^2})/n$, where SEP_i was the standard error of prediction for animal i (obtained by inversion using blupf90); Hii is the self-relationship coefficient in $\mathbf{H}(\boldsymbol{\Gamma})$ for animal i; $\sigma_{A_{LY}}^2$ was the additive genetic variance in crossbreds (unscaled additive genetic variance for crossbred performance) and n was the number of purebred boars studied.

For comparison of the performance of the different models, predictive abilities for crossbred animals in the validation populations were studied in the same different scenarios as in (Xiang et al., 2016). The farrowing date of January 1, 2012 was used as cut-off date to divide recorded sows into training and validation populations. For the genotyped boars the birth date of January 1, 2011 was used as cut-off date. Consequently, training population contained phenotypes recorded in 240,543 L, 139,868 Y and 6,779 LY sows and genomic information of 2,210 L, 2,161 Y and 2,357 LY. The validation population for crossbred performance included 4,195 LY, among which 2,846 were genotyped. The predictive ability was measured by cross-validation as the correlation between phenotypes (corrected for fixed and non-genetic random effects) and estimated additive genetic effects for crossbred animals ($cor(\boldsymbol{Y}_c, \widehat{\boldsymbol{c}_{LY}})$). To compare the predictive abilities with those in (Xiang et al., 2016), the corrected phenotypes $\boldsymbol{Y}_c$ used in that study were also used here.

Additive genetic effects $\widehat{\boldsymbol{c}_{LY}}$ were estimated in the following four scenarios, concerning different amount of genomic information for constructing the relationship matrix $\mathbf{H}(\boldsymbol{\Gamma})$: *Nogen_T*, where $\mathbf{H}(\boldsymbol{\Gamma})$ was constructed based on pedigree information (thus, $\mathbf{H}(\boldsymbol{\Gamma}) = \mathbf{A}(\boldsymbol{\Gamma})$); *Genpure_T*, where $\mathbf{H}(\boldsymbol{\Gamma})$ was constructed based on the combined pedigree and purebred genomic information (2,210 L and 2,161 Y); *Genpc_T*, where the $\mathbf{H}(\boldsymbol{\Gamma})$ was constructed based on pedigree information and genotypes of 2,210 L, 2,161Y and 2,357 LY that belonged to

the training population; *Genall_T*, where all information in *Genpc_T* in combination with extra 2,846 validation crossbred genotyped animals was used to construct $\mathbf{H}(\boldsymbol{\Gamma})$. These scenarios are the same as in Xiang et al. (2016) for comparison purposes, but we note that scenario *Nogen_T* has no practical relevance because (a) the connection across the two breeds is not informative and (b) genotypes were used to estimate matrix $\boldsymbol{\Gamma}$. For *Genall_T*, the validation population was divided into two subgroups of genotyped and non-genotyped animals and the validation correlations were made within the two subgroups as well as in the whole validation population. A Hotelling-Williams t-test at a 5% confidence level was applied to evaluate the significance for the differences of validation correlations between scenarios. Furthermore, the regression coefficients of $\boldsymbol{Y}_c$ on $\widehat{\boldsymbol{c}_{LY}}$ were explored to check the possible biases of predictions.

RESULTS AND DISCUSSION

Estimations of Γ

The self-relationship of the metafounder L or the relationship coefficient across animals in the base population of Landrace ($\hat{\gamma}_L$) was 0.756. For Yorkshire, the $\hat{\gamma}_Y = 0.730$ was close to $\hat{\gamma}_L$. The similar γ's indicated that the additive relationships among base animals for Landrace and for Yorkshire were similar. According to Legarra et al. (2015), γ was determined by (or a measure of) the effective population size, which is, in both breeds, around 55 (personal communication, Tage Ostersen, Danish Pig Research Center). In this study, both γ's were smaller than 1, which lead to a negative inbreeding coefficient of metafounder ($F = \gamma - 1$) (Legarra et al. 2015). This negative inbreeding represents an excess of heterozygotes relative to the average of the population and indicated in most cases, gametes of base animals were not identical (Legarra et al., 2015). In other words, the base population has a large genetic variability. For an infinite population, the inbreeding of the population is -1 (i.e. all animals are heterozygotes). The relationship coefficient between base populations of Landrace and Yorkshire ($\hat{\gamma}_{L,Y}$) was 0.259. The $\hat{\gamma}_{L,Y}$ was larger than 0, suggesting there was an overlap between their ancestor populations, which was in line with the mixture history of Landrace and Yorkshire (King, 1991; Wang et al., 2013). As a whole, the estimated relationship matrix among the base populations (metafounders) is $\boldsymbol{\Gamma} = \begin{bmatrix} \hat{\gamma}_L & \hat{\gamma}_{L,Y} \\ \hat{\gamma}_{L,Y} & \hat{\gamma}_Y \end{bmatrix} = \begin{bmatrix} 0.756 & 0.259 \\ 0.259 & 0.730 \end{bmatrix}$.

Variance components, heritabilities and genetic correlations

Estimates of variance components, calculated heritabilities and genetic correlations between purebred and crossbred performances for Landrace and Yorkshire in each scenario are shown in Table 1. For all the presented genetic parameters, they were scaled to be comparable with the usual genetic variance where the founders of pedigree were assumed to be unrelated (Legarra et al., 2015). From the table, it can be seen that for each breed, genetic parameters for purebred performance were nearly the same across different scenarios.

For crossbred performance, genetic variances in scenarios with genomic information were slightly larger than those in *Nogen* scenario. When comparing these estimated variance components with results in our previous study (Xiang et al., 2016), additive genetic variances for purebred performances were almost identical, but the genetic variances for crossbred performance were slightly different, ranging around 0.1~0.2. As for the non-genetic parameters, estimates were very close to results in the previous study (Xiang et al., 2016). Heritabilities (h^2) were constant for Landrace (0.11) and for Yorkshire (0.09) in different scenarios. Genetic correlations between purebred and crossbred performances (r_{pc}) ranged from 0.73 to 0.80 for Landrace and from 0.63 to 0.70 for Yorkshire. Slight differences of the estimated genetic correlations were observed between the two breeds. Landrace showed a bit higher r_{pc} than that for Yorkshire. These values were similar to our previous results, which show a range of 0.70~0.78 for Landrace and 0.57~0.68 for Yorkshire (Xiang et al., 2016). The genetic correlations between purebred L and Y effects were 0.23 and 0.30 as estimated in *Genpure* and *Genall* scenarios; we consider that the genetic correlations between L and Y with *Nogen* scenario is unreliable because only ancestral relationships in **Γ** are used and do not present the genetic correlations between L and Y in *Nogen* scenario. The genetic correlations between effects of different populations have rarely been estimated; Legarra et al. (2014) estimated values of 0.3 to 0.5 across sheep breeds and Karoui et al. (2012) from 0 to 0.8 for dairy cattle breeds.

Table 1 Variance components [1], heritabilities[2], and genetic correlations between purebred and crossbred performances[3] with standard errors for Landrace and Yorkshire

Scenarios	Breed	σ_a^2	$\sigma_{a,c}$	σ_c^2	σ_{sb}^2	σ_e^2	$\sigma_{e_{LY}}^2$	h^2	r_{pc}
Nogen	L	1.63	0.80	0.73	0.88	12.17		0.11	0.73
		(0.09)	(0.16)	(0.14)	(0.03)	(0.05)	8.46	(0.02)	(0.11)
	Y	1.23	0.60	0.74	0.72	11.67	(0.13)	0.09	0.63
		(0.08)	(0.16)	(0.14)	(0.02)	(0.06)		(0.02)	(0.12)
Genpure	L	1.64	0.85	0.82	0.88	12.17		0.11	0.73
		(0.09)	(0.12)	(0.11)	(0.03)	(0.05)	8.40	(0.02)	(0.09)
	Y	1.22	0.71	0.84	0.71	11.68	(0.13)	0.09	0.70
		(0.08)	(0.14)	(0.11)	(0.02)	(0.06)		(0.02)	(0.11)
Genall	L	1.64	0.93	0.82	0.88	12.29		0.11	0.80
		(0.09)	(0.12)	(0.11)	(0.03)	(0.05)	8.40	(0.02)	(0.09)
	Y	1.22	0.68	0.84	0.71	11.76	(0.13)	0.09	0.67
		(0.08)	(0.13)	(0.11)	(0.02)	(0.06)		(0.02)	(0.10)

[1]Variance components for genetic parameters correspond to the usual genetic variance which is the variance among unrelated individuals in base population: σ_a^2 is additive genetic variance for purebred performance, $= \sigma_{A_L}^2\left(1-\frac{\gamma_L}{2}\right)$ for Landrace and $\sigma_{A_Y}^2\left(1-\frac{\gamma_Y}{2}\right)$ for Yorkshire; σ_c^2 is additive genetic variance for crossbred, $= \sigma_{A_{LY}}^2\left(1-\frac{\gamma_L}{2}\right)$ for Landrace and $\sigma_{A_{LY}}^2\left(1-\frac{\gamma_Y}{2}\right)$ for Yorkshire; $\sigma_{a,c}$ is genetic covariance between purebred and crossbred performances , $= \sigma_{A_L A_{LY}}\left(1-\frac{\gamma_L}{2}\right)$ for Landrace and $\sigma_{A_Y A_{LY}}\left(1-\frac{\gamma_Y}{2}\right)$ for Yorkshire; σ_{sb}^2 = variance of service-boar effect; σ_e^2 = residual variance for purebred performance; $\sigma_{e_{LY}}^2$ = residual variance for crossbred animals. Numbers between brackets are the standard errors of the corresponding parameters.

[2]Heritablity $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_{sb}^2 + \sigma_e^2)$

[3]Genetic correlation $r_{pc} = \frac{\sigma_{a,c}}{\sqrt{\sigma_a^2\sigma_c^2}} = \frac{\sigma_{A_L A_{LY}}}{\sqrt{\sigma_{A_L}^2\sigma_{A_{LY}}^2}}$ for Landrace; $= \frac{\sigma_{A_Y A_{LY}}}{\sqrt{\sigma_{A_Y}^2\sigma_{A_{LY}}^2}}$ for Yorkshire

Model-based reliability and Predictive abilities

Mean model-based reliabilities for purebred boars for their crossbred performance in different scenarios are shown in Table 2. Model-based reliabilities increased from about 0.34 to 0.48 for the Landrace breed and from about 0.27 to 0.40 for the Yorkshire breed from *Nogen* to *Genall.* Scenarios with marker information (*Genpure* and *Genall*) show higher model-based reliabilities than the pedigree-based scenario (*Nogen*). With crossbred genomic information, reliabilities were further improved in scenario *Genall.* The genotyped subgroup presented higher model-based reliabilities than the non-genotyped group across different scenarios. Comparing the results to the previous one (Xiang et al., 2016), results here were 0.03~0.10 higher than results in the corresponding scenarios in that study.

Table 2 Model-based reliabilities for crossbred performance in purebred boars

	All[1]			Genotyped[2]			Non-genotyped[3]		
	Nogen	*Genpure*	*Genall*	*Nogen*	*Genpure*	*Genall*	*Nogen*	*Genpure*	*Genall*
L	0.338	0.368	0.484	0.341	0.371	0.486	0.336	0.356	0.477
Y	0.274	0.376	0.404	0.285	0.377	0.405	0.272	0.373	0.396

[1]All = all the sires of crossbred animals, consisting of 765 Landrace and 465 Yorkshire.
[2]Genotyped = genotyped sires of crossbred animals, consisting of 656 Landrace and 443 Yorkshire.
[3]Non-genotyped = Non-genotyped sires of crossbred animals, consisting of 109 Landrace and 22 Yorkshire.

Table 3 shows predictive abilities for crossbred animals in the validation populations. The validation correlations between corrected phenotypes and estimated additive genetic effect for crossbreds $cor(\boldsymbol{Y}_c, \widehat{\boldsymbol{c}_{LY}})$ ranged from 0.084 in *Nogen_T* to 0.125 in *Genall_T*. No significant differences were detected between scenarios *Nogen_T* and *Genpure_T*, but *Genpc_T* and *Genall_T* presented significantly higher accuracies than those two scenarios. Results here were virtually the same as results in the corresponding scenarios in our previous study (Xiang et al., 2016). For scenario *Genall_T*, the genotyped subgroup (0.132) had higher $cor(\boldsymbol{Y}_c, \widehat{\boldsymbol{c}_{LY}})$ than that in the non-genotyped subgroup (0.109). The value of $cor(\boldsymbol{Y}_c, \widehat{\boldsymbol{c}_{LY}})$ in the non-genotyped subgroup in *Genall_T* was still higher than that in the other three scenarios. The $cor(\boldsymbol{Y}_c, \widehat{\boldsymbol{c}_{LY}})$ for the genotyped subgroup in *Genall_T* was about 30% higher than that in *Genpc_T*. The regression coefficients of estimated additive genetic effects for crossbred animals are also shown in Table 3. In general, the regression coefficients were close to 1. The regression coefficients in scenarios with genomic information were closer to 1 than that in *Nogen_T*. Among scenarios with genomic information, in terms of regression coefficients, there was no clear pattern showing which scenario performed better than others, although regression coefficient in *Genpc_T* was closest to 1. When comparing these regression coefficients with those in our previous study (Xiang et al., 2016), slightly smaller regression coefficients (~0.02) were found in this study.

Table 3 Predictive abilities for crossbred animals in the validation populations

	Nogen_T	*Genpure_T*	*Genpc_T*	*Genall_T*		
				All	Genotyped	Non-Genotyped
$cor(\boldsymbol{Y}_c, \widehat{\boldsymbol{c}_{LY}})$[1]	0.084[a]	0.090[a]	0.099[b]	0.125[c]	0.132	0.109
Regression coefficients[2]	1.204	0.978	1.043	1.044	1.034	1.059

[1] $cor(\boldsymbol{Y}_c, \widehat{\boldsymbol{c}_{LY}})$ is correlation coefficients between corrected phenotypes and estimated additive genetic effects on crossbred animals; different superscripts of small letters among scenarios indicate significant differences ($p<0.05$) by Hotelling-Williams t-test
[2]Regression coefficients of corrected phenotypes on estimated additive genetic effects

Predictive abilities empirically assess the fitness of the models, whereas model-based reliabilities assess individual accuracies. Unfortunately, assessing the predictive ability for breeding values of boars for their crossbred performance was very difficult, as the boars in our study have a small number of crossbred daughters (around 5). Results in this work are similar to Xiang et al. (2016) (except for *Nogen_T*), which is in line with Christensen (2012) that the ssGBLUP with an adjusted pedigree-based relationship matrix $\mathbf{A}(\boldsymbol{\Gamma})$ should perform equal or better than methods based on adjusting the marker-based relationship matrix. This study also shows that the ssGBLUP method with metafounders is applicable for genomic evaluation of purebred and crosses. For scenarios with crossbred genomic information, results of reliabilities and validation correlations in this study seemed to be slightly higher than those in the previous study. In practice, the method used in the previous study requires phasing the data and inferring breed origin, which is a bit cumbersome. Thus, all else being equal, the ssGBLUP method with metafounders (Legarra et al., 2015) is more convenient to implement than ssGBLUP with breed-specific partial relationship matrices (Christensen et al., 2014).

Modelling breeds and their crosses using genetic groups (metafounders in this case) is not common in pigs, but is customary in ruminants, e.g. Arnold et al. (1992). It is more common to use a multiple-trait approach (Wei and Van der Werf, 1994; Lo et al., 1997; Lutaaya et al., 2001). Here we combine both in a way that is at the same time sensible to the specialties of markers (genomic relationship across breeds) and to the fact that the three populations differ in genetic and environmental background. More specifically, biological dominance and epistasis provoke that substitution effects of the causal genes differ from one population to the other. For instance, the substitution effect $\alpha = a + (q - p)d$ differs from matings within the same breed to matings with another breed because the allelic frequencies change. With differences in the environment the allele substitution effects also differ between populations. However, the fact that effects may be similar across breeds is accounted for through correlations. This is markedly different from the assumption in Xiang et al. (2016) that relationships between lines (and, accordingly, the correlation between marker effects) are 0.

All in all, a method of single-step genomic evaluation with metafounders was successfully implemented in the Danish Landrace, Yorkshire and crossbred populations in this study. The estimated variance components were generally similar to those parameters in the previous study and the model-based reliabilities and predictive abilities were at least as good as those obtained by a single-step genomic evaluation using breed-specific partial relationship matrices in the previous study (Xiang et al., 2016). There is good agreement across studies, which is reassuring.

CONCLUSION

Effective population sizes are similar for Landrace and Yorkshire and the ancestor populations for Landrace and Yorkshire are related. In the method of ssGBLUP with metafounders, models with crossbred genomic information improve model-based reliability for crossbred performance for purebred boars and predictive abilities for validated crossbred animals while reducing regression coefficients. The single-step genomic evaluation method with metafounders performs at least as well as the breed-origin based ssGBLUP in prediction.

LITERATURE CITED

Arnold, J. W., J. K. Bertrand, and L. L. Benyshek. 1992. Animal Model for Genetic Evaluation of Multibreed Data. J Anim Sci. 70: 3322–32.

