Linkage disequilibrium and genomic selection in pigs

Renata Veroneze

Thesis committee

Promotors

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

Prof. Dr S.E.F. Guimarães
Professor of Animal Science
Universidade Federal de Viçosa, Brazil

Co-promotor

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

Other members

Prof. Dr F.A. van Eeuwijk , Wageningen University
Dr M.V.G.B. da Silva, EMBRAPA, Brazil
Prof. Dr D.-J. de Koning, Swedish University of Agricultural Sciences, Uppsala, Sweden
Dr H.C.M. Heuven, Genetwister Technologies, Wageningen

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

Linkage disequilibrium and genomic selection in pigs

Renata Veroneze

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Wednesday 16 September 2015
at 1.30 p.m. in the Aula.

Veroneze, R.
Linkage Disequilibrium and genomic selection in pigs
142 pages.

PhD thesis, Wageningen University, Wageningen, NL (2015)
With references, with summary in English

ISBN 978-94-6257-415-1

Abstract

Veroneze, R. (2015). Linkage disequilibrium and genomic selection in pigs. PhD thesis, Wageningen University, the Netherlands

Genomic selection and genomic wide association studies (GWAS) are widely used methods that aim to exploit the linkage disequilibrium (LD) between markers and quantitative trait loci (QTL). Securing a sufficiently large set of genotypes and phenotypes can be a limiting factor when implementing genomic selection that may be overcome by combining data from multiple populations or using crossbred information. The overall objective of this thesis was to characterize LD patterns in different pig populations and to evaluate whether the differences in LD determine the accuracy of genomic predictions when using different reference sets (within-, across- and multi-population) and methodologies. In this thesis I used data from pure lines and crossbred pig populations genotyped with PorcineSNP60 BeadChip. Loess regression provided a better fit to the real LD data, and more accurate LD predictions could be made, compared to nonlinear regression. It was also shown that Loess regression can be used to statistically compare the LD decay of different populations. The persistence of LD phase between crosses and the parental pig lines was found to be high, from which it was hypothesized that similar marker-QTL associations would be found in a cross and in their purebred parent populations and therefore accuracies of genomic prediction across these populations should be high. Between the pure lines the persistence of phase was low, thus higher density panels should be used to have the same marker-QTL associations across these lines. Accuracies obtained from across- and multi-population genomic prediction and from using crossbred data did however not follow the expectations based on LD. Having the same LD phase may therefore not be as important for genomic prediction accuracy as previously thought but rather the interplay between LD, genetic architecture and allele frequencies also plays a major role. Differences in allele frequencies between lines and information from GWAS on the genetic architecture of traits for the different lines were taken into account in analyses developed in the later chapters. The use of weights, based on GWAS results, was expected to lead the GBLUP model towards the real genetic architecture of the traits. This strategy was shown to have some benefit for the genomic predictions with single- and multi-population data sets. Weights obtained from GWAS in different data sets (within and combining populations) did not always lead to increased accuracies of prediction, depending on which lines the weights are applied to. Using weights from GWAS in a combined population was the best approach, resulting in higher accuracy of GBLUP predictions within single- as well as

in multi-population analysis. Understanding and evaluating how the accuracy of within-, across- and multi-population genomic prediction is affected by differences in LD, in genetic architecture and in allele frequencies is key to optimize the accuracy of genomic prediction in pig breeding.

Contents

1

General introduction

1.1 Introduction

Genomic selection was first applied to Holstein cattle, but currently, most major breeding companies have implemented it. Although genomic selection is used in practice, its application presents some challenges. Several knowledge gaps remain to be bridged to enable the creation of practical, feasible methods for applying this new technology. Many aspects of quantitative and population genetics are revisited in the context of genomic selection. In this thesis, I report my research on linkage disequilibrium and on practical strategies and methods for the implementation of genomic selection in pigs.

1.2 Linkage Disequilibrium

Linkage disequilibrium (LD) is a non-random association between alleles at different loci (Ardlie et al., 2002). These allelic associations are mainly due to physical proximity, but they are also influenced by population history and evolutionary forces (Khatkar et al., 2008). For example, the extent of LD depends on local recombination rates. Therefore, the LD is higher in regions with low recombination rates, which for mammals, includes the Y chromosome, parts of the X chromosome, and regions near the centromere in autosomes. Conversely, the LD is lower in regions with high recombination rates, such as euchromatin and small regions known as hotspots (Jeffreys et al., 2001).

The population history, breeding system, and pattern of geographic subdivision are reflected in the LD throughout the genome. In contrast, the history of natural selection, gene conversion, mutations, and other forces that cause gene-frequency evolution can lead to differences in the LD of specific genomic regions (Slatkin, 2008).

Genomic selection and genome wide association studies (GWAS) rely on the LD between DNA markers and quantitative trait loci (QTL) to estimate genomic breeding values (GEBV) or to detect regions that control traits of interest. In evaluating how efficiently GWAS results were transferred across peoples of European and East Asian ancestries, Marigorta and Navarro (2013) suggested that a proportion of the associations found in Europeans failed to replicate in East Asians, due to the heterogeneity in LD between causal variants and tag-SNPs. In genomic selection, it has been shown that the GEBV accuracy depends at least partly on the LD between the DNA markers and the QTL (Hayes et al., 2009).

The LD in commercial pig populations extends over larger distances than in cattle. The currently, widely-used pig SNP panel (Porcine SNP60, Illumina Inc, San Diego, USA) appears to have an adequate number of DNA markers to provide a sufficient level of LD for effective GWAS and genomic selection (Uimari and Tapio, 2011;

Badke et al., 2012; Veroneze et al., 2013; Wang et al., 2013). This high level of LD also benefits the imputation of SNP genotypes (Pei et al., 2008) and opens the possibility of using low density panels in pigs.

In addition to the level of LD, the consistency of LD phase is important for genomic selection and GWAS. LD consistency means that the marker effects are consistent across generations, which is critical for genomic selection implementation. Also, the accuracy of across- and multi-population genomic predictions is influenced by the consistency of LD phase between populations. An inconsistent LD phase can explain why a marker associated with an important effect in one population may not be effective for selection in a second population.

Badke et al. (2012) evaluated the Landrace, Yorkshire, Hampshire, and Duroc pig breeds. They found that the correlations of LD phase ranged from 0.87, between Duroc and Yorkshire pigs, to 0.92, between Landrace and Yorkshire pigs, for markers with a pairwise distance <10 Kb. For markers separated by the same distances, Wang et al. (2013) found a somewhat lower persistence of phase, with correlations of 0.61 between Duroc and Landrace, 0.57 between Duroc and Yorkshire, and 0.66 between Landrace and Yorkshire pigs. Therefore, the current 60K pig marker panel may have insufficient density to maintain LD phase consistency across all pig breeds.

1.3 Modeling for linkage disequilibrium prediction

Linkage disequilibrium can be computed for each pair of loci in the genotype dataset; this procedure can generate a large amount of output data. Summarizing the data provides better comprehension of the LD patterns in different populations. To date, the relationship of LD to the physical distance between markers (LD decay) has been studied, either by calculating simple averages within predefined windows of distance (Uimari and Tapio, 2011; Badke et al., 2012; García-Gámez et al., 2012; Veroneze et al., 2013) or using parametric nonlinear regression models (Heifetz et al., 2005; Amaral et al., 2008; Abasht et al., 2009).

The most commonly used nonlinear regression model was proposed by Sved (1971). This model was developed based on the theory of genetic drift and recombination (Sved, 2009). The model derivation assumed that the population was isolated, mating was random, and the population size remained constant over time. These assumptions are likely to be violated in most natural and selective breeding populations. Moreover, nonlinear regression models generally assume that errors are independent, have homogeneous variance, and are normally distributed. These assumptions are violated, due to the nature of LD data, which is dependent on the distance between markers and is more variable at short distances than at long distances. Consequently, some alternatives have been proposed for modeling LD decay. Instead of using a fixed value of unity for the intercept, Corbin et al. (2010) introduced a new parameter to estimate the intercept in the equation proposed by Sved (1971); this new parameter may provide a better fit to the LD at short distances. LOESS regression is also a good option for describing the LD decay, because it allows the functional form between dependent and independent variables to be determined by the data without requiring strong assumptions (Andersen, 2009).

1.4 Genomic selection in pigs

The genomic selection methods proposed by Meuwissen et al. (2001) exploit the LD that exists between markers and QTLs for the estimation of GEBV. Estimating breeding values with markers can provide a reduction in the generation interval and/or an increase in the accuracy of the breeding values. In pigs, the generation interval is typically short; thus, genomic selection is expected to provide limited improvement in the annual genetic gain by reducing the generation interval. Therefore, the most important advantage of genomic selection in pig breeding is the increased accuracy it can provide.

The accuracy of GEBV depends on several factors, including the LD between the markers and the QTL, the number of animals in the reference population, the heritability of the trait, the distribution of QTL effects (Hayes et al., 2009), and the level of family relationship between the reference population and the selection candidates (Wientjes et al., 2013).