Christensen, O. F. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet Sel Evol. 44: 37.

Christensen, O. F., and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 42: 2.

Christensen, O. F., A. Legarra, M. S. Lund, and G. Su. 2015. Genetic evaluation for three-way crossbreeding. Genet Sel Evol. 47: 98.

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

Cockerham, C. C. 1969. Variance of gene frequencies. Evolution. 23: 72-84.

Colleau, J. J. 2002. An indirect approach to the extensive calculation of relationship coefficients. Genet Sel Evol. 34: 409-21.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to quantitative genetics. Pearson Education Limited, Harlow, UK.

Forneris, N.S., A. Legarra, Z. G. Vitezica, S. Tsuruta, I. Aguilar, I. Misztal, and R. J. Cantet. 2015. Quality control of genotypes using heritability estimates of gene content at the marker. Genetics. 199: 675-81.

García-Baccino, C. A., A. Legarra, O. F. Christensen, I. Misztal, I. Pocrnic, Z. G. Vitezica, and R. J. Cantet. 2017. Metafounders are related to Fst fixation indices and reduce bias in single-step genomic evaluations. Genet Sel Evol. 49: 34.

García-Cortés L. A., M. Á. Toro. 2006. Multibreed analysis by splitting the breeding values. Genet Sel Evol. 38: 601-15.

Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal. 1: 21-28.

Karoui, S., M. J. Carabaño, C. Díaz, and A. Legarra. 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet Sel Evol. 44: 39.

Kennedy, B. W. 1991. CR Henderson: The unfinished legacy. J Dairy Sci. 74: 4067-81.

King, J. 1991. Pig breeds of the world:Their distributions and adaptation. In: K. Maijala (ed.) Genetic resources of pig, sheep and goat. p 52-53. Elsevier Science Publishers, Dunfermline, UK.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 92: 4656-63.

Legarra, A., G. Baloche, F. Barillet, J. M. Astruc, C. Soulas, X. Aguerre, F. Arrese, L. Mintegi, M. Lasarte, F. Maeztu, and I. B. de Heredia. 2014. Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J Dairy Sci. 97:3200-12.

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics 200: 455-68.

Lo, L. L., R. L. Fernando, and M. Grossman. 1997. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. J Anim Sci. 75: 2877-84.

Lutaaya, E., I. Misztal, J. W. Mabry, T. Short, H. H. Timm, and R. Holzbauer. Joint Evaluation of Purebreds and Crossbreds in Swine. J Anim Sci. 80: 2263–66.

Madsen, P. 2012. DMU Trace, A program to trace the pedigree for a subset of animals from a large pedigree file. Version 2., Center for Quantitative Genetics and Genomics. Dept. of Molecular Biology and Genetics. Aarhus University, Denmark.

McPeek, M. S., X. Wu, O. Carole. 2004. Best Linear Unbiased Allele-Frequency Estimation in Complex Pedigrees. Biometrics. 60: 359-67.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. BLUPF90 and related programs (BGF90). In: 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91: 4414-23.

Van Vleck, L. 1993. Variance of prediction error with mixed model equations when relationships are ignored. Theor Appl Genet. 85: 545-49.

Wang, L., P. Sørensen, L. Janss, T. Ostersen, and D. Edwards. 2013. Genome-wide and local pattern of linkage disequilibrium and persistence of phase for 3 Danish pig breeds. BMC Genet. 14: 115.

Wei, M., and Van der Werf, J. H. J. 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim Prod. 59: 401-13.

Xiang, T., P. Ma, T. Ostersen, A. Legarra, and O. F. Christensen. 2015. Imputation of genotypes in Danish purebred and two-way crossbred pigs using low-density panels. Genet Sel Evol. 47: 54.

Xiang, T., B. Nielsen, G. Su, A. Legarra, and O. F. Christensen. 2016. Application of single-step genomic evaluation for crossbred performance in pig. J Anim Sci. 94: 936-48.

CHAPTER 5: PAPER IV.

Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs

Tao Xiang[1,2*], Ole Fredslund Christensen[1], Zulma Gladis Vitezica[3], Andres Legarra[2]

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark

[2]INRA, UR1388 GenPhySE, CS-52627, F-31326 Castanet-Tolosan, France

[3]Université de Toulouse, INP, ENSAT, GenPhySE, F-31326 Castanet-Tolosan, France

[*]Corresponding author

This paper was published in *Genetics selection evolution* (2016) 48: 92. DOI: 10.1186/s12711-016-0271-4

Abstract

Background

Improved performance of crossbred animals is partly due to heterosis. One of the major genetic bases of heterosis is dominance, but it is seldom used in pedigree-based genetic evaluation of livestock. Recently, a trivariate genomic best linear unbiased prediction (GBLUP) model including dominance was developed, which can distinguish purebreds from crossbred animals explicitly. The objectives of this study were: (1) methodological, to show that inclusion of marker-based inbreeding accounts for directional dominance and inbreeding depression in purebred and crossbred animals, to revisit variance components of additive and dominance genetic effects using this model, and to develop marker-based estimators of genetic correlations between purebred and crossbred animals and of correlations of allele substitution effects between breeds; (2) to evaluate the impact of accounting for dominance effects and inbreeding depression on predictive ability for total number of piglets born (TNB) in a pig dataset composed of two purebred populations and their crossbreds. We also developed an equivalent model that makes the estimation of variance components tractable.

Results

For TNB in Danish Landrace and Yorkshire populations and their reciprocal crosses, the estimated proportions of dominance genetic variance to additive genetic variance ranged from 5 to 11%. Genetic correlations between breeding values for purebred and crossbred performances for TNB ranged from 0.79 to 0.95 for Landrace and from 0.43 to 0.54 for Yorkshire across models. The estimated correlation of allele substitution effects between Landrace and Yorkshire was low for purebred performances, but high for crossbred performances. Predictive ability for crossbred animals was similar with or without dominance. The inbreeding depression effect increased predictive ability and the estimated inbreeding depression parameter was more negative for Landrace than for Yorkshire animals and was in between for crossbred animals.

Conclusions

Methodological developments led to closed-form estimators of inbreeding depression, variance components and correlations that can be easily interpreted in a quantitative genetics context. Our results confirm that genetic correlations of breeding values between purebred and crossbred performances within breed are positive and moderate. Inclusion of dominance in the GBLUP model does not improve predictive ability for crossbred animals, whereas inclusion of inbreeding depression does.

Background

Crossbreeding is primarily and intensively applied in meat production systems [1], especially for swine and poultry. Crossbreeding capitalizes on heterosis effects and complementarity between breeds, and results in an increased performance of crossbred animals compared to purebred animals [1]. In terminal crossbreeding systems, selection on purebred animals to maximize their crossbred performance is the ultimate goal [2, 3]. Due to the existence of genotype-by-environment interaction effects and non-additive genetic effects in combination with different allele frequencies in different breeds [3, 4], the genetic correlation of breeding values between purebred and crossbred performances (r_{PC}) is usually lower than 1 [1, 5], and therefore, purebred performance under nucleus conditions may not be an optimal predictor for crossbred performance in commercial animals [4, 6].

One of the major genetic bases of heterosis is dominance [7, 8]. At the level of gene action, dominance is due to interactions between alleles at the same locus [9]. In pedigree-based genetic evaluation, dominance is rarely included because large-scale datasets that comprise a high proportion of full sibs are required to obtain accurate estimates and because the computational complexity is high [10]. With the recent availability of single nucleotide polymorphism (SNP) information and the development of genomic selection, estimation of the dominance effects of SNPs has become more feasible [11, 12].

Genomic evaluation has been successfully used in purebred [13, 14] and crossbred populations [15-17]. However, these studies generally ignore the dominance effects. A number of studies have been carried out on genomic evaluation including dominance effects using either simulated [18] or real purebred data [9, 12].

Recently, several studies [19, 20] have tried to extend genomic evaluation including dominance effects from purebred performance to crossbred performance. However, they either used genomic information on purebred animals only [19] or applied a genomic model that assumed that all animals belong to a single population, and thus the variance components were estimated based only on the genotyped crossbred animals [20]. Nevertheless, combining purebred and crossbred information is essential to implement genetic evaluation for crossbred performance [1, 19]. Furthermore, because of genotype-by-environment interaction effects and different patterns of linkage disequilibrium (LD) between SNPs and quantitative trait loci (QTL), the effects of SNPs may be breed-specific [21]. To overcome these issues, a trivariate genomic best linear unbiased predictor (GBLUP) model that explicitly distinguishes between purebred and crossbred data and includes dominance was recently developed by Vitezica et al. [22]. This model allowed the estimation of different, yet correlated, additive and dominance marker effects in crossbred and purebred individuals. However, the empirical predictive ability of the trivariate GBLUP model has not been evaluated yet.

Thus, the current study had the following objectives:(1) to show how genomic inbreeding can be meaningfully included in GBLUP, even for crossbred animals; (2) to estimate the variance components of additive and dominance genetic effects by using data on total number of piglets born (TNB) in two Danish purebred and one crossbred pig populations using the trivariate GBLUP model; (3) to show how to derive, from variance component estimates, estimated genetic correlations of breeding values between purebred and crossbred performances in each pure breed, and also correlations of allele substitution effects between the two pure breeds; and (4) to evaluate the impact of dominance effects from genomic information on genomic evaluation by comparing accuracies of estimated genomic values in different cross-validation scenarios.

Methods

Animals and genotypes

We begin this section with a short presentation of the data used in the study, with the aim of defining the notation for the methodological developments that follow. For this study, all datasets were provided by the Danish Pig Research Centre. Data from three Danish pig populations were analyzed simultaneously: Landrace (L), Yorkshire (Y) and their reciprocal crosses (LY). Only data on TNB data for the first parity of sows in the three populations were used. In total, there were 2126, 2218 and 5143 genotyped sows with own records on TNB for L, Y and LY, respectively. Instead of using original records, corrected phenotypic values of TNB were used as dependent variables for the trivariate GBLUP model, because the pre-correction for non-genetic effects, such as herd-year-season, month at farrowing, and service sire was more accurately achieved on a larger dataset (293,339 L, 180,112 Y, and 10,974 LY). Among the crossbred animals, 7407 LY had a Landrace sire and a Yorkshire dam, while 3567 LY had a Yorkshire dam and a Landrace sire; L and Y populations were from nucleus farms and LY from a commercial farm. The litters of purebred sows were both purebred and crossbred litters. The relationship between LY-L and LY-Y are comparable since, in both cases, parents of the F1 animals are in the purebred datasets; further details about the model used for the pre-correction are in [17]. All the purebred sows had first farrowing dates between 2003 and 2013, while the crossbred sows first farrowed between 2010 and 2013. Only five of these purebred L and Y sows were dams of the LY.

The pedigrees for both purebred and crossbred sows were available and all crossbred animals were traced back to their purebred ancestors until 1994 by the DMU Trace program [23], as was done for the larger dataset used for pre-correction. Consequently, 8227 L, 9851 Y and 5143 LY individuals were in the pedigree. The dataset of pre-corrected TNB records for genotyped individuals is termed “full genomic dataset” throughout the whole paper, and it should not be confused with the larger dataset used to do the pre-correction.

For the "full genomic dataset", purebred sows were genotyped with the Illumina PorcineSNP60 Genotyping BeadChip [24], while the crossbred sows were genotyped with a 8.5K GGP-Porcine Low Density Illumina Bead SNP chip [25]. SNP quality controls (such as: call rate for individuals ≥ 80%; call rate for SNPs ≥ 90%; minor allele frequencies ≥ 0.01; etc.) were applied on the same dataset in a previous study [26], which provides more details. Then, for the crossbred individuals, imputation from low density to moderate density was done by using a joint reference panel of the two pure breeds [26] using the software Beagle version 3.3.2 [27] (imputation accuracies ≥ 95% in terms of correlation coefficients and ≥ 99% in terms of correct rates between imputed and true genotypes). Finally, 41,009 SNPs were available for all the recorded purebred and crossbred sows.

Considering genomic inbreeding and heterosis

Inbreeding can be defined as the proportion of homozygous SNPs across all loci for each animal, as suggested by several authors (e.g., [28]). If there is directional dominance causing inbreeding depression [29], then inbreeding should be considered in the genetic evaluation models [30]. Otherwise, using pedigree or marker data, estimates of genetic parameters are inflated [30, 31]. In Vitezica et al. [28], genomic inbreeding was fitted as a covariate and, in the current study, we prove this reasoning by using a parametric genomic model, such as a GBLUP.

Theory and evidence of directional dominance (equivalently, inbreeding depression) suggest that dominance effects of genes (here associated to markers) should have *a priori* a positive value for traits that exhibit inbreeding depression or heterosis. If we call **d** the vector of dominance marker effects, the following prior distribution is plausible:

$$(\mathbf{d}) \sim N\left(\mathbf{1}\mu_d, \mathbf{I}\sigma_d^2\right),$$

where μ_d is the overall mean of dominance effects, which should be positive if there is heterosis due to dominance. A typical model for genomic prediction is that in Toro and Varona [11]:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{Wd} + \mathbf{e}, \quad (1)$$

where **y** contains phenotypic values; **Xβ** stands for fixed effects and random effects other than additive and dominance effects; **a** is the vector of "biological" additive SNP effects, **d** is the vector of "biological" dominance SNP effects for each of the markers; matrix **Z** has entries 1, 0, -1, for SNP genotypes AA, Aa and aa, respectively, while matrix **W** has entries 0, 1, 0 for SNP genotypes AA, Aa and aa, respectively. **e** is the vector of overall random residual effects.

Typically, genetic models require **a** and **d** to have zero means, which is not true for **d** when directional dominance exist. Defining $\mathbf{d}^* = \mathbf{d} - \mathrm{E}(\mathbf{d})$, then $\mathrm{E}(\mathbf{d}^*) = \mathbf{0}$, and Equation (1) can be written as:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{W}\left(\mathbf{d}^* + \mathbf{E}(\mathbf{d})\right) + \mathbf{e}$$

$$= \mathbf{X\beta} + \mathbf{Za} + \mathbf{Wd}^* + \mathbf{W1}\boldsymbol{\mu}_\mathbf{d} + \mathbf{e}.$$

The term $\mathbf{W1}\mu_d$ is actually an average of dominance effects for each individual and is equal to $\mathbf{h}\mu_d$, where $\mathbf{h} = \mathbf{W1}$ contains the row-sums of **W**, i.e. individual heterozygosities (it should be noted that **W** has a value of 1 at heterozygous loci for an individual). Inbreeding coefficients **f** can be calculated as:

$$\mathbf{f} = \mathbf{1} - \mathbf{h}/N,$$

where N is the number of SNPs. Then, the prior means $\mathbf{h}\mu_d$ can be rewritten as:

$$\mathbf{h}\mu_d = (\mathbf{1} - \mathbf{f})N\mu_d = \mathbf{1}N\mu_d + \mathbf{f}(-N\mu_d).$$

The term $\mathbf{1}N\mu_d$ is confounded with the overall mean of the model ($\boldsymbol{\mu}$), while the term $\mathbf{f}(-N\mu_d)$ models the inbreeding depression and $b = (-N\mu_d)$ is the inbreeding depression parameter summed over the SNPs, which has to be estimated. Thus, the linear model including genomic inbreeding is, finally:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{f}b + \mathbf{Za} + \mathbf{Wd}^* + \mathbf{e}.$$

Thus, we have proven why fitting overall homozygosity for the individual as a measure of inbreeding depression accounts for directional dominance.

Estimating genetic (co)variances of markers with additive and dominance effects

A trivariate model based on "biological" (genotypic) additive and dominance effects of SNPs [22, 32], and including genomic inbreeding as above, was applied considering TNB as a different trait in each population:

$$\mathbf{y}_L = \mathbf{1}\mu_L + \mathbf{f}_L b_L + \mathbf{Z}_L\mathbf{a}_L + \mathbf{W}_L\mathbf{d}_L + \mathbf{e}_L, \quad (2)$$

$$\mathbf{y}_Y = \mathbf{1}\mu_Y + \mathbf{f}_Y b_Y + \mathbf{Z}_Y\mathbf{a}_Y + \mathbf{W}_Y\mathbf{d}_Y + \mathbf{e}_Y,$$

$$\mathbf{y}_{LY} = \mathbf{1}\mu_{LY} + \mathbf{f}_{LY} b_{LY} + \mathbf{Z}_{LY}\mathbf{a}_{LY} + \mathbf{W}_{LY}\mathbf{d}_{LY} + \mathbf{e}_{LY},$$

where $\mathbf{y}_L$, $\mathbf{y}_Y$ and $\mathbf{y}_{LY}$ contain corrected phenotypic values for purebred L, purebred Y and crossbred LY sows,

respectively; μ_L, μ_Y and μ_{LY} are the respective means; $\mathbf{a}_L$, $\mathbf{a}_Y$ and $\mathbf{a}_{LY}$ are the "biological" additive SNP effects and $\mathbf{d}_L$, $\mathbf{d}_Y$ and $\mathbf{d}_{LY}$ are the "biological" dominance SNP effects for each of the SNPs for L, Y and LY, respectively; matrices $\mathbf{Z}$ and $\mathbf{W}$ are as above; $\mathbf{f}_L b_L$, $\mathbf{f}_Y b_Y$ and $\mathbf{f}_{LY} b_{LY}$ model the inbreeding depression for L, Y and LY populations; $\mathbf{e}_L$, $\mathbf{e}_Y$ and $\mathbf{e}_{LY}$ are the overall random residual effects.