The implementation of genomic selection in pigs is more complicated than in cattle, due to the characteristics of pig breeding. Pig breeding is a pyramidal system, with a small nucleus population and short generation intervals; also, it is typically guided by diverse breeding goals (Ibáñez-Escriche et al., 2014). The small population size complicates the implementation of highly accurate genomic selection, because the accuracy of the GEBV depends on the size of the reference set. The short generation interval inherent in pig breeding removes an important benefit of genomic selection, compared to the situation in cattle. Due to the fact that more generations are produced per unit of time, pig breeding requires frequent re-estimations of marker effects. In addition, the reduction in family relationships will be accelerated. The lack of family relationships between reference and prediction animals will reduce the accuracy of GEBV.

1.5 Across- and multi-population genomic selection

In many livestock populations, the size of the reference population restricts the achievable accuracy of the GEBV. Typically, in pig breeding, more than one population or line is used to produce a viable crossbred pig. Having multiple lines in a breeding program, that each are of limited size, may however point to a solution for increasing accuracies. The size of the reference population for a given line could be doubled, or more, by using data from other lines in multi- or across-population genomic selection and with the use of crossbred information.

The use of multi- and across-population reference sets has been tested in simulation studies (Ibánez-Escriche et al., 2009; Toosi et al., 2010; Zeng et al., 2013) and in real data in cattle (Hayes et al., 2009), sheep (Legarra et al., 2014), pigs (Hidalgo et al., 2015), and chickens (Simeone et al., 2012). Simulation studies have indicated that favorable effects on accuracy could be achieved mainly by using multiple populations. In contrast, studies that use real data have shown both positive and negative outcomes. In a simulation study, de Roos et al. (2009) evaluated the effects of combining multiple populations on the accuracy of genomic selection. Those authors concluded that the greatest benefits of combining populations were achieved when the populations had diverged for only few generations, the marker density was high, and the heritability was low.

In evaluating across- and multi-population reference sets for Jersey and Holstein breeds, Hayes et al. (2009) showed that a limited relationship existed between the breeds. They found that the GEBV accuracies were low when a reference population of one breed was used to predict breeding values of the other breed. However, they found comparable or higher accuracies when multi-population reference sets were used instead of the smaller, purebred reference population. With a reference set that included 5331 Holstein, 1361 Jersey, and 506 Brown Swiss animals, Olson et al. (2012) concluded that the breeds with small reference sets gained the most GEBV accuracy by using a reference set that comprised multiple breeds. On the other hand, Legarra et al. (2014) evaluated genomic predictions for six breeds of sheep, and they concluded that the use of multiple populations only marginally increased the accuracy, and only for a few breeds.

A common outcome of those studies was that the relationships between breeds had an important influence on the accuracy of the GEBV, when using multi- or across-population reference sets. These multi- and across population approaches may provide promising opportunities for genomic prediction in the pig industry, where some lines share a common genetic background. Moreover, the pig industry aims to improve the performance of crossbred animals. Therefore, data from crossbreds could provide powerful additional information to the reference population, because the crossbreds are closely related to purebred candidates.

1.6 Aim and outline of this thesis

This thesis describes research conducted to study LD and genomic selection in pigs. I aimed to characterize LD patterns in different pig populations and to evaluate whether the consistency of LD between populations could be used to indicate the performance of genomic predictions when multiple populations were included in the prediction and/or validation datasets. In addition to LD, I investigated other differences between populations to determine whether they could explain the results achieved in different genomic selection scenarios with across- and multi-population reference datasets. I also implemented various approaches to account for differences between populations, like different allele frequencies and different genetic architectures for traits of interest. I tested these different approaches for their effects on GEBV accuracy. In Chapter 2, I evaluated the LD persistence between populations and LD decay within purebred and crossbred pigs. From those results, I investigated the potential of using crossbreds in reference panels for purebred selections, and the potential of combining pure lines in a reference panel. In Chapter 3, the well-known nonlinear model typically used to fit LD decay (Sved, 1971) was compared to an alternative, LOESS regression designed to describe LD decay. A better description of the LD decay was expected to give better predictions of how much QTL variance would be captured by SNP panels through LD. The LOESS regression was tested, because it makes fewer assumptions about residual normality, residual independence, and heterogeneity of variance, all of which are known to be violated in LD data. In Chapter 4, different reference sets (across- and multi-population) were tested for their utility in predicting GEBVs. Also, crossbred performance was tested for use in genomic prediction. Those empirical results were compared to the expectations based on the results for LD consistency described in Chapter 2. That comparison indicated that factors other than the consistency of LD affected the accuracy of genomic prediction in across- and multi-population scenarios. Therefore, in Chapter 5, a methodology that used information from GWAS was evaluated for genomic prediction accuracy. In chapter 5, the aim was to allow the genomic prediction model to use information from genetic architecture in multi-population genomic predictions. In Chapter 6, the results were placed in a broader context. There, I discuss the practical aspects of LD in breeding, effective population size estimation, the application of genomic selection in small populations, and the challenges of genomic selection, including the use of whole genome sequence data, in pig breeding.

2

Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (*Sus scrofa*) populations

Renata Veroneze[1], John WM Bastiaansen[2], Egbert F Knol[3], Simone EF Guimarães[1], Fabyano F Silva[1], Barbara Harlizius[3], Marcos S Lopes[2,3], Paulo S Lopes[1]

[1]Departamento de Zootecnia, Universidade Federal de Viçosa, Av. P.H. Holfs, 36570-000, Viçosa, MG, Brazil; [2]Animal Breeding and Genomics Centre, Wageningen University, Droevendaalsesteeg 1, Wageningen, 6708 PB, the Netherlands; [3]TOPIGS Research Center IPG B.V., P.O. Box 43, 6640 AA, Beuningen, the Netherlands.

BMC Genetics (2014) 15:126

Abstract

Genomic selection and genomic wide association studies are widely used methods that aim to exploit the linkage disequilibrium (LD) between markers and quantitative trait loci (QTL). Securing a sufficiently large set of genotypes and phenotypes can be a limiting factor that may be overcome by combining data from multiple breeds or using crossbred information. However, the estimated effect of a marker in one breed or a crossbred can only be useful for the selection of animals in another breed if there is a correspondence of the phase between the marker and the QTL across breeds. Using data of five pure pig (*Sus scrofa*) lines (SL1, SL2, SL3, DL1, DL2), one F_1 cross (DLF1) and two commercial finishing crosses (TER1 and TER2), the objectives of this study were: (i) to compare the equality of LD decay curves of different pig populations; and (ii) to evaluate the persistence of the LD phase across lines or final crosses.

Almost all of the lines presented different extents of LD, except for the SL2 and DL3, both of which exhibited the same extent of LD. Similar levels of LD over large distances were found in crossbred and pure lines. The crossbred animals (DLF1, TER1 and TER2) presented a high persistence of phase with their parental lines, suggesting that the available porcine single nucleotide polymorphism (SNP) chip should be dense enough to include markers that have the same LD phase with QTL across crossbred and parental pure lines. The persistence of phase across pure lines varied considerably between the different line comparisons; however, correlations were above 0.8 for all line comparisons when marker distances were smaller than 50 kb.

This study showed that crossbred populations could be very useful as a reference for the selection of pure lines by means of the available SNP chip panel. Here, we also pinpoint pure lines that could be combined in a *multiline* training population. However, if *multiline* reference populations are used for genomic selection, the required density of SNP panels should be higher compared with a single breed reference population.

Key words: nonlinear model, single nucleotide polymorphism, SNP, genomic selection

2.1 Introduction

Linkage disequilibrium (LD) is a nonrandom association between alleles at different loci (Ardlie et al., 2002). There has been a growing interest in LD analysis with the explosion of genomic selection (GS) and genome wide association studies (GWAS) published in recent years. Both GS and GWAS exploit the LD between markers and quantitative trait loci (QTL) to estimate genomic breeding values (GEBV) or to detect regions that control traits of interest.

The accuracy of GEBV depends on the LD between the markers and the QTL, the number of animals in the reference population, the heritability of the trait, the distribution of QTL effects (Hayes et al., 2009a) and the level of family relationship between the reference population and the selection candidates (Wientjes et al., 2013). The number of animals in the reference population is a critical parameter for the accuracy of GS (Daetwyler et al., 2008), and this value can limit the application of GS in certain situations. This constraint may be overcome by increasing the reference population size by combining animals from different breeds or lines (Hayes et al., 2009b). Daetwyler et al. (2012) showed that GEBV are more accurate than pedigree-based best linear unbiased prediction (BLUP) using a multibreed sheep training population.

Another approach that can be used to acquire a larger reference population is the inclusion of crossbred animal information, because large populations are available in commercial farms. Using crossbreds has several advantages: one crossbred population could be used to select more than one pure line, the phenotypes of production animals can be more relevant for breeders and the animals can be selected for traits that are not measured in the nucleus herd (e.g. disease resistance). In addition, using crossbred data it may be possible to account for heterotic effects in the selection. Using marker information, Amuzu-Aweh et al. (2014) showed that it was possible to identify specific sires whose offspring could be expected to show higher levels of heterosis. These approaches are especially attractive for the pig industry, where breeding companies keep a range of sire and dam lines. Using crossbred reference populations could reduce the need to establish separate large reference populations for each pure line.

To evaluate the potential for using a reference population from a different breed or cross, it is essential to know the LD in those breeds and crosses, as well as the persistence of the LD phase across these populations and with the population of selection candidates. Assuming that QTL effects are the same in different breeds, the estimated effect of a marker in one breed can still only be used to select animals in another breed if the phase of the marker and QTL alleles are the same in both breeds (Dekkers and Hospital, 2002). GS uses direct relationships and LD to predict breeding values. When predictions are carried out in populations with distantly related individuals, the accuracy is mainly determined by LD between markers and QTL, while predictions with closely related individuals rely mainly on direct relationships (Daetwyler et al., 2012). Thus, when the relatedness across breeds is small, the accuracy of prediction is mainly reflected in the LD between markers and QTL. In addition, knowledge of the persistence of phase across physical distance between markers for two populations can be used to determine which marker density is needed to provide the same LD phase across these populations (de Roos et al., 2008).

Badke et al. (2012), when evaluating the Landrace, Yorkshire, Hampshire and Duroc breeds, found that the correlation of phase ranged between 0.87 for Duroc-Yorkshire and 0.92 for Landrace-Yorkshire, for markers with a pairwise distance <10 Kb. While, for the same distance, Wang et al. (2013a) found a persistence of phase of 0.61 for Duroc-Landrace, 0.57 for Duroc-Yorkshire and 0.66 for Landrace-Yorkshire. Studies evaluating LD and persistence of phase in crossbred pig lines are scarce, and the comparison of LD decay in different populations has been achieved visually using average LD (de Roos et al., 2008; Uimari and Tapio, 2011; Badke et al., 2012; Veroneze et al., 2013; Wang et al., 2013a), without the application of models or statistical comparisons.

In the present study, we evaluated five pig pure lines (SL1, SL2, SL3, DL1, DL2), one F_1 cross (DLF1) and two commercial finishing crosses (TER1 and TER2) representing the crossbred structure of pork production. The objectives of this study were: (i) to compare the equality of LD decay curves of different populations; and (ii) to evaluate the persistence of phase across populations.

2.2 Methods

This experiment was conducted strictly in line with the Dutch law on the protection of animals.

2.2.1 Data

The data for this study were obtained from animals from five pig pure lines (SL1, n=1,307; SL2, n=643; SL3, n=276; DL1, n=626; DL2, n=1013), one F1 cross (DLF1, n=186) and two commercial finishing crosses (TER1, n=286; TER2, n=330). SL1 and SL2 are synthetic sire lines; SL1 is a combination of Duroc (mostly) and Belgian Landrace created in about 1980. SL2 is a combination of Large White and Pietrain created in about 1975. SL3 is a Pietrain sire line. DL1 is Landrace based dam line and DL2 is a Large White based dam line. DLF1 is a commercial F1 cross resulting from crossing animals of DL1 and DL2. TER1 is a commercial finishing pig resulting from a cross between DLF1 and SL1. TER2 is also a commercial finishing pig that resulted from a cross between DLF1 and SL2. All pure lines were kept under strict inbreeding restrictions, with approximately 40 replacement sires per year and more than 250 gilt replacements per year.

Animals were genotyped using the Illumina Porcine SNP60 Beadchip, and all SNPs with an undefined position in Build 10.2 (Groenen et al., 2012) were excluded, as well the SNPs on the X chromosome. The X chromosome recombines only in females; therefore, it was expected that the X chromosome would show higher LD than the overall genome (Schaffner, 2004), which could cause an overestimation of the LD. The R software (http://www.r-project.org/) was used for within population marker quality control, using the package GenABLE (Aulchenko et al., 2007). Markers with a call rate <90%, MAF <0.05 and/or a p-value for the Hardy-Weinberg equilibrium <0.0001 were excluded. The summary of the quality control of genotype data is presented in supplementary material (Table S2.1).

To estimate the persistence of phase, the data were divided into four groups, according the description shown in supplementary material (Table S2.2), and only SNPs that passed the quality control in all lines of each group were used. In group 1, the F1 (DLF1) cross was compared with its parental lines, while in groups 2 and 3 the finishing crosses (TER1 and TER2) were compared with their parental and grandparental lines. In group 4, which included only pure lines, each line was compared with all other pure lines.

2.2.2 LD

For each pig line, the LD between SNPs was computed as the correlation of gene frequencies (r_{ij}^2) (Hill and Robertson, 1968) using the function LD of the package genetics (Warnes and Leisch, 2005) of the software R (http://www.r-project.org/):

$$r_{ij}^2 = \frac{\left(p_{ij} - p_i p_j\right)^2}{p_i\left(1-p_i\right)p_j\left(1-p_j\right)}$$

where p_i and p_j are the marginal allelic frequencies at the i^{th} and j^{th} SNP, respectively, and p_{ij} is the probability of the marker allele pair ij , which is estimated using maximum likelihood because genotype data were used (Warnes and Leisch, 2005).

2.2.3 LD decay

Decay of LD with the distance between markers was compared between lines. Only SNPs that passed the quality control filtering in all lines were used in this analysis.

The comparison was conducted by adjusting the nonlinear regression model proposed by Sved (1971) to allow for testing a curve equality hypothesis (Bates and Watts, 1988) across the eight populations evaluated. For the curve equality test, the nonlinear model receives a dummy variable that represents each one of the eight populations. This complete model is described as:

$$LD_{ik} = \sum_{k=1}^{8} D_k \left[\frac{1}{1+4\beta_k d_i} \right] + e_{ik}, \qquad (1)$$

where:

LD_{ik} is the observed r_{ij}^2 for marker pair i of line k ;

D_k is an dummy variable, such that :

$$D_k = \begin{cases} 1 \text{ if the observation } LD_{ik} \text{ belong to the group } k \\ 0 \text{ otherwise} \end{cases}$$

d_i is the distance in Kb for marker pair i ;

β_k is the coefficient that describes the decline of LD with distance for line k ;

e_{ik} is a random residual, $e_{ik} \sim N(0,\sigma^2)$;

The complete model is adjusted to test the hypothesis that the same model can describe the LD decay of all lines:

$H_0^{(1)} : \beta_k = \beta \; \forall \; k$ vs $H_a^{(1)} : \beta_k \neq \beta$ for at least one β_k

To test the $H_0^{(1)}$ hypothesis, the following comparison scheme was conducted, considering the complete (1) and the reduced (2) models:

$$LD_{ik} = \sum_{k=1}^{8} D_k \left[\frac{1}{1+4\beta d_i} \right] + e_{ik}, \qquad (2)$$

where a single parameter β for all lines is assumed.

The residual sum of squares of the complete (SQR_Ω) and reduced (SQR_ω) models are used to perform a chi-squared statistic: $\chi^2_{computed} = N\ln(SQR_\Omega / SQR_\omega)$, in which N is the number of observed measures of LD. The hypothesis $H_0^{(1)}$ is rejected if $\chi^2_{computed} = \chi^2_{\alpha(v)}$, where $v = p_\Omega - p_\omega$ is the degree of freedom, where p_Ω and p_ω are the number of parameters of the complete and reduced models, respectively, at a significance level α .

Rejection of the hypothesis $H_0^{(1)}$ implied that at least one parameter β differs from the others, and, subsequently, a pairwise comparison was carried out to identify the lines that are equal or different in relation to the parameter β . Multiple tests were carried out; therefore, the Bonferroni correction was employed to reduce Type I errors. In this case, the significance threshold ($\alpha*$) was obtained by dividing the established significance threshold for a single test ($\alpha = 0.05$) by the number of independent tests (n). Thus, for the present study, the significance level for pairwise comparison was $\alpha* = 0.05/28 = 0.0018$.

The nonlinear models were adjusted using the function *nls* of the software R (http://www.r-project.org/), and the hypothesis tests were also conducted using R scripts.

2.2.4 Persistence of phase

The squared root of r_{ij}^2 was obtained and given the same sign as D, which was calculated as described by de Roos et al. (2008), using the R software (http://www.r-project.org/).

$$D = f_{22} - (f_{12} + f_{22})(f_{21} + f_{22})$$

where:

$$f_{22} = (2p_{A_{22}B_{22}} + p_{A_{22}B_{12}} + p_{A_{12}B_{22}})/\tau$$

$$f_{12} = (2p_{A_{11}B_{22}} + p_{A_{11}B_{12}} + p_{A_{12}B_{22}})/\tau$$

$$f_{21} = (2p_{A_{22}B_{11}} + p_{A_{22}B_{12}} + p_{A_{12}B_{11}})/\tau$$

$$\tau = 2 - 2p_{A_{12}B_{12}}$$

where $p_{A_{12}B_{12}}$ is the proportion of animals with heterozygous genotypes at both loci.

This approach was first described by Goddard et al. (2006), and the setting of the D sign was conducted to consistently define the statistic in all lines. The r_{ij}^2 received the same sign in two breeds if the same haplotype was more common than expected from the allele frequencies in both breeds.

To express the correlation of r_{ij}^2 across populations in relation to the physical distances between SNPs, the Pearson correlations between r_{ij}^2 values were calculated across lines for intervals of 50 kb (from 0 to 5000 kb). The interval of 50 kb was chosen based on the coefficient of variation (CV) of the number of SNP pairs for intervals of 10, 30, 50, 70 and 100 kb [see supplementary material: Table S2.3] to guarantee that the most similar number of observations in each bin were used to calculate the correlation. Based on the CV evaluation, there was no evidence of difference in the use of bins of 30, 50, 70 and 100 kb; thus the value of 50 Kb was chosen to give a more detailed LD description in relation to the bins of 70 and 100 kb, and a better visualization in relation to the bin of 30 kb.