Note that "biological" is used here to refer to the genotypic additive and dominance values of the SNPs, to distinguish them from the traditional treatment of quantitative genetics in terms of "statistical" effects (breeding values and dominance deviations) [32].

The above equations can be reformulated to genotypic values of individuals instead of SNPs, in order to be compatible with the classical GBLUP model and animal breeding software, such as BLUPF90 [33] and DMU [34]:

$$\mathbf{y}_L = \mathbf{1}\mu_L + \mathbf{f}_L b_L + \mathbf{u}_L + \mathbf{v}_L + \mathbf{e}_L, \quad (3)$$

$$\mathbf{y}_Y = \mathbf{1}\mu_Y + \mathbf{f}_Y b_Y + \mathbf{u}_Y + \mathbf{v}_Y + \mathbf{e}_Y,$$

$$\mathbf{y}_{LY} = \mathbf{1}\mu_{LY} + \mathbf{f}_{LY} b_{LY} + \mathbf{u}_{LY} + \mathbf{v}_{LY} + \boldsymbol{e}_{LY}.$$

Note that $\mathbf{u}$ and $\mathbf{v}$ are vectors of genotypic additive and dominance effects and therefore cannot be directly compared to breeding values and dominance deviations in the pedigree-based genetic evaluation. In addition, $\mathbf{f}$ is a vector of genomic inbreeding coefficients and b is a population-specific inbreeding depression parameter per unit of genomic inbreeding, respectively. Note that there is potentially inbreeding depression at the level of the crossbred animals, although, first, the numeric values of the vector $\mathbf{f}$ should be smaller since crossbred animals have a higher level of heterozygosity, and second, the estimates of the inbreeding depression parameters (b) do not need to be identical across the three populations, which thus gives considerable flexibility.

In terms of the genotypic additive effects $\mathbf{u}$, the variances within each breed are:

$$Var(\mathbf{u}_L) = var(\mathbf{Z}_L\mathbf{a}_L) = \mathbf{Z}_L\mathbf{Z}_L'\sigma^2_{a_L},$$

$$Var(\mathbf{u}_Y) = var(\mathbf{Z}_Y\mathbf{a}_Y) = \mathbf{Z}_Y\mathbf{Z}_Y'\sigma^2_{a_Y},$$

$$Var(\mathbf{u}_{LY}) = var(\mathbf{Z}_{LY}\mathbf{a}_{LY}) = \mathbf{Z}_{LY}\mathbf{Z}_{LY}'\sigma^2_{a_{LY}},$$

where $\sigma^2_{a_L}$, $\sigma^2_{a_Y}$ and $\sigma^2_{a_{LY}}$ are the additive variances of SNP effects in breeds L, Y and LY, respectively. The

covariances between the genotypic additive effects **u** are:

$$Cov\begin{pmatrix}\mathbf{u}_L\\ \mathbf{u}_Y\\ \mathbf{u}_{LY}\end{pmatrix}=\begin{pmatrix}\mathbf{Z}_L\mathbf{Z}_L'\sigma_{a_L}^2 & \mathbf{Z}_L\mathbf{Z}_Y'\sigma_{a_{L,Y}} & \mathbf{Z}_L\mathbf{Z}_{LY}'\sigma_{a_{L,LY}}\\ \mathbf{Z}_Y\mathbf{Z}_L'\sigma_{a_{L,Y}} & \mathbf{Z}_Y\mathbf{Z}_Y'\sigma_{a_Y}^2 & \mathbf{Z}_Y\mathbf{Z}_{LY}'\sigma_{a_{Y,LY}}\\ \mathbf{Z}_{LY}\mathbf{Z}_L'\sigma_{a_{L,LY}} & \mathbf{Z}_{LY}\mathbf{Z}_Y'\sigma_{a_{Y,LY}} & \mathbf{Z}_{LY}\mathbf{Z}_{LY}'\sigma_{a_{LY}}^2\end{pmatrix}, \quad (4)$$

where $\sigma_{a_{L,Y}}$, $\sigma_{a_{L,LY}}$ and $\sigma_{a_{Y,LY}}$ are the additive covariances of SNP effects between populations L and Y, populations L and LY, and populations Y and LY, respectively. Analogous structures exist for dominance genotypic effects:

$$Cov\begin{pmatrix}\mathbf{v}_L\\ \mathbf{v}_Y\\ \mathbf{v}_{LY}\end{pmatrix}=\begin{pmatrix}\mathbf{W}_L\mathbf{W}_L'\sigma_{d_L}^2 & \mathbf{W}_L\mathbf{W}_Y'\sigma_{d_{L,Y}} & \mathbf{W}_L\mathbf{W}_{LY}'\sigma_{d_{L,LY}}\\ \mathbf{W}_Y\mathbf{W}_L'\sigma_{d_{L,Y}} & \mathbf{W}_Y\mathbf{W}_Y'\sigma_{d_Y}^2 & \mathbf{W}_Y\mathbf{W}_{LY}'\sigma_{d_{Y,LY}}\\ \mathbf{W}_{LY}\mathbf{W}_L'\sigma_{d_{L,LY}} & \mathbf{W}_{LY}\mathbf{W}_Y'\sigma_{d_{Y,LY}} & \mathbf{W}_{LY}\mathbf{W}_{LY}'\sigma_{d_{LY}}^2\end{pmatrix}.$$

Estimation of marker-based variance components using an equivalent model

The variance components $\sigma_{a_L}^2$, $\sigma_{a_Y}^2$, $\sigma_{a_{LY}}^2$ and $\sigma_{a_{L,Y}}$, $\sigma_{a_{L,LY}}$, $\sigma_{a_{Y,LY}}$ in Equation (4) cannot be estimated by regular methods or software (i.e. REML or Gibbs sampling) because they cannot be factorized out from Equation (4). To fit such a multivariate structure, we used an equivalent model. Additional effects need to be defined, even if they are of no interest *per se*. For instance, the vectors of hypothetical genotypic additive effects of the genotypes of the L breed on the scale of breed Y ($\mathbf{u}_{L,Y}$) and LY ($\mathbf{u}_{L,LY}$) have variance-covariance matrices $\mathbf{Z}_L\mathbf{Z}_L'\sigma_{a_Y}^2$ and $\mathbf{Z}_L\mathbf{Z}_L'\sigma_{a_{LY}}^2$, respectively. Thus, as a whole, the genetic variance and covariance structure for the genotypic additive effects **u** are:

$$Var(\mathbf{u})=\text{var}\begin{bmatrix}\mathbf{u}_L\\ \mathbf{u}_{L,Y}\\ \mathbf{u}_{L,LY}\\ \mathbf{u}_{Y,L}\\ \mathbf{u}_Y\\ \mathbf{u}_{Y,LY}\\ \mathbf{u}_{LY,L}\\ \mathbf{u}_{LY,Y}\\ \mathbf{u}_{LY}\end{bmatrix}=\text{var}\begin{bmatrix}\mathbf{Z}_L\mathbf{a}_L\\ \mathbf{Z}_L\mathbf{a}_Y\\ \mathbf{Z}_L\mathbf{a}_{LY}\\ \mathbf{Z}_Y\mathbf{a}_L\\ \mathbf{Z}_Y\mathbf{a}_Y\\ \mathbf{Z}_Y\mathbf{a}_{LY}\\ \mathbf{Z}_{LY}\mathbf{a}_L\\ \mathbf{Z}_{LY}\mathbf{a}_Y\\ \mathbf{Z}_{LY}\mathbf{a}_{LY}\end{bmatrix}$$

$$=\begin{bmatrix}\mathbf{Z}_L\mathbf{Z}_L' & \mathbf{Z}_L\mathbf{Z}_Y' & \mathbf{Z}_L\mathbf{Z}_{LY}'\\ \mathbf{Z}_Y\mathbf{Z}_L' & \mathbf{Z}_Y\mathbf{Z}_Y' & \mathbf{Z}_Y\mathbf{Z}_{LY}'\\ \mathbf{Z}_{LY}\mathbf{Z}_L' & \mathbf{Z}_{LY}\mathbf{Z}_Y' & \mathbf{Z}_{LY}\mathbf{Z}_{LY}'\end{bmatrix}\otimes\begin{bmatrix}\sigma_{a_L}^2 & \sigma_{a_{L,Y}} & \sigma_{a_{L,LY}}\\ \sigma_{a_{Y,L}} & \sigma_{a_Y}^2 & \sigma_{a_{Y,LY}}\\ \sigma_{a_{LY,L}} & \sigma_{a_{LY,Y}} & \sigma_{a_{LY}}^2\end{bmatrix}$$

$$= \mathbf{ZZ'} \otimes \begin{bmatrix} \sigma^2_{a_L} & \sigma_{a_{L,Y}} & \sigma_{a_{L,LY}} \\ \sigma_{a_{Y,L}} & \sigma^2_{a_Y} & \sigma_{a_{Y,LY}} \\ \sigma_{a_{LY,L}} & \sigma_{a_{LY,Y}} & \sigma^2_{a_{LY}} \end{bmatrix},$$

where matrix $\mathbf{Z}$ contains elements 1, 0, -1 for the three genotypes, and is defined across the three breeds, $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_L \\ \mathbf{Z}_Y \\ \mathbf{Z}_{LY} \end{bmatrix}$.

To construct a relationship matrix similar to the classical $\mathbf{G}$-matrix of GBLUP [35], Vitezica et al. [22] introduced a normalized genomic relationship matrix $\mathbf{G} = \frac{\mathbf{ZZ'}}{\{\mathrm{tr}[\mathbf{ZZ'}]\}/\mathrm{n}}$, where n is the number of animals across the three populations and the division by $\{\mathrm{tr}[\mathbf{ZZ'}]\}/\mathrm{n}$ scales the matrix such that the average of the diagonal elements equals 1. This alters the variances across genotypic additive effects $\mathbf{u}$ in the following way:

$$Var(\mathbf{u}) = var\begin{bmatrix} \mathbf{u}_L \\ \mathbf{u}_{L,Y} \\ \mathbf{u}_{L,LY} \\ \mathbf{u}_{Y,L} \\ \mathbf{u}_Y \\ \mathbf{u}_{Y,LY} \\ \mathbf{u}_{LY,L} \\ \mathbf{u}_{LY,Y} \\ \mathbf{u}_{LY} \end{bmatrix} = \mathbf{ZZ}' \otimes \begin{bmatrix} \sigma^2_{a_L} & \sigma_{a_{L,Y}} & \sigma_{a_{L,LY}} \\ \sigma_{a_{Y,L}} & \sigma^2_{a_Y} & \sigma_{a_{Y,LY}} \\ \sigma_{a_{LY,L}} & \sigma_{a_{LY,Y}} & \sigma^2_{a_{LY}} \end{bmatrix}$$

$$= \left(\mathbf{G} \times \frac{\{\mathrm{tr}[\mathbf{ZZ'}]\}}{\mathrm{n}}\right) \otimes \begin{bmatrix} \sigma^2_{a_L} & \sigma_{a_{L,Y}} & \sigma_{a_{L,LY}} \\ \sigma_{a_{Y,L}} & \sigma^2_{a_Y} & \sigma_{a_{Y,LY}} \\ \sigma_{a_{LY,L}} & \sigma_{a_{LY,Y}} & \sigma^2_{a_{LY}} \end{bmatrix}$$

$$= \mathbf{G} \otimes \begin{bmatrix} \sigma^2_{A_L} & \sigma_{A_L A_Y} & \sigma_{A_L A_{LY}} \\ \sigma_{A_Y A_L} & \sigma^2_{A_Y} & \sigma_{A_Y A_{LY}} \\ \sigma_{A_L A_{LY}} & \sigma_{A_Y A_{LY}} & \sigma^2_{A_{LY}} \end{bmatrix} = \mathbf{G} \otimes \mathbf{G}_0, \quad (5)$$

where $\mathbf{G}_0$ are variance components associated to the genotypic additive effects $\mathbf{u}$. This structure (a Kronecker product) is compatible with animal breeding software for BLUP and REML and the variance-covariance component $\mathbf{G}_0$ can be estimated in a straightforward manner. Then, the (co)variances of additive genotypic effects of SNPs across populations can be obtained as:

$$\begin{bmatrix} \sigma^2_{a_L} & \sigma_{a_{L,Y}} & \sigma_{a_{L,LY}} \\ \sigma_{a_{Y,L}} & \sigma^2_{a_Y} & \sigma_{a_{Y,LY}} \\ \sigma_{a_{LY,L}} & \sigma_{a_{LY,Y}} & \sigma^2_{a_{LY}} \end{bmatrix} = \mathbf{G}_0 \Big/ \{\mathrm{tr}[\mathbf{ZZ'}]\}/\mathrm{n}$$

$$= \begin{bmatrix} \sigma^2_{A_L} & \sigma_{A_L A_Y} & \sigma_{A_L A_{LY}} \\ \sigma_{A_Y A_L} & \sigma^2_{A_Y} & \sigma_{A_Y A_{LY}} \\ \sigma_{A_L A_{LY}} & \sigma_{A_Y A_{LY}} & \sigma^2_{A_{LY}} \end{bmatrix} \Big/ \{\text{tr}[\mathbf{ZZ}']\}/\text{n}. \quad (6)$$

The variances across genotypic dominance effects **v** are altered in a similar way:

$$Var(\mathbf{v}) = var \begin{bmatrix} \mathbf{v}_L \\ \mathbf{v}_{L,Y} \\ \mathbf{v}_{L,LY} \\ \mathbf{v}_{Y,L} \\ \mathbf{v}_Y \\ \mathbf{v}_{Y,LY} \\ \mathbf{v}_{LY,L} \\ \mathbf{v}_{LY,Y} \\ \mathbf{v}_{LY} \end{bmatrix}$$

$$= \mathbf{D} \otimes \begin{bmatrix} \sigma^2_{D_L} & \sigma_{D_L D_Y} & \sigma_{D_L D_{LY}} \\ \sigma_{D_Y D_L} & \sigma^2_{D_Y} & \sigma_{D_Y D_{LY}} \\ \sigma_{D_L D_{LY}} & \sigma_{D_Y D_{LY}} & \sigma^2_{D_{LY}} \end{bmatrix} = \mathbf{D} \otimes \mathbf{D}_0, \quad (7)$$

where $\mathbf{D}_0$ contains variances and covariances associated to the genotypic dominance effects **v** and $\mathbf{D} = \frac{\mathbf{WW}'}{\{\text{tr}[\mathbf{WW}']\}/\text{n}}$, where the matrix **W** contains elements 0, 1, 0 for the three genotypes, and is defined across the three breeds ($\mathbf{W} = \begin{bmatrix} \mathbf{W}_L \\ \mathbf{W}_Y \\ \mathbf{W}_{LY} \end{bmatrix}$) and $\mathbf{W}' = [\mathbf{W}'_L \quad \mathbf{W}'_Y \quad \mathbf{W}'_{LY}]$. Then, the (co)variances of dominance genotypic effects of SNPs are:

$$\begin{bmatrix} \sigma^2_{d_L} & \sigma_{d_{L,Y}} & \sigma_{d_{L,LY}} \\ \sigma_{d_{Y,L}} & \sigma^2_{d_Y} & \sigma_{d_{Y,LY}} \\ \sigma_{d_{LY,L}} & \sigma_{d_{LY,Y}} & \sigma^2_{d_{LY}} \end{bmatrix} = \mathbf{D}_0 \Big/ \{\text{tr}[\mathbf{WW}']\}/\text{n}$$

$$= \begin{bmatrix} \sigma^2_{D_L} & \sigma_{D_L D_Y} & \sigma_{D_L D_{LY}} \\ \sigma_{D_Y D_L} & \sigma^2_{D_Y} & \sigma_{D_Y D_{LY}} \\ \sigma_{D_L D_{LY}} & \sigma_{D_Y D_{LY}} & \sigma^2_{D_{LY}} \end{bmatrix} \Big/ \{\text{tr}[\mathbf{WW}']\}/\text{n} \quad (8)$$

This approach, which is an extension of Vitezica et al. [22], makes it possible to estimate (co)variances of genotypic effects of SNPs in purebred and crossbred populations under a genomic model with additive and non-additive (dominance) inheritance.

Matrices **Z** and **W**, their crossproducts and the inverses of **G** and **D** were built using own programs. Genetic

parameters were estimated by using average information REML with software airemlf90 [33]. Standard errors on functions of genetic parameters (i.e. standard errors on correlations) were estimated from the average information matrix using the REML-MVN method of Houle and Meyer [36].

Additive and dominance variances in purebred and crossbred populations

The additive and dominance (co)variances of genotypic effects of SNPs, either within breed or between breeds, were calculated using Equations (6) and (8), respectively. Using these calculated additive and dominance (co)variances of SNPs across all the SNPs, the corresponding traditional, individual-based genetic parameters can be obtained as follows. The genetic parameters obtained are directly comparable to pedigree-based estimates [32].

Consider the allele substitution effect $\alpha = a + (q - p)d$. According to [32], the additive genetic variances for purebred performance (mating animals in the same breed) for breed L ($\sigma^2_{AP_L}$) and Y ($\sigma^2_{AP_Y}$) are:

$$\sigma^2_{AP_L} = \sum\left(2p_i^L q_i^L\right)\sigma^2_{a_L} + \sum\left(2p_i^L q_i^L\left(q_i^L - p_i^L\right)^2\right)\sigma^2_{d_L},$$

$$\sigma^2_{AP_Y} = \sum\left(2p_i^Y q_i^Y\right)\sigma^2_{a_Y} + \sum\left(2p_i^Y q_i^Y\left(q_i^Y - p_i^Y\right)^2\right)\sigma^2_{d_Y},$$

where σ^2_a and σ^2_d are the variances of additive and dominance genotypic effects of SNPs in either breed L or Y; p_i and q_i are allele frequencies for SNP i; indices L and Y denote the breeds Landrace and Yorkshire, respectively. For crossbred performance of say, Landrace, the allele substitution effect is $\alpha_{AC_L} = a_{AC_L} + (q^Y - p^Y)d_{AC_L}$. Thus, the additive genetic variances within purebred L and Y for crossbred performance (due to gametes from the L or Y individuals in the crossbred population) are equal to:

$$\sigma^2_{AC_L} = \sum\left(2p_i^L q_i^L\right)\sigma^2_{a_{LY}} + \sum\left(2p_i^L q_i^L\left(q_i^Y - p_i^Y\right)^2\right)\sigma^2_{d_{LY}},$$

$$\sigma^2_{AC_Y} = \sum\left(2p_i^Y q_i^Y\right)\sigma^2_{a_{LY}} + \sum\left(2p_i^Y q_i^Y\left(q_i^L - p_i^L\right)^2\right)\sigma^2_{d_{LY}},$$

where the $\sigma^2_{AC_L}$ represents the additive genetic variance of animals in breed L when mated to animals in breed Y; the $\sigma^2_{AC_Y}$ represents the additive genetic variance of animals in breed Y when mated to animals in breed L; and $\sigma^2_{a_{LY}}$ and $\sigma^2_{d_{LY}}$ are the variances of additive and dominance genotypic effects of SNPs in the crossbred LY population, respectively. The additive genetic variance for animals in the crossbred LY population ($\sigma^2_{AC_{LY}}$) is the sum of the additive genetic variance of Landrace alleles and that of Yorkshire alleles in the crossbred animals [22] as follows:

$\sigma^2_{AC_{LY}} = \frac{1}{2}\sigma^2_{AC_L} + \frac{1}{2}\sigma^2_{AC_Y}$.

Note that this variance is not the additive genetic variance of the crossbred animals acting as reproducers (i.e., creating an F2) [37].

The additive genetic covariances between purebred and crossbred performances within breeds L (σ_{AP_L,AC_L}) and Y (σ_{AP_Y,AC_Y}) are:

$$\sigma_{AP_L,AC_L} = \sum\left(2p_i^L q_i^L\right)\sigma_{a_{L,LY}} + \sum\left(2p_i^L q_i^L\left(q_i^L - p_i^L\right)\left(q_i^Y - p_i^Y\right)\right)\sigma_{d_{L,LY}},$$

$$\sigma_{AP_Y,AC_Y} = \sum\left(2p_i^Y q_i^Y\right)\sigma_{a_{Y,LY}} + \sum\left(2p_i^Y q_i^Y\left(q_i^Y - p_i^Y\right)\left(q_i^L - p_i^L\right)\right)\sigma_{d_{Y,LY}},$$

where $\sigma_{a_{L,LY}}$ and $\sigma_{d_{L,LY}}$ are the covariances of SNP effects between purebred L and crossbred LY populations for additive and dominance, respectively; $\sigma_{a_{Y,LY}}$ and $\sigma_{d_{Y,LY}}$ are the covariances of SNP effects between purebred Y and crossbred LY populations for additive and dominance, respectively.

Therefore, the genetic correlations of breeding values between purebred and crossbred performances within L (r_{PC_L}) and Y (r_{PC_Y}) are:

$r_{PC_L} = \frac{\sigma_{AP_L,AC_L}}{\sqrt{\sigma^2_{AP_L}\sigma^2_{AC_L}}}$ and $r_{PC_Y} = \frac{\sigma_{AP_Y,AC_Y}}{\sqrt{\sigma^2_{AP_Y}\sigma^2_{AC_Y}}}$.

According to [22], the dominance genetic variances within purebred populations L and Y are $\sigma^2_{D_L} = \sum\left(2p_i^L q_i^L\right)^2\sigma^2_{d_L}$ and $\sigma^2_{D_Y} = \sum\left(2p_i^Y q_i^Y\right)^2\sigma^2_{d_Y}$, respectively. The dominance genetic variance in crossbred LY animals is $\sigma^2_{D_{LY}} = \sum(4p_i^L q_i^L p_i^Y q_i^Y)\sigma^2_d$.

The broad sense heritabilities for purebred performance (H^2_P) were calculated as the ratio of total genetic variances for purebred performance ($\sigma^2_{AP} + \sigma^2_D$) to phenotypic variances ($\sigma^2_{AP} + \sigma^2_D + \sigma^2_e$).

Correlations of allele substitution effects between two breeds

The breeding value of an individual includes the allele substitution effects of all genes and the allele frequencies. For purebred performance, the allele substitution effects of one locus for breed L and Y are:

$$\alpha_L = a_L + \left(q_i^L - p_i^L\right)d_L,$$

$$\alpha_Y = a_Y + \left(q_i^Y - p_i^Y\right)d_Y,$$

where a is the additive effect and d is the dominance effect for each SNP; p_i and q_i are allele frequencies for SNP i, with superscripts denoting breeds L or Y. In the case of purely additive gene action, the covariance between α_L and α_Y is $\sigma_{a_{L,Y}}$, which can be interpreted as a genetic correlation among populations [38-40]. Then, the covariance between the allele substitution effects of one locus is:

$$cov(\alpha_L, \alpha_Y) = cov\left(a_L + \left(q_i^L - p_i^L\right)d_L, a_Y + \left(q_i^Y - p_i^Y\right)d_Y\right)$$

$$= cov(a_L, a_Y) + \left(q_i^L - p_i^L\right)\left(q_i^Y - p_i^Y\right)cov(d_L, d_Y)$$

$$= \sigma_{a_{L,Y}} + \left(q_i^L - p_i^L\right)\left(q_i^Y - p_i^Y\right)\sigma_{d_{L,Y}},$$

where $\sigma_{a_{L,Y}}$ and $\sigma_{d_{L,Y}}$ are the additive and dominance covariances of SNP effects between breeds L and Y for additive and dominance, respectively. If we assume that SNP effects (both additive and dominance) are independent across loci, then the covariance between the allele substitution effects across all n loci is:

$$cov(\alpha_L, \alpha_Y) = \sigma_{\alpha_{L,Y}} = \sigma_{a_{L,Y}} + \frac{1}{n}\sum\left(\left(q_i^L - p_i^L\right)\left(q_i^Y - p_i^Y\right)\right)\sigma_{d_{L,Y}}.$$

Also, the variances of allele substitution effects across all n loci for breeds L and Y are:

$$var(\alpha_L) = \sigma_{\alpha_L}^2 = \sigma_{a_L}^2 + \frac{1}{n}\sum\left(\left(q_i^L - p_i^L\right)^2\right)\sigma_{d_L}^2,$$

$$var(\alpha_Y) = \sigma_{\alpha_Y}^2 = \sigma_{a_Y}^2 + \frac{1}{n}\sum\left(\left(q_i^Y - p_i^Y\right)^2\right)\sigma_{d_Y}^2,$$

where σ_a^2 and σ_d^2 are the additive and dominance variance of SNPs. Then, the correlation of allele substitution effects for purebred performance between populations L and Y is $r_{\alpha P_L, \alpha P_Y} = \frac{\sigma_{\alpha_{L,Y}}}{\sigma_{\alpha_L}\sigma_{\alpha_Y}}$. If there is no dominance variation, the $r_{\alpha P_L, \alpha P_Y}$ relates to additive genetic variances as $r_{\alpha P_L, \alpha P_Y} = \frac{\sigma_{a_{L,Y}}}{\sigma_{a_L}\sigma_{a_Y}}$.

The correlation of allele substitution effects for crossbred performance between populations L and Y is similar to that for purebred performance, but the allele frequencies are swapped, as:

$$r_{\alpha C_L, \alpha C_Y} = \frac{\sigma_{\alpha_{L\ in\ LY, Y\ in\ LY}}}{\sigma_{\alpha_{L\ in\ LY}}\sigma_{\alpha_{Y\ in\ LY}}} = \frac{\sigma_{a_{LY}}^2 + \frac{1}{n}\sum\left(\left(q_i^L - p_i^L\right)\left(q_i^Y - p_i^Y\right)\right)\sigma_{d_{LY}}^2}{\sqrt{\sigma_{a_{LY}}^2 + \frac{1}{n}\sum\left(\left(q_i^Y - p_i^Y\right)^2\right)\sigma_{d_{LY}}^2}\sqrt{\sigma_{a_{LY}}^2 + \frac{1}{n}\sum\left(\left(q_i^L - p_i^L\right)^2\right)\sigma_{d_{LY}}^2}},$$

where $\sigma^2_{a_{LY}}$ and $\sigma^2_{d_{LY}}$ are the additive and dominance variance of SNPs in the crossbred LY population. If there is no dominance variation, the $r_{\alpha C_L,\alpha C_Y}$ is equal to 1, by assumption in the model.

Scenarios

Variance components, genetic correlations of breeding values between purebred and crossbred performances (r_{PC}) within each pure breed and correlations of allele substitution effects for purebred ($r_{\alpha P_L,\alpha P_Y}$) and crossbred ($r_{\alpha C_L,\alpha C_Y}$) performance between two pure breeds were first investigated using the full genomic dataset. To explore the effects of using genomic information and the inclusion of dominance deviation on the genetic evaluation of crossbred performance in the trivariate model, three different scenarios were compared.

Nogen:

The statistical model was a trivariate BLUP model, similar to Equation (3), but the dominance deviation was excluded. Instead of using a genomic relationship matrix, a single relationship matrix **A** was constructed across the three breeds, assuming that they form a single population. Thus, the genetic (co)variances of additive genetic effects **u** were:

$$Var(\mathbf{u}) = \mathbf{A} \otimes \begin{bmatrix} \sigma^2_{A_L} & \sigma_{A_L A_Y} & \sigma_{A_L A_{LY}} \\ \sigma_{A_Y A_L} & \sigma^2_{A_Y} & \sigma_{A_Y A_{LY}} \\ \sigma_{A_L A_{LY}} & \sigma_{A_Y A_{LY}} & \sigma^2_{A_{LY}} \end{bmatrix} = \mathbf{A} \otimes \mathbf{A_0},$$

where $\mathbf{A_0}$ were variance components associated to genetic additive effects and not the genotypic additive effects in Equation (5). Pedigree-based inbreeding depression was also included in the model. The pedigree-based inbreeding coefficients were calculated as in [41] using the software inbupgf90 [33].

Gen_AM:

The statistical model was similar to Equation (3), but without dominance deviations. Genomic information was used to construct the additive genomic relationship matrix.

Gen_ADM:

The statistical model includes additive and dominance effects as in Equation (3). Genomic information was used to construct the additive and dominance genomic relationship matrices.

To explore the impact of genomic information and dominance effects on genomic evaluation for crossbred performance, the full genomic dataset was split into training and validation populations and the predictive ability for crossbred animals in the validation population was investigated in different scenarios. The farrowing date of January 1, 2013 was used as the cut-off date to divide recorded purebred and crossbred sows into training and validation populations. As a result, 6769 sows (1270 L, 1405 Y and 4094 LY) were included in the training population, while the remaining 2716 sows (854 L, 813Y and 1049 LY) were included in the validation population. Predictive ability of crossbreds was measured as the correlations $cor(y_c, \hat{y})$ in the validation population for each scenario, where y_c is the corrected phenotypic records of TNB for crossbred animals; $\hat{y}$ is the predicted corrected observations of TNB for crossbred animals and is equal to the sum of the estimated population mean ($\hat{\mu}$), inbreeding ($f\hat{b}$) and genotypic values ($\hat{g}$); the genotypic value $\hat{g}$ was calculated as the sum of additive and dominance genetic effects in the scenario *Gen_ADM*. In the other two scenarios, the genotypic value $\hat{g}$ only included the additive genetic effect. Hotelling-Williams t-test at a confidence level of 5% was applied to evaluate the significance of the differences in validation correlations in each scenario. Furthermore, to detect the possible biases in the predictions, the regression coefficients of y_c on $\hat{y}$ were explored. Note that no bias implies that a regression coefficient equals 1. In addition, to measure the uncertainty associated with the predictions, 1000 bootstrap samples [42] was applied to estimate the means and standard errors.

For comparison, the predictive ability of crossbred animals was also investigated in a model without inbreeding depression effects, for all three scenarios. The predictive ability was measured as the correlation $cor(y_c, \hat{y})$, where $\hat{y}$ is the sum of the estimated population mean ($\hat{\mu}$) and genotypic value ($\hat{g}$).

Results

Variance components, heritabilities and correlations

Table 1 shows the estimates of variance components for additive genetic effects for purebred (σ^2_{AP}) and crossbred (σ^2_{AC}) performance in different scenarios, and dominance variations (σ^2_D) in the *Gen_ADM* scenario. For all scenarios, the additive genetic variances for purebred performance (σ^2_{AP}) were larger than those for their crossbred performance (σ^2_{AC}). Estimated variance components in the scenarios *Gen_AM* and *Gen_ADM* were very close, but different from those obtained in scenarios without using genomic information. In general, estimates had large standard errors in all scenarios, but no obvious differences in standard errors were detected between different scenarios. Residual variance for purebred animals (σ^2_e) was larger than for crossbred animals ($\sigma^2_{e_{LY}}$) in each scenario. For the scenario *Gen_ADM*, the ratios of dominance genetic variance to additive genetic variance ranged from 5 to 11% for both purebred and crossbred populations.

Table 1 Variance components of additive and dominance genetic effects for purebred and crossbred animals

Scenario	Breed	σ^2_{AP}	$\sigma_{AP,AC}$	σ^2_{AC}	σ^2_D	σ^2_e	$\sigma^2_{AC_{LY}}$	$\sigma^2_{D_{LY}}$	$\sigma^2_{e_{LY}}$
Nogen	L	0.99	0.17	0.05	-	10.82	0.05	-	7.35
		(0.31)	(0.07)	(0.02)		(0.43)	(0.02)		(0.15)
	Y	1.07	0.15	0.05	-	8.96			
		(0.33)	(0.07)	(0.02)		(0.38)			
Gen_AM	L	0.87	0.47	0.28	-	10.89	0.28	-	7.11
		(0.22)	(0.10)	(0.07)		(0.38)	(0.07)		(0.15)
	Y	0.55	0.17	0.28	-	9.42			
		(0.20)	(0.10)	(0.07)		(0.33)			
Gen_ADM	L	0.86	0.46	0.28	0.04	10.86	0.28	0.02	7.11
		(0.21)	(0.10)	(0.06)	(0.03)	(0.38)	(0.06)	(0.01)	(0.15)
	Y	0.54	0.17	0.28	0.06	9.35			
		(0.18)	(0.09)	(0.06)	(0.05)	(0.33)			

σ^2_{AP} is the additive genetic variance for purebred performance; $\sigma_{AP,AC}$ is the additive genetic covariance between purebred and crossbred performance; σ^2_{AC} is the additive genetic variance for crossbred performance; σ^2_D is the dominance genetic variance for either purebred animals; σ^2_e is the residual variance for purebred animals; $\sigma^2_{AC_{LY}}$is the additive genetic variance for the F1 crossbred animals LY; $\sigma^2_{D_{LY}}$ is the dominance genetic variance for the F1 crossbred animals LY; and $\sigma^2_{e_{LY}}$is the residual variance for the F1 crossbred animals LY.
L: Landrace and Y: Yorkshire breeds.
Numbers in brackets are the standard errors of the corresponding parameters.