2.3 Results

2.3.1 LD decay

The nonlinear model for the decay of LD with distance was adjusted to simultaneously describe multiple lines. The model parameter β_k describes the decline of LD with distance for each line. The estimates of $\hat{\beta}_k$ ranged from 1.25×10^{-3} to 2.92×10^{-3} and were all significantly different from zero (p-value < 0.01) (Table 2.1).

Table 2.1 Parameter estimate ($\hat{\beta}_k$), standard error and p-value for the nonlinear fitted model for each line.

Line	$\hat{\beta}_k$	Std. Error	p-value
SL1	1.78×10^{-3}	4.76×10^{-6}	$<10^{-3}$
SL2	1.25×10^{-3}	2.89×10^{-6}	$<10^{-3}$
SL3	1.69×10^{-3}	4.42×10^{-6}	$<10^{-3}$
DL1	2.12×10^{-3}	6.09×10^{-6}	$<10^{-3}$
DL2	1.71×10^{-3}	4.49×10^{-6}	$<10^{-3}$
DLF1	2.44×10^{-3}	7.46×10^{-6}	$<10^{-3}$
TER1	2.92×10^{-3}	9.63×10^{-6}	$<10^{-3}$
TER2	2.03×10^{-3}	5.75×10^{-6}	$<10^{-3}$

The adjusted model to describe the LD permits a statistical comparison of the lines with respect to the decline of LD with distance, which is important to infer the size of single nucleotide polymorphism (SNP) panels for GS and GWAS in these lines. To compare the lines, the equality of the LD curves was tested. The first hypothesis tested ($H_0^{(1)} : \beta_k = \beta \, \forall \, \mathrm{k}$) states that the model to describe the LD decay is the same for all lines. This hypothesis was rejected (p-value $<10^{-3}$), which implies that at least one parameter β differs from the other parameters. Next, a pairwise comparison was carried out that aimed to identify which lines are equal or different regarding the parameter β . All pairwise comparisons were significantly different [see supplementary material: Table S2.4], with the exception of the comparison between β_{SL3} and β_{DL2} (p-value 0.0117 > Bonferroni corrected significance α^*). These results suggested that the same model could be used to describe the LD decay of these two lines. In addition, SL2 showed the smallest β

value, which implied that this line has the largest extent of LD, while TER1 showed the largest β value and consequently the shortest LD.

The test of the equality of the LD decay curves showed that the overall pattern of LD decay differed between lines. The predicted LD was reported at specific marker distances (Table 2.2), with the highest values of predicted LD observed for SL2 at various distances, while TER1 presented the lowest values. SL3 and DL2 presented the same values of predicted LD, because the β parameters of these lines did not differ statistically. All lines presented low values of LD for marker distances above 3000 kb. At these large marker distances, the crossbreds exhibited similar levels of LD compared with the pure lines.

Table 2.2 Predicted r^2 at various distances (Kb) for eight pig populations.

Distance (Kb)	50	250	500	1000	2000	3000
SL1	0.74	0.36	0.22	0.12	0.07	0.04
SL2	0.80	0.44	0.28	0.17	0.09	0.06
SL3	0.75	0.37	0.23	0.13	0.07	0.05
DL1	0.70	0.32	0.19	0.11	0.06	0.04
DL2	0.75	0.37	0.23	0.13	0.07	0.05
DLF1	0.67	0.29	0.17	0.09	0.05	0.03
TER1	0.63	0.26	0.15	0.08	0.04	0.03
TER2	0.71	0.33	0.20	0.11	0.06	0.04

Most of the studies on LD presented the average r^2 at various distances to compare populations. To facilitate comparison with other studies and also to make a comparison with the predicted LD, the average and standard deviation of LD at various distances are shown in Table 2.3. The standard deviation of r^2 tended to decrease when the distance between markers increased in all lines, which is expected, because at short distances the r^2 values are much more variable. The average LD for markers less than 50 Kb apart ranged from 0.55 for SL2 to 0.46 for TER1, both of which are smaller than the predicted LD at the same marker distance. Similar to the predicted LD, SL2 presented the highest values of average LD at various distances, thus showing the same tendency for predicted and average values. However, the predicted LD was higher than the average for short distances (>50 Kb) and smaller for the largest distances (3000–3050 Kb) for all lines.

Table 2.3 Average and standard deviation r² at various distances (Kb) for eight pig populations.

	0–50	200–250	500–550	1000–1050	2000–2050	3000–3050
SL1	0.49 ± 0.37	0.30 ± 0.31	0.23 ± 0.27	0.18 ± 0.23	0.12 ± 0.19	0.10 ± 0.16
SL2	0.55 ± 0.37	0.35 ± 0.33	0.28 ± 0.30	0.21 ± 0.25	0.14 ± 0.21	0.11 ± 0.18
SL3	0.50 ± 0.37	0.29 ± 0.30	0.24 ± 0.27	0.18 ± 0.23	0.13 ± 0.19	0.10 ± 0.17
DL1	0.49 ± 0.36	0.29 ± 0.30	0.21 ± 0.26	0.16 ± 0.22	0.11 ± 0.18	0.09 ± 0.16
DL2	0.51 ± 0.37	0.31 ± 0.31	0.24 ± 0.27	0.18 ± 0.24	0.12 ± 0.19	0.09 ± 0.16
DLF1	0.47 ± 0.36	0.27 ± 0.29	0.20 ± 0.24	0.15 ± 0.21	0.10 ± 0.16	0.08 ± 0.14
TER1	0.46 ± 0.35	0.25 ± 0.28	0.18 ± 0.23	0.14 ± 0.19	0.09 ± 0.15	0.07 ± 0.13
TER2	0.50 ± 0.35	0.29 ± 0.29	0.22 ± 0.26	0.16 ± 0.22	0.11 ± 0.17	0.08 ± 0.15

2.3.2 Persistence of linkage disequilibrium phase

In pig production, crossbred animals are used for reproduction on commercial farms. The line DLF1 represents these crossbred females, and crossing the dam lines DL1 and DL2 produces these animals. DLF1 presented a similar LD compared with lines DL1 and DL2, with high persistence of phase, a correlation of >0.9 for marker distances up to 150 Kb and a correlation of >0.8 for marker distances up to 1200 Kb.

Commercial finishing pigs TER1 and TER2 are the end product of the pig industry, and are based on a cross between DLF1 and either SL1 or SL2, respectively. TER1 showed higher persistence of phase with SL1 and DLF1 compared with lines DL1 and DL2 (Figure 2.1b); this result was expected because the haplotype sharing is different between TER1 and these four populations. TER1 showed a correlation of phase of >0.9 for markers at distances below 200 Kb in relation to lines SL1, DLF1, DL1 and DL2 (Figure 2.1b).

Similar to TER1, TER2 showed greater persistence of phase with SL2 and DLF1 compared with lines DL1 and DL2 (Figure 2.1c). The distance at which the correlation of phase remained >0.9 was higher for TER2 compared with TER1, with distances of 1050 Kb, 400 Kb, 150 Kb and 50 Kb in relation to the lines SL2, DLF1, DL1 and DL2, respectively (Figure 2.1c).

Interestingly, for TER1, a higher persistence of phase was observed with DL1 than with DL2 (Figure 1b), but the reverse was observed for TER2, with a higher persistence of phase with DL2 than with DL1 (Figure 2.1c). These results can be explained by the contributions of different breeds to the different lines. SL1 and DL1 have contributions from the Landrace breed, while SL2 and DL2 have contributions from the Large White breed.

Persistence of phase across pure lines was evaluated to provide information towards the use of a multiline reference population for GS. The highest persistence of phase was observed between SL2 and SL3, and between SL2 and DL2, which exhibited a correlation of >0.9 for markers at distances up to 50 kb, and the persistence remained high at larger distances (Figure 2.1d). The lowest correlation was observed between SL1 and SL2 (0.81) for markers at distances up to 50 kb. Persistence of phase showed a considerable variation between the different line comparisons; however, correlations were above 0.8 for all line comparisons when marker distances were smaller than 50 kb. Common breeds in the line genetic background resulted in a higher persistence of phase. For multiline reference populations, a SNP panel denser than the currently available is necessary to keep the same phase across pure lines.

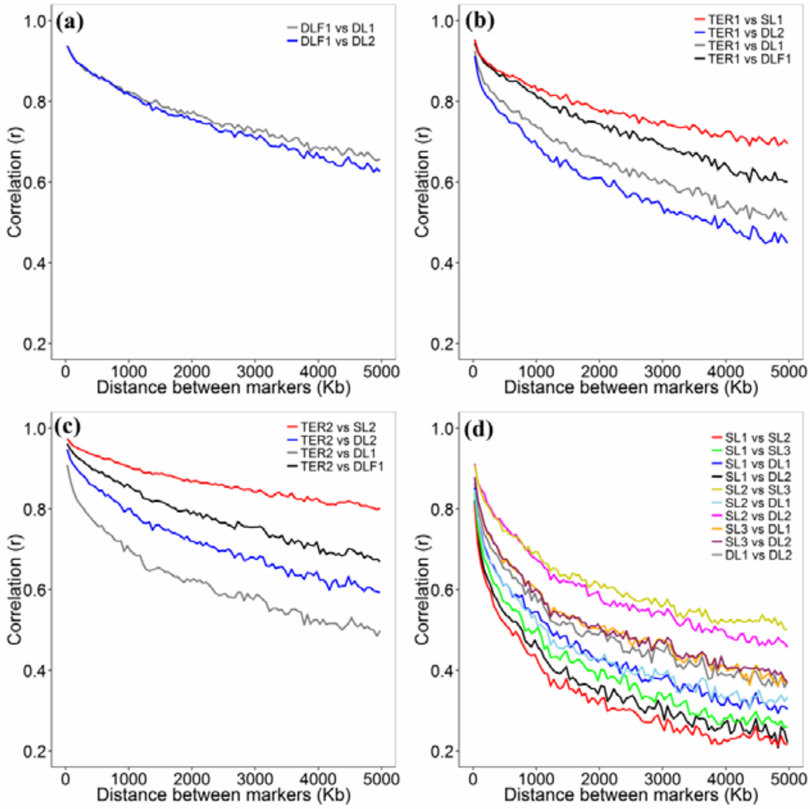

Figure 2.1 Correlation of phase (r_{ij}) in relation to the distance. a. Correlation between F1 (DLF1) and its parental lines (DL1 and DL2). b. Correlation between terminal cross (TER1) and its (grand)parental lines (SL1, DLF1, DL1 and DL2). c. Correlation between terminal cross (TER2) and its (grand)parental lines (SL2, DLF1, DL1 and DL2). d. Correlation across all pure lines (SL1, SL2, SL3, DL1, DL2).

2.4 Discussion

Using the equality of curves test, the LD decay was found to differ significantly for all except one of the pairwise comparisons between the pig lines. Persistence of phase was found to be highest between pure lines, especially for short distances below 50 kb. The persistence of LD between crossbreds and their (grand)parental lines followed the expectations based on the contributions that the different breeds made to each of the lines.

2.4.1 Equality of LD curves

A formal comparison of the level of LD decay was made possible by our adjustments to the nonlinear model described by Sved (1971). All of the lines studied followed the same pattern of a rapid decrease in LD as the distance increased. Previous comparisons of LD decay between breeds or lines was performed using average r² in distance bins (de Roos et al., 2008; Uimari and Tapio, 2011; Badke et al., 2012; Veroneze et al., 2013) and/or adjusting a linear model to test the breed effect (Amaral et al., 2008; Megens et al., 2009). The equality of curves test permits not only the identification of the existence of line differences, but also allows for a pairwise comparison across all lines. The test revealed that six of the eight evaluated lines differ with respect to LD decay. Only the comparison between $\hat{\beta}_{SL3}$ and $\hat{\beta}_{DL2}$ was not rejected, which implied that the decrease in LD with the distance is the same for these two lines. The extent of LD provides an insight into the number of SNPs required for GS and GWAS. Lines SL3 and DL2 presented the same predicted LD; therefore, an identical marker density could be used for genomic studies in both lines. However, this does not imply that the same marker set is suitable in both lines, because different markers may be segregating in different lines. The test also revealed different extents of LD for six of the evaluated lines, with a higher LD observed for SL2 and a lower LD for TER1. This information implied that different marker densities should be used for GS and GWAS for these lines, which could also influence the accuracy of GS.

2.4.2 LD in crossbreds

According to Reich et al. (2001), the extent of LD depends on the number of generations that have passed since the occurrence of an LD-generating event. In crossbred populations, LD comprises the existing LD in the parent populations and new LD generated in the cross as a result of different allele frequencies in the parental breeds (Toosi et al., 2010). The average LD for markers at distances up to 50 kb ranged from 0.47 to 0.50 in crossbreds and from 0.49 to 0.55 in the pure lines, while for markers at distances between 3000 and 3050 Kb, the LD ranged from 0.07 to 0.08 in crossbreds and from 0.09 to 0.11 in the pure lines. Surprisingly, the LD over large distances was not higher in crossbreds. A possible explanation for these similar LD levels in crossbred and pure lines may be the similarities in allele frequencies, or in LD phase, between the (grand)parental lines of the crossbreds. With similar frequencies, limited LD is created because of crossing (Toosi et al., 2010). Similarity in the allele frequencies could be caused by the fact that the minor allele frequency (MAF) was one of the criteria used to select markers for the 60K beadchip, which may have reduced the differences in allele frequency across lines (Ramos et al., 2009).

2.4.3 LD in pigs from the literature

By evaluating LD in Finnish Landrace and Finnish Yorkshire pigs, Uimari and Tapio (2011) found an average r^2 of 0.47 and 0.49 for markers 30 kb apart, and these results are similar to our findings for DL1 (Landrace based line) and DL2 (Large White based line). In addition, Uimari and Tapio (2011) reported r^2 values of 0.09 and 0.12 for SNPs that were 5 Mb apart in Finnish Landrace and Finnish Yorkshire pigs, respectively, which is higher than the average r^2 of 0.06 for DL1 and DL2 found in the present study.

By studying Duroc, Hampshire, Landrace and Large White from the USA, Badke et al. (2012) detected average r^2 values of 0.26, 0.25, 0.19 and 0.21 for SNPs that were 500 Kb apart, respectively, while in the present work, the lines SL1 (a combination of Duroc and Belgian Landrace), DL1 and DL2 presented average LDs of 0.23, 0.21 and 0.24, respectively. The differences regarding Duroc and SL1 could be explained by the breed composition of SL1, which contains Landrace genes, while differences in population structure, such as inbreeding and effective population size, could explain the LD differences of the Landrace and Large White breeds evaluated by Badke et al.(2012) and between DL1 and DL2. At large distances (5 Mb), the LD levels were similar to those found by Badke et al. (2012).

Evaluating the LD in Danish Landrace, Large White and Duroc, Wang *et al.* (2013a) found average LDs of 0.32, 0.32 and 0.35 for markers at a distance of 500 Kb, respectively, and these values are much higher than the values found in the present paper for DL1, DL2 and SL1 (0.21, 0.24 and 0.23, respectively). Parameters that are specific for a population, such as the inbreeding, effective population size and selection, can also result in different LD levels across populations. Studying the LD of local Spanish and Portuguese pig breeds and of wild pig populations, Herrero-Medrano et al. (2013) found that the decay of LD was greater in wild boars than in the domestic breeds. Evaluating the LD of Chinese and Western pigs, Ai et al. (2013) found that Chinese breeds have lower extents of LD than Western pigs.

2.4.4 Implications for GS

An average LD greater than 0.2 has been reported to be required for GS (Meuwissen et al., 2001), and this LD level was observed for most of the evaluated lines at marker distances between 500 and 550 kb. All lines exhibited an average r^2 higher than 0.3 for markers 100–150 kb apart. Qanbari et al. (2010) found an average $r^2 = 0.30$ for markers at distances <25 kb for German Holstein cattle, and Bohmanova et al. (2010) found $r^2 > 0.3$ for markers at distances of 60 kb in American Holstein cattle. Thus, in agreement with Veroneze et al. (2013) and Badke et al. (2012), it seems that LD extends further in European commercial pig breeds than in Holstein cattle, which implies that the use of less dense SNP panels is possible for GWAS and GS in pigs. Evaluating the use of low density panels associated with genotype imputation in pig sire lines, Wellmann et al. (2013) recommended that a panel with 384 markers could be used for genotyping selection candidates if at least one parent was genotyped at high-density. However, if multibreed reference populations are used for GS, the required density of SNP panels should be higher compared with a single breed reference population. Persistence of phase is essential for the success of across lines GS. In the present paper, the persistence of LD phase was evaluated for eight commercial pig populations, thus representing the crossbreeding structure of pig production design.

The high persistence of phase for SNPs with a 150 Kb distance when comparing DLF1, DL1 and DL2 implies that similar marker effects may be expected across the evaluated lines. The available porcine SNP chip should be dense enough to include markers that have the same LD phase with QTL across DLF1, DL1 and DL2. The persistence of phase with DLF1 shows the potential use of an F1 commercial cross as a reference population to select purebred lines. However, using pig purebreds to predict crossbred performance, Hidalgo (personal communication) found that the accuracies of the breeding values are trait-dependent, which challenges the use of the crossbred information in breeding programs.

In a simulation study using a crossbred (F1) as the reference population to select purebred animals, Toosi et al. (2010) found that using 10 markers per cM (a density approximately equal to the present work) resulted in an accuracy of GEBV of 0.78, while training in the same breed as the validation population resulted in a accuracy of 0.83. The authors concluded that crossbreds could be used to select purebreds without significant loss of accuracy. Crossbred animals can also be used as a source of information for genotype imputation, because of the high persistence of phase. Evaluating multi-breed imputations in Canadian dairy cattle breeds, Larmer et al. (2014) found that multi-breed populations resulted in increased imputation accuracy for the breeds Guernsey and Ayrshire, where consistency of gametic phase was high.