The broad sense heritabilities for purebred and crossbred animals, genetic correlations between breeding values for purebred and crossbred performances within pure breeds and correlations of allele substitution effects across the two breeds are in Table 2. In different scenarios, the heritabilities of purebred performance (H^2_P) ranged from 0.07 (0.03) to 0.08 (0.03) and from 0.06 (0.03) to 0.10 (0.03) for breeds L and Y, respectively. Standard errors of H^2_P were almost consistent across scenarios. Estimated genetic correlations of breeding values between purebred and crossbred performances (r_{PC}) increased from 0.76 (0.20) (*Nogen*) to 0.95 (0.06) (*Gen_AM*) for breed L and from 0.43 (0.22) (*Gen_ADM*) to 0.54 (0.30) (*Nogen*) for breed Y. The r_{PC} was higher for breed L than for breed Y in all scenarios, but the standard errors of r_{PC} were always higher for breed Y than for breed L. With genomic information, the correlations of allele substitution effects between purebred ($r_{\alpha P_L,\alpha P_Y}$) and crossbred ($r_{\alpha C_L,\alpha C_Y}$) performance between breeds L and Y were estimated, as shown in Table 3. For purebred performance, $r_{\alpha P_L,\alpha P_Y}$ was equal to 0.14 and 0.19 in *Gen_AM* and *Gen_AM*D, respectively. However, the standard errors were large, around 0.2 in both scenarios. For crossbred performance, $r_{\alpha C_L,\alpha C_Y}$ was equal to 0.98 in *Gen_ADM*. This high correlation is a byproduct of assuming that additive biological effects in crossbred animals are the same regardless of the Yorkshire or Landrace origin of the allele. However, the same allele has potentially different effects in the respective Landrace or Yorkshire genetic backgrounds, and the difference is modeled through the correlations, hence the low values of $r_{\alpha P_L,\alpha P_Y}$.Without including the dominance effects in the model *Gen_AM*, $r_{\alpha C_L,\alpha C_Y}$ was equal to 1 by definition.

Table 2 Heritabilities and genetic correlations between breeding values for purebred and crossbred performances

Scenario	Breed	r_{PC}	H_P^2
Nogen	L	0.76 (0.20)	0.08 (0.03)
	Y	0.54 (0.30)	0.10 (0.03)
Gen_AM	L	0.95 (0.06)	0.07 (0.03)
	Y	0.44 (0.20)	0.06 (0.03)
Gen_ADM	L	0.93 (0.05)	0.08 (0.03)
	Y	0.43 (0.22)	0.06 (0.03)

L: Landrace Y: Yorkshire.
r_{PC} is the genetic correlation of breeding values between purebred and crossbred performances within the Landrace or Yorkshire breeds.
H_P^2 is the broad sense heritability for purebred performance for the Landrace and Yorkshire breeds in different scenarios.
Numbers between brackets are the standard errors of the corresponding parameters.

Table 3 Correlations of allele substitution effects for purebred and crossbred performance between Landrace and Yorkshire breeds

Scenario	$r_{\alpha P_L,\alpha P_Y}$	$r_{\alpha C_L,\alpha C_Y}$
Nogen	-	-
Gen_AM	0.14 (0.22)	1
Gen_ADM	0.19 (0.24)	0.98 (0.02)

$r_{\alpha P_L,\alpha P_Y}$ is the correlation of allele substitution effects for purebred performance between the Landrace and Yorkshire breeds.
$r_{\alpha C_L,\alpha C_Y}$ is the correlation of allele substitution effects for crossbred performance between the Landrace and Yorkshire breeds. For *Gen_AM*, $r_{\alpha C_L,\alpha C_Y}$ is equal to 1 by definition.
Numbers between brackets are the standard errors of the corresponding parameters.

Predictive abilities

Predictive abilities for crossbred pigs in the validation population are in Table 4. The correlation between the corrected phenotypic values and the predicted observations for TNB ($cor(y_c, \hat{y})$) ranged from 0.010 in the scenario *Nogen* to 0.056 in scenarios *Gen_AM* and *Gen_ADM*. Standard errors of $cor(y_c, \hat{y})$ based on 1000 bootstrap samples were equal to 0.03 across all scenarios. No significant differences in predictive ability between scenarios were detected by the Hotelling-Williams t-test at the confidence level of 5%.

The regression coefficients of corrected phenotypic values on the predicted corrected observations for TNB are in the second row of Table 4. Regression coefficients were smaller than 1 for the three scenarios. Among these scenarios, regression coefficients for scenarios with genomic information (*Gen_AM* and *Gen_ADM*) were slightly closer to 1 than that for the pedigree-based scenario (*Nogen*). Except for the *Nogen* scenario, standard errors of regression coefficients were around 0.39. For the *Nogen* scenario, the standard error was around 5 times larger than that for other scenarios. Overall, there was no clear trend towards a scenario with

less bias.

For comparison, predictive abilities $cor(y_c, \hat{y})$ for crossbred pigs in the validation population for the models without the inbreeding depression effect were equal to -0.08 in scenario *Nogen*, 0.045 in scenario *Gen_AM* and 0.046 in scenario *Gen_ADM*. In all cases, these are lower than the predictive abilities in Table 4, and these differences are statistically significant according to the Hotelling-Williams t-test.

Table 4 Predictive ability for crossbred animals in the validation population

	Nogen	*Gen_AM*	*Gen_ADM*
$cor(y_c, \hat{y})$[1]	0.010 (0.031)	0.056 (0.031)	0.056 (0.031)
Regression coefficient[2]	0.703 (2.218)	0.736 (0.386)	0.730 (0.385)

[1]Predictive ability ($\mathrm{cor}(y_c, \hat{y})$) is given by the correlation coefficient between the corrected phenotypes (y_c) and their predictions ($\hat{y}$) for total number of piglets born (TNB) in crossbred animals.
[2]Regression coefficient of the corrected phenotypes (y_c) on the predicted observations ($\hat{y}$) in crossbred animals
Numbers between brackets are the standard errors of the corresponding parameters.

Inbreeding depression

Marker-based and pedigree-based inbreeding coefficient (f) for each population and their estimated corresponding inbreeding depression parameters (*b*) in the different scenarios are in Table 5. Marker-based inbreeding coefficients were almost identical for breeds L and Y, but they were larger than those for LY, which was expected because crossbred animals have a higher level of heterozygozity than purebred animals. However, according to the pedigree-based inbreeding coefficients, the Landrace population was slightly more inbred than the Yorkshire population. In terms of inbreeding depression parameters (*b*), they were all negative (thus, genomic inbreeding has detrimental effects for TNB even in crossbred animals) but not of the same magnitude across the three populations. Note that for the scenario *Nogen*, *b* was estimated based on the pedigree-based inbreeding coefficients. As a whole, breed L had the most negative *b*, while breed Y had the least negative *b*, regardless of the scenario. Thus, TNB was more negatively affected by inbreeding in breed L than in breed Y and population LY.

Table 5 Marker-based and pedigree-based inbreeding coefficients f and estimated inbreeding depression parameter *b* (piglets per 100% of inbreeding) in different scenarios for each breed

	L	Y	LY
Marker-based inbreeding coefficient f [1]	0.695 (0.019)	0.698 (0.020)	0.565 (0.012)
Pedigree-based inbreeding coefficient f [2]	0.111 (0.032)	0.078 (0.031)	0
Nogen (b)	-4.821	-3.561	0
Gen_AM (b)	-9.656	-1.924	-5.122
Gen_ADM (b)	-9.731	-1.878	-5.055

[1] is calculated as the proportion of homozygous loci per individual.
[2] is calculated as in Meuwissen and Luo [41].
In this table, the inbreeding coefficient is the mean inbreeding coefficient across individuals within each breed. Numbers between brackets are the standard deviations of the mean inbreeding coefficient.
For *Nogen*, the inbreeding depression parameter *b* is the regression of phenotype on pedigree-based inbreeding.
For *Gen_AM* and *Gen_ADM*, the inbreeding depression parameter *b* is the regression of phenotype on marker-based inbreeding.

Discussion

This study extended the trivariate GBLUP model of Vitezica et al. [22] in order to obtain (co)variances of effects of SNPs, genetic correlations of breeding values between purebred and crossbred performances and correlations of allele substitution effects under dominance. We also evaluated this model using different scenarios for the genetic evaluation of crossbred performance in Danish purebred and crossbred pigs. Scenarios that included or not genomic information were studied to estimate the genetic correlations of breeding values between purebred and crossbred performances. To our knowledge, this is the first study to report correlations of allele substitution effects between two breeds in the presence of dominance effects. The results show that the Vitezica model [22] is a tool that can be used for the genomic evaluation of crossbred performance in genotyped animals. In this study, for TNB, models with dominance deviations did not improve the genomic evaluation of crossbred performance with regard to both predictive ability and unbiasedness, but the inclusion of an inbreeding depression effect in the models significantly improved predictive ability.

Phenotypic variances were larger for purebred animals (11.76 for breed L and 9.99 for breed Y) than for crossbred animals (7.30 for LY). This could be the reason why the estimated additive genetic variances for purebred performance (σ^2_{AP}) were larger than those for crossbred performance (σ^2_{AC}). However, compared to results in a previous study that used a much larger Danish purebred and crossbred dataset [17], both estimated additive genetic variances and phenotypic variances in the current study were smaller, which is due to three reasons. (1) The dataset in the current study was a genotyped subset of the population used in the previous study. Purebred genotyped individuals were pre-selected and their performances were more

homogeneous than that of the whole population. The preselection process resulted in a loss of about 15% of the purebred phenotypic variation. However, the genotyped crossbred animals were an almost random sample of the whole population and there was only a small loss of about 5% of phenotypic variation for crossbred animals. (2) The phenotypic values for TNB in the current study were pre-corrected for fixed and non-genetic random effects. This pre-correction led to a loss of about 11 and 17% of phenotypic variation for purebreds and crossbreds, respectively. (3) During the pre-correction, some genetic variation may have been allocated to other random effects (e.g. service boar effects), in particular because TNB is a lowly heritable trait.

The estimated heritabilities of TNB for purebred performance (H_P^2) were slightly lower than those previously reported (0.11 and 0.09 for breeds L and Y, respectively) [17, 22, 43]. Large standard errors of H_P^2 implied that the current dataset was not large enough. The consistent standard errors across scenarios indicated that even when genomic information was included, the uncertainty of H_P^2 did not decrease. Taking the standard errors into account, the estimated H_P^2 across scenarios were not very different. Compared to the results of [17], the lower H_P^2 found in the current study was due to the sharp decrease in additive genetic variances (σ_{AP}^2).

The ratios of estimated dominance genetic variances to additive genetic variances in the current study (5 to 11%) were generally a little smaller than in other studies on TNB. Vitezica et al. [22] reported that this ratio was equal to about 20% for litter size in both purebred and crossbred lines by using the same trivariate GBLUP model. Esfandyari et al. [19] stated that, by using purebred genomic information in a univariate Bayesian mixture model at the SNP level, the ratio between dominance variance and additive variance for TNB was equal to 15 and 18% for breeds L and Y, respectively. Based on pedigree information, Misztal et al. [10] reported a ratio that reached about 25% for number of piglets born alive in a Yorkshire population. However, there are some studies that did report smaller ratios than those reported here. For instance, Hidalgo et al. [20] reported that, based on genotyped crossbred animals, the dominance variance for TNB accounted for nearly zero of the total genetic variance and concluded that TNB was not affected by dominance effects in the Dutch Landrace and Yorkshire populations. For other traits or species, different ratios of dominance genetic variance to additive genetic variance were also reported. For average daily gain in Duroc pigs, Su et al. [9] estimated a ratio of 15%, but their results were based on genotypic variance components and cannot be directly compared to genetic variance components [32]. For average daily weight gain in Yorkshire and Landrace pigs, Lopes et al. [44] reported ratios of 13.8 and 28%, respectively by including genomic information. For Fleckvieh cattle, Ertl et al. [12] calculated ratios that ranged from 3.4% for stature to 69% for protein yield by using a univariate SNP-BLUP model. Overall, these different ratios of dominance genetic variance to additive genetic variance may reflect differences in the traits analyzed and in the type of information used for the estimation [9], and also uncertainty in the estimates.

The genetic correlation of breeding values between purebred and crossbred performances (r_{PC}) is a key parameter in crossbreeding schemes [2]. In the current study, the estimated r_{PC} was in line with results reviewed by Wei et al [3]. Lutaaya et al. [5] also reported r_{PC} that ranged from 0.32 to 1. Such differences in r_{PC} may reflect differences in the extent of GxE interactions and the distance across breeds. In our study, estimated r_{PC} did not vary dramatically across the scenarios, when the standard errors were taken into account. These standard errors were very large, which indicated that the amount of available information was too small to ensure accurate r_{PC} estimates. Across scenarios, standard errors of r_{PC} decreased when genomic information was included, which indicates that including genomic information may reduce the uncertainty of the estimations. r_{PC} was larger for breed L than for breed Y, which was in agreement with a previous study [17] and may be due to the data structure. Among the 5143 crossbred animals, the number of Yorkshire sires (N = 1125) was much smaller than that of Landrace sires (N = 4018). Such a different amount of information affects the accuracy of the estimates, and thus the standard error of r_{PC} was larger for breed Y than for breed L (see Table 2). However, compared to the results reported in [17], the r_{PC} for breed L increased by about 10% while that for breed Y did not change much. Both pre-correction of data and the genotyped subset of original data used may play a role in the differences observed between the current and previous results [17]. In the previous study, a single-step method, which can use pedigree information and genomic information simultaneously, was used. In this study, the use of only phenotypic records on genotyped individuals affected the accuracy of estimates. Our results confirmed the moderate value of the r_{PC} for TNB in breeds L and Y.

To our knowledge, this is the first time that correlations of allele substitution effects for both purebred ($r_{\alpha P_L,\alpha P_Y}$) and crossbred ($r_{\alpha C_L,\alpha C_Y}$) performance between two breeds in the presence of dominance variation are estimated. In genomic selection, SNPs are assumed to be in LD with QTL along the whole genome [45]. The correlation of allele substitution effects between breeds measures the degree of average similarities between SNP effects assuming that the QTL effects are the same in breeds 1 and 2 [38-40]. In practice, the correlation of allele substitution effects between two breeds can be interpreted as indicating "how consistent the SNP substitution effects are across two breeds". For purebred performance, the estimated SNP substitution effects were based on the within-breed allele frequencies. A high $r_{\alpha P_L,\alpha P_Y}$ correlation means that the estimated SNP substitution effects based on allele frequencies from breed L can be used for breed Y and vice versa. However, $r_{\alpha P_L,\alpha P_Y}$was not significantly different from 0 in the current study, which demonstrates that SNP effects estimated from a reference population that consists of one pure breed (e.g. Landrace) cannot be readily applied to the other breed (e.g. Yorkshire). This was in agreement with the findings of [46] who reported that prediction based on an across-population reference panel was worse than within-population prediction. In other species, estimated correlations of allele substitution effects between breeds based on models without dominance, oscillate between 0 and 0.8, and are trait-dependent [38, 47]. For crossbred performance, an $r_{\alpha C_L,\alpha C_Y}$ close to 1 was found in the current study, which indicated that the allele

substitution effects based on the allele frequencies from the opposite breeds were very similar for the L and Y breeds. In practice, this suggests that SNP substitution effects that are estimated based on a reference population consisting of crossbred animals can be used to estimate crossbred breeding values for both breeds L and Y.

It was expected that genomic evaluations obtained by including dominance deviations in the model would be improved, especially when records of crossbred animals were included [9]. However, our results showed that inclusion of dominance deviations did not increase the predictive ability for crossbreds. This result was in line with conclusions in [9, 12, 20], but was opposite to those in [18, 19, 48, 49]. Theoretically, estimating dominance genetic effects should be useful because ignoring them will result in less accurate estimates of allele substitution effects and consequently less accurate estimated breeding values in genomic prediction [11]. However, regarding the additive genetic variance, estimates were nearly the same in scenarios *Gen_AM* and *Gen_ADM*, which demonstrated that the additive variances were already well captured by the additive model. Thus, the accuracy of the estimated additive genetic effects was not affected when dominance effects were included in the model [12]. Moreover, a simulation study at the level of the gene action showed that when all gene actions were purely additive, including dominance in addition to the additive effects in the model was not advantageous compared to using an additive model. Hidalgo et al. [20] showed that TNB was not affected by dominance in the Dutch crossbred population. In the current study, we also observed similar results, and dominance variation accounted for a small proportion of the total genetic variation (4 to 10%). The lack of change in predictive ability also indicated the difficulty of distinguishing dominance genetic effects from additive genetic effects [9], but it confirmed a previous simulation study that concluded that the use of a dominance model did not negatively affect genomic evaluation even if the trait was purely additive [18].

Scenarios in which genomic information was included (*Gen_AM* and *Gen_ADM*) showed higher predictive abilities than the pedigree-based scenario (*Nogen*). For the *Nogen* scenario, the relationship matrix was constructed based on a base population that was considered as a mixture of L and Y animals, which was not the case. Therefore, the results of the *Gen_AM* and *Gen_ADM* scenarios were more reliable than those of the *Nogen* scenario. Although predictive abilities were not significantly different according to the Hotelling-Williams t-test, the results from 1000 bootstrap samples still showed that the predictive abilities of about 90% of the crossbred animals would be higher when genomic information was available (894 of 1000 bootstrap samples showed higher predictive abilities in scenarios that included genomic information than those in the *Nogen* scenario; results not shown). Comparison of the predictive abilities that were estimated in the current study with those from a previous study [17] indicated that the single-step model [16] might be more robust than the Vitezica model [22] used in this paper in terms of both predictive ability and unbiasedness for the crossbred performance. Our results suggested that using a small set of genotyped animals and pre-corrected

data to implement genetic evaluation for crossbred performance was less powerful than using the whole dataset, which is similar to the conclusions for purebred performance [43].

The regression coefficients obtained with the Vitezica model were less than 1, which suggests that variations in total genetic effects could be overestimated (inflated). In terms of unbiasedness, there was no clear trend among the scenarios examined, regardless of whether genomic information was included or not. Overall, unbiasedness was not a problem in the current study because the regression coefficients in all scenarios did not significantly differ from 1.

Inbreeding depression for litter size in pigs is a well-known phenomenon [50, 51], and we found that inclusion of inbreeding effects in the model improved predictive abilities of crossbred animals. Estimates of inbreeding depression effects are rarely reported, but our estimates agree with those previously reported for commercial and Iberian pigs [52]. Inbreeding depression was, for the same amount of marker-based inbreeding, more detrimental in the Landrace than in the Yorkshire breed. There are many possible explanations among which the purging of lethal recessive alleles [53]. We also report an estimate of the inbreeding depression parameter for the crossbred animals, which is between the estimates for the parental breeds. To our knowledge, this estimate has never been reported.

The correlation between breeding values and dominance deviations is of theoretical concern [30]. However, this does not apply to the current marker-based analyses for the following reasons. (1) In a pedigree-based analysis, mating in an inbred population produces deviations from the Hardy-Weinberg equilibrium, which generate correlations between breeding values and dominance deviations [30]. However, in our study, SNPs are in Hardy-Weinberg equilibrium if allele frequencies are considered in the current generation. (2) Such a correlation occurs because the pedigree information forces the genetic model to refer to the base population, since the state of alleles is not known, i.e. only probabilities of IBD are known. In our study, the states of alleles are known and the model can be described as referring to the current generation instead. (3) The equivalent GBLUP models in Equation (3) used genotypic additive and dominance values, not breeding values and dominance deviations. A reasonable assumption in the model is that additive and dominance effects are unrelated at each SNP. Thus, covariance between additive and dominance genetic effects was ignored in the current study.

Conclusions

We present for the first time the use of genomic inbreeding in crossbred and purebred genomic evaluation. Estimates are biologically sound and are relevant even for crossbred animals. We also report for the first time, estimated correlations of allele substitution effects in the presence of dominance. For TNB, the dominance genetic variance accounts for only a small proportion of the total genetic variation (4 to 10%). A moderate, positive genetic correlation between breeding values for TNB for purebred and crossbred performances was confirmed. Inclusion of dominance in the GBLUP model did not improve predictive ability for crossbred animals, whereas inclusion of inbreeding depression effects did. An additive GBLUP model is sufficient to capture the additive genetic variances and for genomic evaluation. The GBLUP model [22] was applied successfully for genetic evaluations for crossbred performance in pigs. This model can potentially be a useful tool in genetic evaluation for crossbred performance.

Declarations

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TX performed data analysis and wrote the manuscript. All authors participated in the derivation of the theory. AL and OFC coordinated the project, conceived the study, made substantial contributions for the interpretation of results and revised the manuscript. ZGV improved the manuscript and added valuable comments during the study. All authors read and approved the manuscript.

Author details

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830 Tjele, Denmark. [2]INRA, UR1388 GenPhySE, CS-52627, F-31326 Castanet-Tolosan, France. [3]Université de Toulouse, INP, ENSAT, GenPhySE, F-31326 Castanet-Tolosan, France.

Acknowledgements

The work was funded through the Green Development and Demonstration Programme (grant no. 34009-12-0540) by the Danish Ministry of Food, Agriculture and Fisheries, the Pig Research Centre and Aarhus University. The first author benefits from a joint grant from the European Commission and Aarhus University, within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG". AL and ZGV thank financial support from the INRA SelGen metaprogram projects X-Gen and SelDir. OFC acknowledges funding from the GenSAP project. We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing resources. Discussions with Luis Varona are gratefully acknowledged.

References

1. Wei M, Van der Werf JHJ. Maximizing genetic response in crossbreds using both purebred and crossbred information. Anim Sci. 1994;59:401-13.

2. Bijma P, Bastiaansen JWM. Standard error of the genetic correlation: how much data do we need to estimate a purebred-crossbred genetic correlation? Genet Sel Evol. 2014;46:79.

3. Wei M, Van der Steen HAM. Comparison of reciprocal recurrent selection with pure-line selection systems in animal breeding (a review). Anim Breeding Abs. 1991;59:281-98.

4. Dekkers JCM. Marker-assisted selection for commercial crossbred performance. J Anim Sci. 2007;85:2104-14.

5. Lutaaya E, Misztal I, Mabry JW, Short T, Timm HH, Holzbauer R. Genetic parameter estimates from joint evaluation of purebreds and crossbreds in swine using the crossbred model. J Anim Sci. 2001;79:3002-7.

6. Lo LL, Fernando RL, Grossman M. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. J Anim Sci. 1997;75:2877-84.

7. Charlesworth D, Willis JH. The genetics of inbreeding depression. Nat Rev Genet. 2009;10:783-96.

8. Falconer DS, Mackay TFC. Introduction to quantitative genetics. New York: Longman Group Ltd; 1981.

9. Su G, Christensen OF, Ostersen T, Henryon M, Lund MS. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One. 2012;7:e45293.

10. Misztal I, Varona L, Culbertson M, Bertrand JK, Mabry J, Lawlor TJ, et al. Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. Biotechnol

Agron Soc Environ. 1998;2:227-33.

11. Toro MA, Varona L. A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol. 2010;42:33.

12. Ertl J, Legarra A, Vitezica ZG, Varona L, Edel C, Emmerling R, et al. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet Sel Evol. 2014;46:40.

13. Fulton JE. Genomic selection for poultry breeding. Anim Front. 2012;2:30-6.

14. Loberg A, Dürr JW. Interbull survey on the use of genomic information. Interbull Bull. 2009;39:3-14.

15. Christensen OF, Legarra A, Lund MS, Su G. Genetic evaluation for three-way crossbreeding. Genet Sel Evol. 2015;47:98.

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

17. Xiang T, Nielsen B, Su G, Legarra A, Christensen OF. Application of single-step genomic evaluation for crossbred performance in pig. J Anim Sci. 2016;94:936-48.

18. Zeng J, Toosi A, Fernando RL, Dekkers JCM, Garrick DJ. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol. 2013;45:11.

19. Esfandyari H, Bijma P, Henryon M, Christensen OF, Sorensen AC. Genomic prediction of crossbred performance based on purebred Landrace and Yorkshire data using a dominance model. Genet Sel Evol. 2016;48:40.

20. Hidalgo AM. Exploiting genomic information on purebred and crossbred pigs. PhD Thesis. Swedish University of Agricultural Sciences, Uppsala. 2015.

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

22. Vitezica ZG, Varona L, Elsen MJ, Misztal I, Herring W, Legarra A. Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genet Sel Evol. 2016;48:6.

23. Madsen P. DMU Trace, A program to trace the pedigree for a subset of animals from a large pedigree file. Version 2. Center for Quantitative Genetics and Genomics. Dept. of Molecular Biology and Genetics. Aarhus University; 2012.

24. Ramos AM, Crooijmans RP, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a high

density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One. 2009;4:e6524.

25. GeneSeek Company. GGP-for Porcine LD (GeneSeek Genomic Profiler for Porcine Low Density). 2012, http://www.neogen.com/Genomics/pdf/Slicks/GGP_PorcineFlyer.pdf.

26. Xiang T, Ma P, Ostersen T, Legarra A, Christensen OF. Imputation of genotypes in Danish purebred and two-way crossbred pigs using low-density panels. Genet Sel Evol. 2015;47:54.

27. Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet. 2008;124:439-50.

28. Silió L, Rodríguez M, Fernández A, Barragán C, Benítez R, Óvilo C, et al. Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. J Anim Breed Genet. 2013;130:349-60.

29. Lynch M, Walsh B. Genetics and analysis of quantitative traits. 1st ed. Sunderland: Sinauer Assoc; 1998.

30. de Boer IJM, Hoeschele I. Genetic evaluation methods for populations with dominance and inbreeding. Theor Appl Genet. 1993;86:245-58.

31. Aliloo H, Pryce JE, Gonzalez-Recio O, Cocks BG, Hayes BJ. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genet Sel Evol. 2016;48:8.

32. Vitezica ZG, Varona L, Legarra A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics. 2013;195:1223-30.

33. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH. BLUPF90 and related programs (BGF90). In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production: Montpellier;19-23 August, 2002.

34. Madsen P, Jensen J. A user's guide to DMU. Version 6, release 5.2. Center for Quantitative Genetics and Genomics, Dep. of Molecular Biology and Genetics, Aarhus University. Tjele; 2013.

35. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414-23.

36. Houle D, Meyer K. Estimating sampling error of evolutionary statistics based on genetic covariance matrices using maximum likelihood. J Evol Biol. 2015;28:1542-9.

37. Lo LL, Fernando RL, Grossman M. Covariance between relatives in multibreed populations: additive model. Theor Appl Genet. 1993;87;423-30.

38. Karoui S, Carabaño MJ, Díaz C, Legarra A. Joint genomic evaluation of French dairy cattle breeds using

multiple-trait models. Genet Sel Evol. 2012;44:39.

39. Wientjes YC, Veerkamp RF, Bijma P, Bovenhuis H, Schrooten C, Calus MP. Empirical and deterministic accuracies of across-population genomic prediction. Genet Sel Evol. 2015;47:5.

40. Porto-Neto LR, Barendse W, Henshall JM, McWilliam SM, Lehnert SA, Reverter A. Genomic correlation: harnessing the benefit of combining two unrelated populations for genomic selection. Genet Sel Evol. 2015;47:84.

41. Meuwissen THE, Luo Z. Computing inbreeding coefficients in large populations. Genet Sel Evol. 1992;24;305-13.

42. Mäntysaari EA, Koivula M. GEBV validation test revisited. Interbull Bull. 2012;45:11-6.

43. Guo X, Christensen OF, Ostersen T, Wang Y, Lund MS, Su G. Improving genetic evaluation of litter size and piglet mortality for both genotyped and nongenotyped individuals using a single-step method. J Anim Sci. 2015;93:503-12.

44. Lopes MS, Bastiaansen JWM, Janss L, Knol EF, Bovenhuis H. Estimation of additive, dominance, and imprinting genetic variance using genomic data. G3 (Bethesda). 2015;5:2629-37.

45. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819-29.

46. Hidalgo AM, Bastiaansen JWM, Lopes MS, Harlizius B, Groenen MAM, de Koning D-J. Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3 (Bethesda). 2015;5:1575-83.

47. Legarra A, Baloche G, Barillet F, Astruc JM, Soulas C, Aguerre X, et al. Within-and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. J Dairy Sci. 2014;97:3200-12.

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

49. Sun C, VanRaden PM, Cole JB, O'Connell JR. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PLoS One. 2014;9: e103934.

50. Dickerson GE. Inbreeding and heterosis in animals. In Proceedings of the Animal Breeding and Genetics Symposium in Honor of Jay L. Lush: 29 July 1972; Blacksburg. 1973:54-77.

51. Leroy G. Inbreeding depression in livestock species: review and meta-analysis. Anim Genet. 2014;45:618-28.

52. Silió L, Barragán C, Fernández AI, García-Casco J, Rodríguez MC. Assessing effective population size, coancestry and inbreeding effects on litter size using the pedigree and SNP data in closed lines of the Iberian pig breed. J Anim Breed Genet. 2016;133:145-54.

53. Hinrichs D, Meuwissen THE, Ødegard J, Holt M, Vangen O, Woolliams JA. Analysis of inbreeding depression in the first litter size of mice in a long-term selection experiment with respect to the age of the inbreeding. Heredity. 2007;99:81-8.

CHAPTER 6

General discussion

Introduction

Genomic selection (Meuwissen et al., 2001) has been widely used in the pig industry in different countries. In pig industry, selection intensities are already high and the generation intervals are not long. However, the accuracies of pedigree-based conventional EBVs are low, especially for the traits with sex-limitation, low heritabilities and that cannot be recorded early in life or are difficult to be measured (Muir, 2007). Therefore, when compared with dairy cattle (Hayes et al., 2009), the increased genetic gain through using GS is mainly due to the improved accuracies of EBVs in pig.

For purebred performance, pig breeding companies, such as DanAvl, PIC and Topigs Norsvin, commonly chose the single-step GBLUP method (Legarra et al., 2009; Christensen and Lund, 2010) as the standard approach in their routine genetic evaluation. Due to the economical restrictions, it is currently unfeasible to genotype all animals. In DanAvl, potential breeding candidates must be tested and breeding indices are calculated. The breeding indices are used as criterion for selecting which animals will be genotyped (Eskildsen and Weber, 2016). Then more accurate breeding indexes are calculated based on both pedigree and genomic information, by using single-step GBLUP method.

Recently, the single-step GBLUP method has been theoretically extended to crossbred performance (Christensen et al., 2014; Christensen et al., 2015; Legarra et al., 2015), because pig production largely relies on the crossbreeding system and performance of purebred animals is not an optimal predictor for its crossbred performance (Dekkers, 2007). This method had to be verified in real datasets. Hence, this thesis was initially inspired by the question on how to apply single-step method for crossbred performance in a feasible way.

To find out solutions, in chapter 2, genotype imputation was first implemented in both Danish purebred and crossbred pigs using low-density panels. This chapter was the precondition for the applications of GS in other chapters. Weigel et al. (2010) and Cleveland and Hickey (2013) demonstrated that accuracies of GS did not decline significantly when imputed SNP markers were used instead of real genotyped markers, whereas the costs of genotyping reduced sharply. In chapter 2, differences in imputation accuracies between purebred and crossbred animals were compared from 5K to 8K panels first and then, according to the "optimal" imputation strategy, imputation was processed from 8K to 60K panels in crossbred animals as the second step. The study confirmed that to impute two-way crossbred genotypes correctly, both parental breeds should be included in the reference population, which was in line with Moghaddar et al. (2015), and imputation worked almost as well in crossbred animals as in purebred animals. Then both purebred and (imputed) crossbred genomic information was used in chapter 3, 4 and 5. In chapter 3 and 4, single-step GBLUP method was used to investigate both the purebred and crossbred performances. These chapters considered either two breed-specific partial relationship matrices or one relationship matrix with two

metafounders to account for genomic relationships across three pig populations. Results were mutually consistent, which was reassuring. They both confirmed the existence of a moderate, positive genetic correlation for breeding values between purebred and crossbred performances for TNB. Models with genomic information, especially from crossbred animals, can improve model-based reliabilities for crossbred performance of purebred boars and also improve predictive abilities for validated crossbred animals. The common genetic approach in chapter 4 performed at least as well as the partial genetic approach in chapter 3, and was also easier to implement. Chapter 3 and 4 indicated that the single-step method is applicable for genomic evaluation for both purebred and crossbred performances. However, single-step method is difficult to extend to account for dominance effects, but dominance is the main genetic basis for heterosis. Hence, investigation on models including dominance effects was studied in chapter 5. The estimated correlations of allele substitution effects of markers across breeds were reported for the first time. Results indicated that TNB was only slightly affected by the dominant gene actions and the accuracies of predictions were not improved by including dominance effects in the model, but genomic inbreeding depressions are relevant, even in crossbreds, and their inclusion can efficiently improve the performance of prediction.

In this general discussion part, I will mainly discuss the shortcomings or possible improvements for the thesis. For those issues that have been discussed in the respective chapters, I will not repeat them here.

Genotype Imputation

Imputation of genotypes in both purebred and two-way crossbred pigs using low-density panels was done in chapter 2. Results indicated that to impute crossbred genotypes accurately, the reference population should be composed of animals from both parental breeds. We also concluded that close relatives in the reference population to the target animals in the imputed population play a much higher important role than distant relatives, in line with Huang et al. (2012) and Pszczola et al. (2012). Although the strategy of imputing crossbred genotypes was considered as “optimal” in chapter 2, to enhance the imputation accuracies for crossbred genotypes further, the reference population can still be optimized. In chapter 2, mainly purebred sires of crossbred animals were genotyped and included in the reference population, but neither their dams nor sibs were. Moghaddar et al. (2015) successfully used crossbred reference population to impute crossbred Merino sheep and they obtained ~0.88 correlation coefficients between imputed and real genotypes. They concluded that crossbreds need larger imputation reference sets that included genotypes for all relevant breeds than purebreds. However, with additional funds, dams should be genotyped in a higher priority than the crossbred sibs for two reasons: 1) higher proportion of haplotypes can be shared between the parents and imputed offspring than between crossbred sibs; 2) purebred parents are closely related to the purebred selection candidates.