Using crossbred animals in the reference population is expected to have a number of advantages. First, the utilization of crossbred performance to select purebreds enables selection for traits that cannot be measured at nucleus farms, such as disease resistance (Ibánez-Escriche et al., 2009). Second, a crossbred reference population may allow for reduced costs of GS when the same crossbred performance can be used as information for selection in two or more pure lines. Third, the use of crossbreds permits exploitation of the heterotic effects, which cannot be done when the selection is performed exclusively in purebreds. However, for the use of crossbred information, pig breeding programs need to adapt their data collection to obtain the phenotypes of F1 sows and finishing pigs, which can be challenging, because these animals are held on commercial farms.

2.4.5 Interpretation of the correlations between lines

The higher correlation of the LD phase in TER1 with SL1 and DLF1 compared with the correlation of TER1 with DL1 and DL2 was expected because the persistence of the LD phase tended to decrease when a smaller proportion of the genome is shared. TER1 shares 50% with both SL1 and DLF1, and only 25% with both DL1 and DL2. TER2 showed the same tendencies, showing a higher correlation with its parent lines, SL2 and DLF1, than with its grandparent lines, DL1 and DL2. The correlation of phase with TER2 was higher over much longer distances between markers compared with TER1. Correlations above 0.9 were observed for the LD between markers at distances up to 1050 kb and 400 kb when comparing TER2 with SL2 and DLF1. Our assumption was that the higher persistence of phase of TER2 with its paternal line SL2 is caused by the higher LD observed in SL2.

With the marker density provided by the pig 60K SNP panel, the data from TER1 could be used in GS strategies for SL1. Similarly, the data from TER2 could be used to select in SL2. The 60K SNP panel provides a marker density that shows a high persistence of LD phase between these lines. A much higher marker density would be necessary to ensure a persistence of phase between the lines TER1 and TER2 and between the dam lines DL1 and DL2.

While the correlations between crossbreds and their parental lines should allow for GS with a crossbred reference population using the SNP60Beadchip, the question remains whether the correlation of phase between pure lines is also high enough for a multibreed reference population design. The persistence of phase between pure lines depends on the time since their divergence took place (de Roos et al., 2008); i.e., the consistency of LD is directly related to the degree of relationship between lines (Andreescu et al., 2007). The highest persistence of phase was observed between SL2 *vs*. SL3 and SL2 *vs*. DL2. As described in the material and methods section, SL2 is a synthetic line resulting from the combination of the Large White and Pietrain breeds. SL3 is a Pietrain pure sire line and DL2 is a Large White pure dam line. Thus, the higher persistence of phase observed between SL2 *vs*. SL3 and SL2 *vs*. DL2 could be explained by the common breeds in the composition of these lines.

The persistence of phase of Duroc, Hampshire, Landrace and Large White breeds was studied by Badke et al. (2012). A correlation of phase of 0.92 was found between the breeds Landrace and Large White for markers at distances of 10 kb, which is similar to the correlation observed between the lines DL1 and DL2 (which are Landrace- and Large White-derived lines, respectively) for markers at the same distance (0.93). The persistences of phase between SL1 *vs*. DL1 and SL1 *vs*. DL2 were higher (0.92 and 0.90, respectively) than the values found by Badke et al. (2012) between Duroc *vs*. Large White and Duroc *vs*. Landrace (0.87 for both) for markers at distances of 10 kb. Some difference was expected, because SL1 is a synthetic line of Duroc (mostly) and Landrace, so the highest persistence of phase in relation to the study of Badke et al. (2012) could be caused by the presence of the Landrace breed in SL1.

The lowest correlations of phase were observed between all lines and SL1. By evaluating the persistence of phase in Landrace, Large White and Duroc, Wang et al. (2013b) found a closer relationship between Landrace and Yorkshire and a more distant relationship between Duroc and Landrace/Large. By studying genetic diversity in native and commercial pig breeds in Portugal, including Duroc, Landrace, Large White and Pietrain, Vicente et al. (2008) concluded that Duroc is the more distant breed relative to the others. This could explain why the lowest correlations were observed between SL1 and the other lines.

Reference populations must be large for accurate prediction in GEBV, and the use of a combined reference population would be desirable. However, the correlation of phase across pure lines was low, suggesting the need for a SNP panel with a higher density than the 60K panel, even when combining SL2 and SL3 or SL2 and DL2, which presented the highest correlation of phase across the pure lines (>0.9 for markers at distances up to 50 Kb).

The utilization of multibreed reference panels has been studied as a method to increase the reference population size (de Roos et al., 2008; Hayes et al., 2009; Daetwyler et al., 2012). Hayes et al. (Hayes et al., 2009) indicated that multi-breed reference populations will be a valuable resource to fine mapping of QTL. de Roos et al. (2008) concluded that multi-breed reference panels could increase the reliability of the GEBV when at least some animals of the target breed are included, and the benefit of combining populations increased when the populations have diverged for fewer generations. In addition, Daetwyler et al. (2012) showed that GEBV are more accurate than pedigree-based BLUP, using a multibreed sheep training population. According to Daetwyler et al. (2012), across breed accuracy depends on the LD between markers and QTL because the impact of the relatedness between the breeds is expected to be minimal. Thus, persistence of phase studies provide information for shaping multibreed, or in the case of the pig industry, *multiline* reference panels. Knowing the persistence of phase allows us to identify the lines that have diverged more recently and would provide higher relationship between reference and validation populations, a factor that plays a large role in the accuracy of the predictions.

2.5 Conclusions

This work evaluated the persistence of LD and LD decay of pure and crossbred pig lines using real data, and by representing the crossbreeding structure of pig production. Our data demonstrated the potential of crossbreds as reference panels for purebred selection and also pinpointed the pure lines that could be combined in a multiline training population. This study proposed an equality of LD decay curves to evaluate significant differences regarding LD decay. Useful LD (>0.3) seems to extend over larger distances in pigs than in Holstein cattle, which implied that less dense SNP panels are needed in GS and GWAS in pigs. However, if multiline reference populations are used for GS, the required density of SNP panels should be higher compared with a single breed reference population.

3

Comparison of nonlinear and loess regression models for prediction of linkage disequilibrium decay curves

Renata Veroneze[1,2], John WM Bastiaansen[2], Paulo S Lopes[1], Egbert F Knol[3], Naomi Duijvestein[3], Marcos S Lopes[2,3], Simone EF Guimarães[1], Fabyano F Silva[1]

[1]Departamento de Zootecnia, Universidade Federal de Viçosa, Av. P.H. Holfs, 36570-000, Viçosa, MG, Brazil; [2]Animal Breeding and Genomics Centre, Wageningen University, Droevendaalsesteeg 1, Wageningen, 6708 PB, the Netherlands; [3]Topigs Norsvin, P.O. Box 43, 6640 AA, Beuningen, the Netherlands

Genetics Selection Evolution (Submitted)

Abstract

Knowledge about the relationship between linkage disequilibrium (LD) and physical distance can be used to infer the number of markers required to achieve a certain level of LD, which is useful for customization of SNP chips. Nonlinear and loess regression can be used to describe the relationship of LD with physical distance between markers; however, the impact of the regression models on LD predictions has not been investigated. Moreover, comparison of LD decay between different populations has been performed empirically without the application of a hypothesis test to determine whether curves differ significantly. Thus, proposals of comparison tests arise as a relevant point to be exploited in the field of statistical genomics. The objective of this study was to compare the nonlinear and loess regression models to describe LD decay and evaluate the impact of the estimation method on LD predictions and application of hypothesis tests for equality of LD curves.

The comparison of regression methods to describe LD decay showed that loess regression provided a better fit than did nonlinear models because loess suffered less from the lack of normality, heterogeneity of variance and residual dependence. However, when the LD decay of two populations was compared, the same result was found using either a test for equality of nonlinear curves or a nonparametric ANOVA-type statistic because the LD decay of lines SL1 and DL2 were not significantly different. The predicted number of markers to achieve an average r_{ij}^2 >0.3 between flanking SNPs (which has been recommended for genomic selection and genome wide association studies) differed widely between nonlinear and loess regression (11,667 and 62,222, respectively). The prediction of the nonlinear model was found to be an underestimate.

Loess regression is less influenced by the lack of residual normality, residual dependence and heterogeneity of variance than were nonlinear models when fitting LD decay curves. Moreover, the loess fit provides more reliable LD predictions that are more appropriate for the design of customized SNP chips than are the nonlinear models.

3.1 Introduction

Interest in the description of linkage disequilibrium (LD) has increased due to its importance in genomic selection (GS) (Goddard and Hayes, 2007), genome wide association studies (GWAS) (Corbin et al., 2010) and its contribution to better understanding the evolutionary history of a population (Slatkin, 2008). Additionally, LD information can be used to customize SNP chips because it can indicate the number of SNPs required to achieve a certain average level of LD (Carlson et al., 2004; de Roos et al., 2008; Veroneze et al., 2013).

To date, the relationship of LD with physical distance between markers (LD decay) has been studied using either simple averages in predefined windows of distance (Uimari and Tapio, 2011; Badke et al., 2012; García-Gámez et al., 2012; Veroneze et al., 2013) or parametric nonlinear regression models (Heifetz et al., 2005; Amaral et al., 2008; Abasht et al., 2009; Wang et al., 2013).