Comparison of different imputation software was beyond the scope of chapter 2. Beagle version 3.3.2 used algorithms that depend crucially on local LD patterns across markers and thus, its performance may be breed-specific. Although Ma et al. (2013) recommended Beagle as the most robust software for imputation, the family structure information was not used during the process of imputation. Johnston and Kistemaker (2011) and Ventura (2013) showed that the family based software Fimpute (Sargolzaei et al., 2011) slightly outperformed Beagle, and imputation time was reduced to one-twentieth of that in Beagle. Thus, it would be interesting to test other software in imputing crossbred genotypes as well.

In chapter 2, only autosomes were imputed and they were used for genomic evaluation in the other chapters. However, Hickey and Kranis (2013) pointed out that a large portion of genome is in sexual chromosomes. Although X chromosomes were considered more difficult to impute accurately than autosomes, the markers on the X chromosomes contributed to the accuracy of genomic evaluation and inclusion of genomic information of the X chromosomes can increase the reliability of genomic prediction (Su et al., 2014). Thus, genotype imputation should also be carried out on the X chromosome in chapter 2 if the low density genotypes were available for X chromosome.

Models

We investigated different approaches and scenarios of genomic evaluation for crossbred performance in pigs in chapter 3, 4 and 5. Single-step GBLUP was applied in chapter 3 (partial genetic approach) and chapter 4 (common genetic approach) and GBLUP method including dominance and genomic inbreeding was used in chapter 5. Chapter 3 and 4 used additive models where non-additive effects at the individual level were ignored, while dominance effects and inbreeding depressions were included explicitly in GBLUP model in chapter 5. It is difficult to estimate dominant relationships accurately based on the pedigree information. In combination with computational complexities, single-step method has not been extended to account for dominance effects yet (Legarra et al., 2014). Although the additive models in chapter 3 and 4 did not explicitly contain dominance effects, it may still capture parts of dominant gene actions and other non-additive gene actions (Christensen et al., 2014), because substitution effects involve functional dominance and epistatic effects. For instance, the r_{pc} smaller than 1 could partly be due to the dominant gene actions.

The model used in chapter 3 was the model from Christensen et al (2014), which was an extension of full model in Wei and Van der Werf (1994). Christensen et al. (2014) chose to extend the full model rather than reduced model in Wei and Van der Werf because crossbred marker genotypes provide information on the Mendelian sampling of the additive genetic effects for crossbred animals, and such genomic information cannot be incorporated in the reduced model.

In chapter 3, the model contains two breeding values for purebred and crossbred performance and these two breeding values are correlated by r_{pc}. This model is more sophisticated than models in other studies (Dekkers, 2007; Ibánẽz-Escriche et al., 2009; Hidalgo, 2015; Lopes, 2016) because those models can only be used for evaluating either purebred or crossbred performance for each animal at one time.

In chapter 3, it was assumed that genetic variances were different in the two pure breeds. It used two breed-specific variance and covariance structures between additive genetic effects for purebred and crossbred breeding values, which were $\begin{bmatrix} \sigma^2_{a_L} & \sigma_{a_L c_L} \\ \sigma_{c_L a_L} & \sigma^2_{c_L} \end{bmatrix}$ and $\begin{bmatrix} \sigma^2_{a_Y} & \sigma_{a_Y c_Y} \\ \sigma_{c_Y a_Y} & \sigma^2_{c_Y} \end{bmatrix}$ for Landrace and Yorkshire, respectively. The additive genetic variance for crossbred animals is $\sigma^2_{A_{LY}} = 0.5 * (\sigma^2_{c_L} + \sigma^2_{c_Y})$. In chapter 3, it was assumed that the covariance between Landrace and Yorkshire was 0, which was in line with the pedigree information that base individuals in different populations were unrelated. A special case in chapter 3 is when the additive genetic variances for crossbred performance are same for Landrace and Yorkshire and thus, $\sigma^2_{A_{LY}} = \sigma^2_{c_L} = \sigma^2_{c_Y}$ (Christensen et al. 2014). In such case, instead of using breed-specific variance and covariance matrices, one combined (co)variance matrix ($\mathbf{G_{special}}$) across three populations can be used as

$$\mathbf{G_{special}} = \begin{bmatrix} \sigma^2_{a_L} & 0 & \sigma_{a_L c_L} \\ 0 & \sigma^2_{a_Y} & \sigma_{a_Y c_Y} \\ \sigma_{c_L a_L} & \sigma_{c_Y a_Y} & \sigma^2_{A_{LY}} \end{bmatrix},$$

and there would be no need to trace the breed-origin of crossbred genotypes. This (co)variance structure looks similar to the $\mathbf{G_0}$ matrix ($\begin{bmatrix} \sigma^2_{A_L} & \sigma_{A_L A_Y} & \sigma_{A_L A_{LY}} \\ \sigma_{A_Y A_L} & \sigma^2_{A_Y} & \sigma_{A_Y A_{LY}} \\ \sigma_{A_{LY} A_L} & \sigma_{A_{LY} A_Y} & \sigma^2_{A_{LY}} \end{bmatrix}$) used in chapter 4, but actually with several differences. Two differences were discussed in the method section in chapter 4: a) with the concept of metafounder, the base animals in different populations become related, and the additive genetic covariance between Landrace and Yorkshire is not 0 anymore; b) genetic parameters in $\mathbf{G_{special}}$ were the usual genetic variances, but in chapter 4, those parameters corresponded to the unusual situation that base populations were related. Another difference is that, in this special case, the compatibility between pedigree-based and marker-based relationship matrices could be a problem because it is difficult to adjust the marker-based relationship matrix for both two pure breeds at the same time. However, with the concept of metafounder in chapter 4, pedigree-based matrix is compatible with the marker-based matrix automatically. Conversely, having metafounders in the special case in chapter 3 is also possible. An underlying assumption is that the relatedness across base populations in Landrace and Yorkshire is 0 (that is $\gamma_{L,Y} = 0$). $\mathbf{G_0}$ matrix in chapter 4 can be regarded as an extension of such special case $\mathbf{G_{special}}$.

It has to be understood that although the notation of $\mathbf{G_0}$ matrix was also used in chapter 5, it had totally different meanings from in chapter 4. In chapter 4, the $\mathbf{G_0}$ matrix was associated with breeding values, but in chapter 5, it was associated with genotypic additive effects. Vitezica et al. (2013) compared the differences between breeding values and genotypic additive effects and concluded that parameters from these two scopes were not comparable directly. Based on Vitezica et al. (2016), $\mathbf{G_0}$ matrix in chapter 5 was transferred to the additive (co)variances of SNP effects first and then transferred to the individual scaled genetic parameters, which is comparable with traditional pedigree-based genetic parameters directly. Note that as we mentioned above, $\mathbf{G_0}$ matrix in chapter 4 had to be transferred to the scale that base populations were unrelated, and then it can also be comparable with pedigree-based genetic parameters directly.

Comparing the magnitudes of predictive abilities and the unbiasedness across chapters 3, 4 and 5, single-step genomic model worked better than the GBLUP model with precorrected data. However, it does not mean GBLUP model always works poorer than the single-step method. It depends on how many animals are genotyped and which animals are genotyped. Furthermore, the proportion of genetic variances that can be captured by genotypes also affect the performance of different models. For traits that are not much affected by dominant gene actions, single-step additive model seems to be a reasonable choice.

Genetic correlations between purebred and crossbred performances

The parameter of genetic correlation between breeding values for purebred and crossbred performances (r_{pc}) is crucial in a crossbreeding system, since it determines the need of including crossbred information in genetic evaluation. Due to the genotyped by environment interactions, dominance effects, and different allele frequencies in different populations, the r_{pc} is usually not equal to 1. Thus, in chapter 3, 4 and 5, the same "biological" trait (TNB) was regarded as different phenotypes in purebreds and crossbreds.

The r_{pc} varied largely in different chapters (0.43~0.70 for YY and 0.70~0.95 for LL) due to the different amount of information that was used to estimate the r_{pc}. The reported r_{pc} in other studies also ranged dramatically depending on the specific traits and the associated environments. Usually, it is difficult to estimate r_{pc} accurately and the SE of r_{pc} are high. Bijma and Bastiaansen (2014) showed that if only pedigree relationships were available, the SE of r_{pc} was affected by the number of sire families and the reliabilities of sire EBVs. For a trait with $h^2 = 0.30$, more than 100 half-sib families are needed to make the SE of r_{pc} be smaller than 0.05. In this thesis, the number of sire families were all smaller than 10 and thus, it was hard to have accurate r_{pc}. The r_{pc} is related to both estimated additive and dominant genetic (co)variances and allele frequencies in different populations and thus, it changes after long-term selection. For a trait that is purely affected by additive genetic effects, the r_{pc} would not change under selection (Wei, 1992). However, if the trait is partly influenced by dominance effects, r_{pc} changes in different directions, e.g:

a) r_{pc} tends to increase after the purebred selection because the differences between allele frequencies in different populations decline (Wei and Steen, 1991); b) if overdominance exists, with combined purebred and crossbred selection, r_{pc} usually decreases because alleles with opposite effects tend to be neutral (Swan, 1992). Wei and Steen (1991) thus considered the change of r_{pc} as a detector for the existence of overdominance. Generally, the r_{pc} should be re-estimated frequently so that the efficiency of selection in purebred nucleus herd for crossbred performance can be kept.

The r_{pc} is always associated with a certain environment (Wei, 1992). An interesting topic is how to decompose the r_{pc} into different components because this can help to improve the design of a reference population (Wei, 1992; Esfandyari, 2016). If the environment can be identical in the purebred and crossbred herds and the calculated r_{pc} is still lower than 1, it indicates that using a model that accounts for dominance effects could be an efficient way of improving genomic evaluation for crossbred performance; if the same traits that recorded in the diverse environments showed there is a strong G by E interactions, it indicates that using purebred reference cannot achieve optimal accuracy of r_{pc} for crossbred performance and crossbred animals should be included in the reference population; if both G by E interactions and dominance contribute to the $r_{pc} < 1$, model including dominance effects in combination with a combined purebred and crossbred reference population was recommended to be used (Dufrasne, 2015).

In chapter 5, both r_{pc} and correlations of allele substitution effects between two breeds were calculated based on the substitution effects summarized over all the marker loci. If there are no dominance effects existing, the $r_{pc_L} = \frac{\sigma_{a_{L,LY}}}{\sigma_{a_L}\sigma_{a_{LY}}}$, and the correlation of allele substitution effects for purebred performance between L and Y is $r_{\alpha P_L,\alpha P_Y} = \frac{\sigma_{a_{L,Y}}}{\sigma_{a_L}\sigma_{a_Y}}$. They have similar structure and it is easy to replace one parameter by another one to estimate both these two genetic correlations. If dominance effects exist, how to relate these two correlations could be an interesting topic.

Obstacles in applying genomic selection for crossbred performance

In this thesis, we used data recording from both purebred and crossbred animals. The single-step genomic evaluation approach requires different types of information. Thus, to implement the approaches in the thesis, several difficulties need to be overcome beforehand.

First, some traits are difficult to measure accurately in either purebred or crossbred herds. Usually, phenotypes of purebred animals living in nucleus herds are recorded by automatized devices, which are standard and objective. However, in the commercial herds, due to the differences of recording systems and environments, biases depending on the technologies and technicians often appear (Dufrasne, 2015).

Furthermore, among the commercial herds, the environments and devices may not be homogeneous. Additionally, for some traits, like resistance to diseases, they can only be recorded in the commercial herds because in the nucleus herds, the diseases have been eliminated by bio-security (Esfandyari, 2016). Therefore, the first challenge is how to collect phenotypes in both purebred and crossbred herds in consistent criteria.

Second, in practice, the pedigree of crossbred animals is usually unknown. Normally, companies have a large amount of commercial animals and it is not easy to record the pedigree. For those genotyped purebred animals, some of them are distant ancestors for the crossbred animals and several generation gaps may exist between purebred and crossbred animals. This increases the difficulties of tracing pedigree. However, to implement single-step approach, accurate pedigree information from crossbred offspring to purebred ancestors is essential. One possible solution is to use parentage assignation chips (Clarke et al. 2014) to correct pedigree information, which is low cost.

The third one is how to trace the breed-origin of crossbred alleles. Ibáñez-Escriche et al. (2009) applied a model with breed-specific SNP effects to fit crossbred phenotypes, which required the breed-origin of crossbred alleles are known. Esfandyari et al. (2015) found that when the pure breeds in training population were distantly related, the tracing of the breed origin of alleles in crossbreds can improve genomic prediction for crossbred performance. In Christensen model (2014), the breed-origin of crossbred alleles is assumed to be known. Thus, tracing the breed-origin of crossbred alleles is crucial. Although an algorithm was applied in chapter 3, the accuracy of assigning crossbred alleles to its breed origin is hard to evaluate. This algorithm largely depended on the accuracies of phasing (by Beagle 3) in crossbred animals. Therefore, even if reliabilities in chapter 4 were higher than those in chapter 3, we cannot conclude whether it is due to the robustness of the model with metafounders or due to the lack of accuracies in the alleles tracing in chapter 3. Several other algorithms have also been suggested to trace crossbred alleles. For instance, Bastiaansen et al., (2014) used a long-range phasing method without the need of pedigree information of crossbred animals, which was easy to implement in practice. The advantage of such method is that even if the relationships between purebred and crossbred animals are low, the long-range method can still work. This approach was based on results from software AlphaPhase (Hickey et al., 2012) and they considered that if more than 90% of a haplotype was located to one breed, the whole haplotype was from that breed. This approach has similar problem as the approach used in chapter 3 because wrong phasing (crossing-over between parental haplotypes) cannot be totally avoided. To avoid the possible errors from crossing-over, Sevillano et al. (2016) and Vandenplas et al. (2016) showed a "BOA" approach that assign breed origin to phased haplotypes. To ensure the correctness of phasing, pedigree information was needed and they showed that the average accuracy of correctly assigning crossbred alleles to its breed-origin is around 90%. These accuracies increased with distance increasing between parental breeds, but still difficult to reach accuracy of 100%.

Simulation studies can be used to compare different approaches and the consequences of error allele tracing on genomic evaluation can be further investigated.

Problems on convergence appeared during the process of parameter estimation. REML variance component estimation in models with non-additive effects converged much slower than the additive model, e.g: model in chapter 5 took around one month to converge but it took only 3 days to converge for the model in chapter 3. However, for chapter 4, although an additive model was used, due to the increased number of genetic parameters within one relationship, one day was needed for each REML iteration. Thus, Gibbs sampling was first used to get ballpark estimates of starting values for REML, but this was still inefficient. With the increased amount of genomic information, it seems to be a challenge to run the genomic evaluation for crossbred performance efficiently.

Future perspectives

In this thesis, each individual had two breeding values: one for purebred performance and another one for crossbred performance. A question is how to set the selection criteria for the next generation. Bijma and Van Arendonk (1998) argued that the breeding goal of the pig industry is to maximize the crossbred performance and thus, selection should be done based on breeding values for crossbred performance. Dekker (2007) and Lutaaya et al. (2001) found that selection based on crossbred performance can still make genetic gains in purebred lines. However, Wei et al. (1991) showed that if selection is based on crossbred performance, purebred performance would be improved more slowly, sometimes even reduced, but the inbreeding levels may increase quickly, which is not desired. Thus, they suggested breeders to use the combined purebred and crossbred breeding values as the selection criteria. The combined breeding values were calculated as $\mathrm{BV} = \mathrm{w}_p * BV_p + \mathrm{w}_c * BV_c$, where the w_p and w_c were weights of breeding values for purebred and crossbred performances; BV_p and BV_c were breeding values for purebred and crossbred performances, respectively. The relative weights were trait and species dependent. For the reproductive traits in pigs, w_c usually was much larger than w_p because the number of commercial animals were much larger than the number of breeding animals in purebred lines. For cattle and sheep, due to the low reproductive rate, magnitude of w_p is usually higher than the w_c (Wei and van der Werf, 1994). However, in practice, industries prefer to keep the breeding system simple and they pay much attention to the genetic gains within purebred lines and therefore, they usually just focus on purebred performance. If the r_{pc} is close to 1, there are no obstacles for improving crossbred performance as well as the purebred performance, but if the r_{pc} is close to 0, selection makes no sense for crossbreds. If the r_{pc} is between 0 and 1, selection on purebred performance can also improve crossbred performance to some extent.

In this thesis, investigations focused on two-way crossbred animals. However, usually commercial pig producers use three-way crossbreeding animals instead of two-way crossbreeding animals in a terminal system. In this system, two-way F1 sows perform well in reproduction traits. They are mated to a third breed of purebred boars that perform well in production traits to reproduce the three-way crossbred pigs (Christensen et al., 2015). In Denmark, Duroc is used as boar line and sow lines are crosses between Landrace and Yorkshire. Therefore, breeding values of purebred pigs for crossbred performance in the three-way crosses need to be estimated. Christensen et al. (2015) have extended the single-step GBLUP method from two-way crossbreeding to three-way crossbreeding, but it has not been applied in real dataset yet. The model for three-way crossbreeding is a four-variate model, three for recordings from each pure breed and one for the recording from crossbreds. Similar to the two-way crossbreeding system, either partial or common genetic approach can be applied to construct the relationship matrices across the purebred and crossbred animals. When partial genetic approach applies, three breed-specific partial relationship matrices need to be specified for breed A, B and C, respectively. For instance, breed A partial relationship matrix accounts for relationships across purebred A animal, breed A gametic effects in F1 and breed A gametic effects in three-way crossbreds. In this approach, totally, 10 genetic parameters have to be estimated simultaneously and also, breed origin of crossbred alleles need to be traced, which will be a problem especially for the three-way crossbred animals. Alternatively, common genetic approach, which uses one relationship matrix across all the populations with three metafounders, can be used. Totally, 10 genetic parameters will be contained in the relationship matrix and 6 gamma parameters need to be estimated beforehand. Moreover, four-way crossbred animals where F1 sires are mated to F1 dams to reproduce F2 pigs as the terminal products are also used in the crossbreeding system. The single-step genomic evaluation approach can also be extended to four-way crossbreeding system (Christensen et al. 2015).

In this thesis, only TNB was studied. In DanAvl, the trait of litter size at day 5 after birth (LS5) is actually set as breeding goal because piglet mortality has a positive genetic correlation with TNB (Lund et al., 2012; Su et al., 2007). In practice, genomic evaluation for traits with similar genetic background usually uses a multiple-trait model because the inclusion of genetic and environmental correlations among traits may achieve higher reliabilities on EBVs. Single-step approach used in this thesis also works for multiple traits. The additive relationship matrices do not need any changes, but the number of genetic parameters increases. For instance, if there are two traits (trait 1 and trait 2, e.g: TNB and LS5) being considered in the partial genetic approach simultaneously, one of the breed-specific partial relationship matrices will be changed to a 4 by 4 matrix, as:

$$\begin{bmatrix} \sigma^2_{a_1} & \sigma_{a_1a_2} & \sigma_{a_1c_1} & \sigma_{a_1c_2} \\ \sigma_{a_2a_1} & \sigma^2_{a_2} & \sigma_{a_2c_1} & \sigma_{a_2c_2} \\ \sigma_{c_1a_1} & \sigma_{c_1a_2} & \sigma^2_{c_1} & \sigma_{c_1c_2} \\ \sigma_{c_2a_1} & \sigma_{c_2a_2} & \sigma_{c_2c_1} & \sigma^2_{c_2} \end{bmatrix},$$

where a and c are breeding values for purebred and crossbred performances, respectively. It still needs to be investigated whether this model with multiple traits performs better than the model with single trait. If no, EBVs for each trait should be estimated separately and then weighted by their economic values, being a single index value used for breeding finally.

In chapter 5, it is assumed that additive and dominance effects at each SNP are unrelated. However, mating in an inbred population will produce deviations from Hardy-Weinberg equilibrium, and these deviations generate correlations between breeding values and dominant deviations (Fernández et al., 2017). Furthermore, it has been found that magnitudes of additive and dominant genetic effects in quantitative trait loci are related by the dominance coefficients $\delta = d/|a|$, and two ways of directional relationships between additive and dominance effects, such as $\mathrm{cor}(|a|, \delta) = 0$ (BayesD2) and $\mathrm{cor}(|a|, \delta) > 0$ (BayesD3) were suggested (Wellmann and Bennewitz, 2012). Nevertheless, these relationships in magnitude cannot be fitted in individual scales in the animal model and they are not compatible with standard mixed model theory and animal breeding software. Thus, a current study is about fitting the correlation between genotypic additive and dominant effects in genomic evaluation.

In chapter 5, the single-step model was replaced by a GBLUP model because of the difficulty of extending dominance genomic relationship matrix to a combined pedigree and genomic dominance relationship matrix. This could be feasible in the future if methodological and computational problems can be overcome. With the development of new technologies (e.g: genotyping by sequencing), it will further reduce the genotyping cost, making large-scale genotyping being possible. Thus, much more genomic information from both purebred and crossbred animals is expected. Also, with the use of new phenotyping equipment and recording systems, more reliable datasets can be anticipated in the crossbreeding system in the near future (Dufrasne, 2015). As large numbers of animals have genomic information and phenotypic records, GBLUP model with non-additive genetic effects would become more attractive.

References

Bastiaansen, J. W. M., H. Bovenhuis, M. S. Lopes, F. Silva, H. J. W. C. Megens, and M. P. L. Calus. 2014. SNP Effects depend on genetic and environmental context. In: Proceedings of the 10th World congress on genetics applied to livestock production.

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

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

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

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

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

Clarke, S. M., H. M. Henry, K. G. Dodds, T. W. Jowett, T. R. Manley, R. M. Anderson, and J. C. McEwan. 2014. A high throughput single nucleotide polymorphism multiplex assay for parentage assignment in New Zealand sheep. PloS one 9(4): e93392.

Cleveland, M., and J. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. Journal of Animal Science 91: 3583-3592.

Dekkers, J. 2007. Marker-assisted selection for commercial crossbred performance. Journal of Animal Science 85: 2104-2114.

Dufrasne, M. 2015. Genetic improvement of pig sire lines for production performances in crossbreeding. PhD thesis, Université De Liège Gembloux, Belgium.

Esfandyari, H. 2016. Genomic selection for crossbred performance. PhD thesis, Wageningen University, the Netherlands.

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

Eskildsen, M., and A. V. Weber. 2016. Pig production. SEGES Publishing. Aarhus, Denmark.

Fernández, E., A. Legarra, R. Martínez, J. P. Sánchez, and M. Baselga. 2017. Estimation of covariance between dominance deviations and additive genetic effects in closed rabbit lines using an equivalent model. Journal of Animal Breeding and Genetics : Accepted.

Hayes, B., P. Bowman, A. Chamberlain, and M. Goddard. 2009. Invited review: genomic selection in dairy cattle: progress and challenges. Journal of Dairy Science 92: 433 - 443.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. van der Werf, and M. A. Cleveland. 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genetics Selection Evolution 44: 9.

Hickey, J. M., and A. Kranis. 2013. Extending long-range phasing and haplotype library imputation methods to impute genotypes on sex chromosomes. Genetics Selection Evolution 45: 4.

Hidalgo, A. M. 2015. Exploiting genomic information on purebred and crossbred pigs. PhD Thesis, Swedish University of Agricultural Sciences, Sweden.

Huang, Y., C. Maltecca, J. P. Cassady, L. J. Alexander, W. M. Snelling, and M. D. MacNeil. 2012. Effects of reduced panel, reference origin, and genetic relationship on imputation of genotypes in Hereford cattle. Journal of Animal Science 90: 4203-4208.

Ibánẽz-Escriche, N., R. Fernando, A. Toosi, and J. Dekkers. 2009. Genomic selection of purebreds for crossbred performance. Genetics Selection Evolution 41: 12.

Johnston, J., and G. Kistemaker. 2011. Success rate of imputation using different imputation approaches. Canadian Dairy Network.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92: 4656-4663.

Legarra, A., O. F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. Livestock Science 166: 54-65.

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

Lopes, M. 2016. Genomic selection for improved crossbred performance. PhD thesis, Wageningen University and research center, the Netherlands.

Lund, M. S., M. Puonti, L. Rydhmer, and J. Jensen. 2002. Relationship between litter size and perinatal and pre-weaning survival in pigs. Animal Science 74: 217-22.

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

Ma, P., R. F. Brøndum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. Journal of Dairy Science 96: 4666-4677.

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

Moghaddar, N., K. P. Gore, H. D. Daetwyler, B. J. Hayes, and J. H. Werf. 2015. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. Genetics Selection Evolution 47: 97.

Muir, W. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics 124: 342-355.

Pszczola, M., T. Strabel, H. Mulder, and M. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of Dairy Science 95: 389-400.

Sargolzaei, M., J. Chesnais, and F. Schenkel. 2011. FImpute-An efficient imputation algorithm for dairy cattle populations. Journal of Dairy Science 94: 421.

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

Su, G., M. S. Lund, and D. Sorensen. 2007. Selection for litter size at day five to improve litter size at weaning and piglet survival rate. Journal of Animal Science 85:1385–1392.

Su, G., B. Guldbrandtsen, G. P. Aamand, I. Strandén, and M. S. Lund. 2014. Genomic relationships based on X chromosome markers and accuracy of genomic predictions with and without X chromosome markers. Genetics Selection Evolution 46: 1.

Swan, A. A. 1992. Multibreed evaluation procedures. PhD thesis, University of New England, Australia.

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

Ventura, R. 2013. Accuracy of imputation to high density SNP data in multibreed beef cattle. In: Plant and Animal Genome XXI Conference

Vitezica, Z. G., L. Varona, J. M. Elsen, I. Misztal, W. Herring, and A. Legarra. 2016. Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. Genetics Selection Evolution 48: 6.

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

Wei, M. 1992. Combined crossbred and purebred selection in animal breeding. PhD thesis, Wageningen University and Research Centre, the Netherlands.

Wei, M., and H. A. M. v. d. Steen. 1991. Comparison of reciprocal recurrent selection with pure-line selection systems in animal breeding (a review). Animal Breeding Abstracts 59: 281-298.

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

Wei, M., J. H. J. Werf, and E. W. Brascamp. 1991. Relationship between purebred and crossbred parameters. 2. Genetic correlation between purebred and crossbred performance under the model with 2 loci. Journal of Animal Breeding and Genetics 108: 262-269.

Weigel, K. A., G. de Los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of Dairy Science 93: 5423-5435.

Wellmann, R., and J. Bennewitz. 2012. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genetics Research 94: 21 - 37.

CHAPTER 7: CONCLUSIONS

This thesis realized the two proposed objectives of the PhD project. It successfully applied single-step and GBLUP genomic evaluation for both purebred and crossbred performances in different scenarios with data recording and genomic information in Danish Landrace, Yorkshire and F1 crossbred pig populations, confirming the existence of a moderate and positive genetic correlation between breeding values for purebred and crossbred performances and showing the importance of combined purebred and crossbred genomic information on the reliability of predictions. It also successfully investigated the impact of non-additive genetic on improvement of genomic evaluation for crossbred performance and first reported the importance of including genomic inbreeding depressions in the genomic evaluation. This thesis can be regarded as a preliminary experiment for the large-scale applications of genomic evaluation for both purebred and crossbred performances in the breeding industry.

INDIVIDUAL TRAINING (30 ECTS minimum)		
Mandatory courses	**When/where**	**ECTs**
Welcome to EGS-ABG	October,2013 / Ethiopia	2.0
EGS-ABG Research School 1	October,2013 / Ethiopia	2.0
EGS-ABG Research School 2	February,2017 / the Netherlands	2.0
Professional Skills support courses (≥6 ECTS)		
Academic English	March 2014 / Aarhus	3.0
QGG research skill for PhD students	January 2015 / Foulum	1.0
QGG Scientific Writing course	May 2017/ Foulum	2.0
MBG first PhD conference	October 2016 / Aarhus	1.0
Advanced scientific courses (≥18 ECTS)		
Programing and computer algorithms in animal breeding with focus on genomic selection and single step GBLUP	May 2014 / US	10.0
Advanced quantitative genetics for animal breeding	June 2014 / Finland	3.0
Linear models in animal breeding (MSc course)	Jan-Mar, 2015 / Aarhus	5.0
Winter school: An introduction to Bayesian analysis and MCMC	February 2015 / Italy	2.5
A short introduction to computer intensive methods for genetic analysis	September 2016 / Foulum	4.0
Design of Genetic Improvement Programs	June 2017 / Viborg	3.0
Total credits (≥30 ECTS)		**40.5**

DISSEMINATION OF KNOWLEDGE	
International conferences (minimum of 3)	**Where/When**
10th World Congress on Genetics Applied to Livestock Production (WCGALP) – Poster	Vancouver, Canada / August 2014
65th European Association of Animal Production (EAAP) - Oral	Copenhagen, Denmark / September 2014
66th European Association of Animal Production (EAAP) - Oral	Warsaw, Poland / August 2015
68th European Association of Animal Production (EAAP) - Abstract	Tallinn, Estonia / August 2017
Seminars and workshop (minimum 1)	
2nd Genomic Selection in Animals and Plants (GenSAP) - Oral	Korsør, Denmark / October 2014
Workshop on Genomic selection in pigs - Oral	Copenhagen, Denmark / February 2016
Séminaire des doctorants du métaprogramme SelGen - Poster	Paris, France / May 2016
4th Genomic Selection in Animals and Plants (GenSAP) - Oral	Aarhus, Denmark / October 2016

Title : COMBINED PUREBRED AND CROSSBRED INFORMATION FOR GENOMIC EVALUATION IN PIG

Abstract: This PhD thesis has two aims: first, apply single-step genomic evaluation method for purebred and crossbred performances in different scenarios with data records and genotypes in Danish Landrace, Yorkshire and F1 crossbred pig populations; second, investigate the impact of non-additive genetic effects on genomic evaluation for crossbred performance. In chapter 2, performances of genotype imputation in low density SNP-panels were compared in both purebred and crossbred populations. Imputation for crossbreds worked as well as for purebreds if both parental breeds were included in the reference population. In chapter 3, the single-step GBLUP method was applied to a combined purebred and crossbred dataset, focusing on evaluating genetic ability for crossbred performance of total number of piglets born (TNB). Additive genetic effects in crossbred animals were split into two breed-specific gametic effects. The analysis confirmed the existence of a moderate, positive genetic correlation between purebred and crossbred performances for TNB. Models with genomic information, especially from crossbred animals, improved model-based reliabilities for crossbred performance of purebred boars and also improve predictive abilities on crossbred animals in a validation population. This method requires tracing the breed origin of crossbred alleles, which may be inconvenient. Therefore, in chapter 4, this dataset was reanalysed using a single relationship matrix with metafounders to relate all the involved animals in the three populations. This method did not need tracing the breed origin of crossbred alleles. Estimates of genetic parameters were similar to those in chapter 3 and the predictive abilities for crossbred performance were at least as good as in chapter 3. Both chapters 3 and 4 indicate that the single-step method for combined purebred and crossbred performances is applicable for genomic evaluation. In chapter 5, genomic evaluation using a model including dominance effects and inbreeding depression was investigated for genotyped animals in GBLUP context. The estimated correlations between allele substitution effects of markers for different breeds were reported for the first time. Results indicated that the accuracies of predictions were not improved by including dominance effects in the model, but inclusion of genomic inbreeding depression effects did improve the performance of prediction.

Keywords: Genomic evaluation, crossbred, pig

Résumé : Cette thèse de doctorat a donc deux objectifs : Premièrement, l'application de la méthode en une seule étape pour l'évaluation génomique du rendement des animaux de race pure et des hybrides dans de différentes scénarios, en appliquant des recensements de données et génotypes de Danish Landrace, Yorkshire et F1 populations de porc; Deuxièmement, l'examen du résultat des effets génétiques non-additifs sur l'évaluation génomique du rendement des hybrides. Dans le chapitre 2, les rendements sur l'imputation génotypique des panneaux à SNP [SNP-panels] à basse densité sont comparés dans des populations de pure race et d'hybrides. L'imputation d'animaux hybrides a fonctionné aussi bien que l'imputation d'animaux de pure race, si les deux races parentales sont inclues dans la population de référence. Dans le chapitre 3, la méthode GBLUP en une seule étape est appliquée sur un jeu de données combiné, incluant des animaux de pure race et des hybrides, et en focalisant sur l'évaluation de la valeur génétique au croisement pour la taille de portée. Des effets additifs et génétiques en hybrides sont divisés en deux effets gamétiques et spécifiques de la race. L'analyse confirme l'existence d'une corrélation modérée, positive et génétique entre le rendement des animaux de pure race et des hybrides en ce qui concerne la taille de portée. Les modèles avec information génomique provenant particulièrement d'hybrides améliorent la précision du modèle pour le rendement hybride de verrats de pure race, et améliorent également la précision de prédiction pour hybrides dans une population de validation. Cette méthode demande une détection de l'origine de race des allèles en hybrides, ce qui peut être un désavantage. Dans le chapitre 4, ce jeu de données est analysé en utilisant une seule matrice de parenté avec méta-fondateurs pour rélier tous les animaux impliqués dans les trois populations. Avec cette méthode, il n'est pas nécessaire de tracer l'origine de race des allèles dans des hybrides. Les estimations de paramètres génétiques correspondent à celles décrites dans le chapitre 3, et la capacité de prédiction pour le rendement des hybrides est au moins aussi bonne qu'en chapitre 3. Les chapitres 3 et 4 indiquent que la méthode en une seule étape pour le rendement combiné des animaux de pure race et des hybrides est applicable pour l'évaluation génomique. Dans le chapitre 5, l'évaluation génomique, dans laquelle on applique un modèle incluant les effets de dominance et la dépression de consanguinité, est examinée en utilisant GBLUP. Les corrélations estimées entre les effets de substitution d'allèles de marqueurs pour des races différentes sont ainsi documentées pour la première fois dans cette thèse. Les résultats indiquent que la capacité de prédiction ne s'est pas améliorée en incluant des effets de dominance dans le modèle. Par contre, l'inclusion des effets de dépression de consanguinité dans le modèle a amélioré la capacité de prédiction.

Mots-clés : évaluation génomique, hybride, porc