The most commonly used nonlinear regression model was proposed by Sved (1971). This model applies to an isolated population with random mating and constant population size, which are assumptions that are not fulfilled by most current livestock populations. Moreover, nonlinear regression models assume that errors are independent with homogeneous variance and normal distribution. These assumptions are violated due to the nature of LD data, which is dependent on the distance between markers and is more variable at short distances.

Loess regression (Cleveland, 1979) is a nonparametric regression model that fits smooth curves and is often used to provide a graphical view of the relationship between variables. Loess regression is also characterized as a flexible method that provides predictions of dependent variables without requiring the establishment of a functional relationship with an independent variable. In other words, this method allows the functional form between dependent and independent variables to be determined by the data without requiring strong assumptions (Andersen, 2009); therefore, it is a good alternative to describe LD decay curves.

Although the nonlinear and loess regression models can both be used to describe the decay of LD as a function of physical distance, the impact of these regression models on LD predictions has not been investigated. Furthermore, comparison of LD decay between different populations has only been performed empirically without application of a hypothesis test to determine whether the curves differ significantly. Thus, it is desirable to establish statistical approaches that test the equality of LD decay curves.

The objective of this study was to compare nonlinear and loess regression models for prediction of LD decay and to evaluate the impact of the regression models on LD predictions including statistical hypothesis tests.

3.2 Methods

3.2.1 Data

Data used in this study consisted of animals from two commercial pig lines (SL1, n=1,307 and DL2, n=1,013). SL1 is a synthetic sire line that was created around 1980 as a combination of the Duroc and Belgian Landrace. DL2 is a Large White based dam line. All animals were genotyped using the Illumina Porcine SNP60 Beadchip. However, only markers located on chromosome 18 (SSC18, n=1,456 SNPs) were used in this study. R software (http://www.R-project.org/) was used for marker quality control within lines using the package GenABEL (Aulchenko et al., 2007). Markers with a genotype call rate <90%, minor allele frequency (MAF) <0.05, and strong deviation from the Hardy-Weinberg equilibrium (HWE) (*P*<0.0001) were excluded. Only SNPs that passed quality control in both lines were included in the analysis, resulting in a marker set of 830 SNPs.

3.2.2 Linkage disequilibrium

For each population, the LD between SNPs was computed as the correlation of gene frequencies (r_{ij}^2) (Hill and Robertson, 1968) using the function LD in the R package *genetics* (Warnes and Leisch, 2005):

$$r_{ij}^2 = \frac{\left(p_{ij} - p_i p_j\right)^2}{p_i\left(1-p_i\right)p_j\left(1-p_j\right)}$$

where p_i and p_j are the marginal allelic frequencies at the i^{th} and j^{th} SNP, respectively, and p_{ij} is the probability of the marker allele pair ij (Warnes and Leisch, 2005).

3.2.3 Nonlinear regression

The pairwise r_{ij}^2 were regressed on the distance between the marker pairs based on the nonlinear model described by Sved (1971):

$$LD_{ijk} = 1/(1 + 4\beta_k d_{ij}) + e_{ijk} \qquad (1)$$

where LD_{ijk} was the observed r_{ij}^2 between SNPs i and j in line k;

d_{ij} was the distance in Kb (kilo-base pair) between SNPs i and j;

β_k was the coefficient that describes the decline of LD with distance for line k, and

e_{ijk} was a random residual defined as $e_{ijk} \overset{iid}{\sim} N(0,\sigma^2)$. Smaller values of β_k indicate a higher extent of LD.

A test for the equality of curves (Bates and Watts, 1988) was implemented to compare the nonlinear curves of the two evaluated lines. This test allows a statistical comparison of the LD decay parameter (β). Considering the following hypothesis test:

$H_0 : \beta_k = \beta \, \forall \, k$ vs $H_a : \beta_k \neq \beta$ for at least one β_k

for $k = 1,...,g$ (the number of populations), a dummy variable (D_k) is attributed to the model, such that:

$$D_k = \begin{cases} 1 \text{ if the observatio n } LD_{ik} \text{ belong to the group } k \\ 0 \text{ otherwise} \end{cases}$$

Thus, equation (1) can be written as:

$$LD_{ijk} = \sum_{k=1}^{g} D_k \left[1/(1 + 4\beta_k d_{ijk}) \right] + e_{ijk} \qquad (2)$$

which is a complete model (Ω) without restrictions on the parametric space. To conduct the statistical comparison, a reduced model (ω) was fitted with the restrictions imposed on H_0 :

$$LD_{ijk} = \sum_{k=1}^{g} D_k \left[1/(1 + 4\beta d_{ijk}) \right] + e_{ijk} \qquad (3)$$

where a single parameter (β) for all lines was assumed.

Thus, the statistics of the likelihood ratio test (L) can be written as:

$$L = \left(\frac{\hat{\sigma}^2_\Omega}{\hat{\sigma}^2_\omega} \right)^{N/2}$$

where N was the number of observations and $\hat{\sigma}^2_\Omega$ and $\hat{\sigma}^2_\omega$ were the maximum likelihood estimates for the residual sum of squares (RSS) of the complete and reduced models, respectively. According to Rao (1973), this can for large samples be described as:

$$-2\ln L = -N\ln\left(\frac{\hat{\sigma}^2_\Omega}{\hat{\sigma}^2_\omega} \right) \xrightarrow[N \to \infty]{} \chi^2_v$$

For the likelihood test:

$$\chi^2_{computed} = -N\ln\left(\frac{\hat{\sigma}^2_\Omega}{\hat{\sigma}^2_\omega} \right) = -N\ln\left(\frac{RSS_\Omega}{RSS_\omega} \right)$$

where RRS_Ω and RRS_ω are the residual sum of squares of the complete and reduced models.

The hypothesis H_0 is rejected if $\chi^2_{computed} \geq \chi^2_{\alpha(v)}$, where $v = g - 1$. Rejection of the hypothesis H_0 implies that at least one parameter β differs from the others. The nonlinear models were solved using the function nls of the software R v.2.14.2 (http://www.r-project.org/) and the hypothesis tests were implemented using custom R scripts.

3.2.3 Loess regression

Locally estimated regression and smoothing scatterplots (loess) uses a smooth curve to describe the relationship between variables without assuming a functional relationship between them. Assuming a simple regression as follows:

$LD_{lk} = g_k(x_{lk}) + e_{lk}, \mathrm{k} = 1,...,\mathrm{g}\ \mathrm{and}\ \mathrm{l} = 1,...,\mathrm{n_k}$

LD_{lk} is the linkage disequilibrium of the marker pair l of line k;

x_{lk} is the distance between markers of the pair l of the line k;

$g(.)$ is a unknown function; and

e_{lk} is the random residual of the marker pair l of line k.

In the loess regression model, the estimation is fragmented to remove noise from the data. A function $g(.)$ is estimated in the neighborhood of each point of interest $x = x_0$. The smoothing span (f) defines the size of such a neighborhood, which is a critical point for the estimation. The maximum value of f is 1, indicating that all data will be used for the fit. Values of f smaller than 1 indicate that a subset of the data will be used for the estimation. The improved Akaike Information Criterion (Hurvich et al., 1998) was used to select the smoothing span; this method avoids large variability and undersmoothing. The analyses were conducted using the function loess.as of the R package fANCOVA (Wang, 2010). An ANOVA-type statistic (Dette and Neumeyer, 2001) was used to test the equality of the nonparametric curves fitted for both lines evaluated in this study. In this methodology, the equality of smoothing curves is tested according to the hypothesis:

$H_0 : g_k(.) = g(.) \forall \mathrm{k}\ \mathrm{vs}\ H_a : g_k(.) \neq g(.); \mathrm{for}\ \mathrm{k} = 1,2,...,\mathrm{g}$

The test motivated by one way ANOVA is given by:

$$Y_N = \frac{N}{\hat{S}^2} T_N$$

where

$$T_N = \frac{1}{N} \sum_{k=1}^{g} \sum_{l=1}^{n_k} \left[\hat{g}(x_{lk}) - \hat{g}_k(x_{lk})\right]^2$$

and

$$\hat{S}^2 = \frac{1}{2(N-g)}\sum_{k=1}^{g}\sum_{l=1}^{n_k}\left(LD_{k,l+1} - LD_{k,l}\right)^2$$

being that $N = \sum_{k=1}^{g} n_k$ denotes the total sample size and n_k is the number of observations of each population k.

The function T.aov of the R package fANCOVA (Wang, 2010) was used to perform the analysis.

3.2.4 Comparison of models

Coefficients of determination (R²) and a range of residual graphs were used to compare the results obtained using the nonlinear (parametric) and loess (nonparametric) regression models. The plots were used to evaluate whether the residuals hold the assumptions of normal distribution, homogeneity variance and error independence. Histograms and QQ-plots were used to assess whether the residuals were normally distributed. Homogeneity of residual variance was evaluated by plotting the residuals against the fitted values, and residual independence was verified by plotting the residuals against the distance between markers.

3.2 Results

3.2.1 Nonlinear regression

The parameter (β) that describes the decline of LD was 0.0033 and 0.0031 for lines SL1 and DL2, respectively; both values were significantly different from zero (*P* <0.001). The comparison of the β values of the two lines was performed using an equality of curves test, which revealed that for SSC18 the parameters did not differ significantly (i.e., the same extent of LD was observed for both lines).

3.2.2 Loess regression

The estimation of LD using the loess regression model depends on the smoothing span, which was chosen using the improved Akaike Information Criterion. The resulting span was the same for the two populations (0.05), revealing that the same control of smoothness was applied for both. The equality of the nonparametric curves for the two lines was tested using an ANOVA-type statistic. The curves did not differ statistically, which is in agreement with the test for nonlinear estimation.

3.2.3 Comparison of models

To evaluate which approach is the most appropriate to predict LD decay, we compared the fit of the nonlinear (parametric) and loess (nonparametric) regression models using the coefficient of determination (R^2) and residual plots. Although both models presented small R^2 values, a slightly higher value was observed for loess regression (0.27 and 0.33 for nonlinear and loess regression in SL1, respectively). This indicates that a small proportion of the total variation of LD decay is explained by these models (Figure 3.1). The same pattern was observed for both lines, with the nonlinear regression model predicting higher values of LD than the loess model when distances between markers were small. Additionally, the nonlinear regression model predicted a faster decay of LD in comparison to the loess regression model.

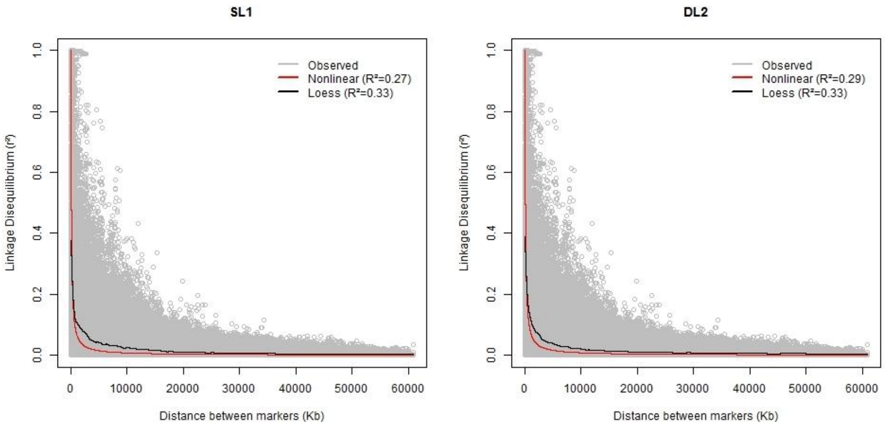

Figure 3.1 Observed and predicted values of linkage disequilibrium (r^2) in relation to the distance on SSC18 using nonlinear and loess regression for two pig lines (SL1 and DL2).

The histograms of residuals were not symmetric or bell-shaped for either model (Figure 3.2). An inflated density of residuals with values close to zero was observed. However, in the QQ-plot a better fit of the loess model was observed compared to the nonlinear model. The points of the nonlinear model deviated further from the straight line compared to the loess function.

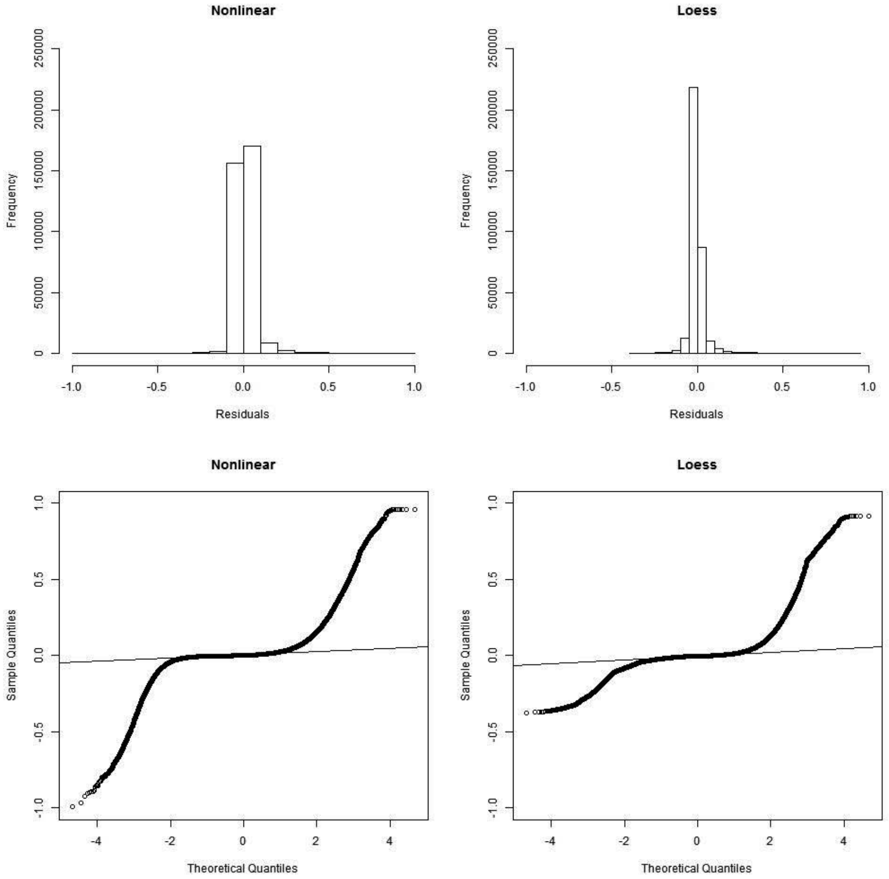

Figure 3.2 Residual frequency and QQ plot for nonlinear and loess regression in SL1.

The homogeneity of residual variance was evaluated by plotting the residuals against the fitted values. A pattern with larger residuals at lower distances was observed for both models (Figure 3.3). However, the deviation from homogeneity was more pronounced with the nonlinear model. Independence of errors was determined by plotting the residuals against the distance between markers. A clear influence of the distance between markers on the residuals was observed, with an increase in variability in residuals when markers were close together (Figure 3.3). As observed for the other residual plots, the nonlinear model performed worse, exhibiting evident large negative residual values when the distance between markers was small. The DL2 residual plots were similar to Figures 3.2 and 3.3 [see Supplementary material: Figures S3.1 and S3.2].

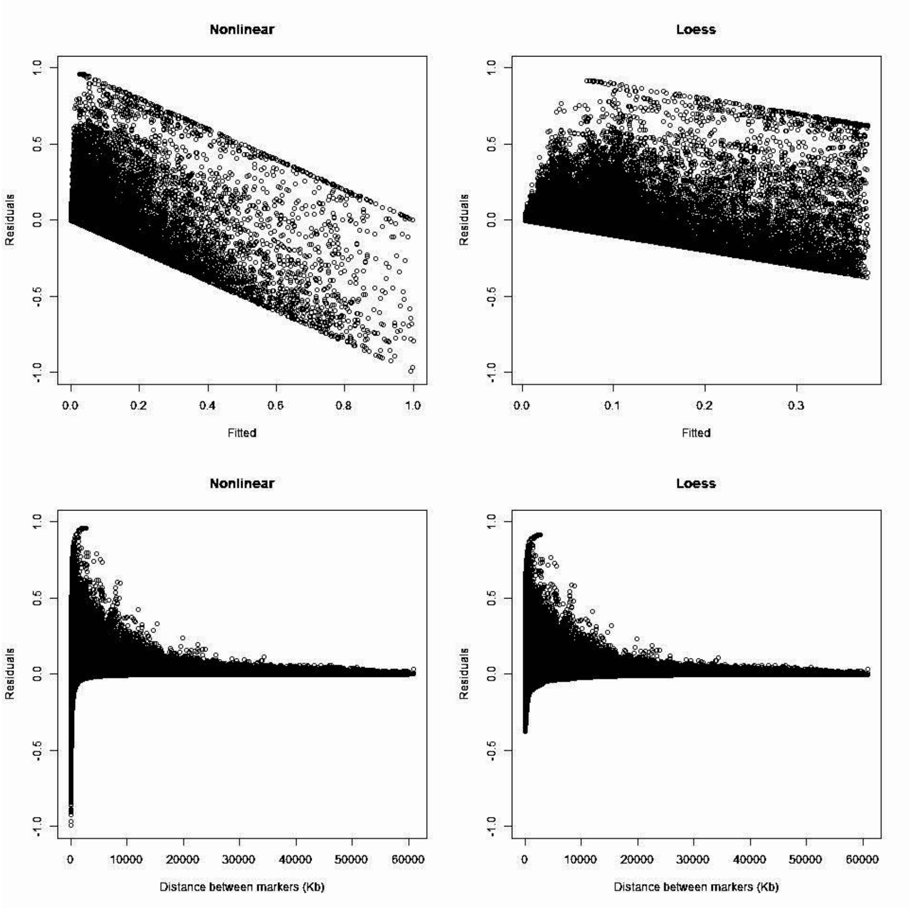

Figure 3.3 Plots of the residuals against fitted values (top) and against the distance between markers (bottom) for nonlinear (left) and loess regression (right) in line SL1.

We predicted the LD for different marker densities using nonlinear and loess regression. The average LD between markers for different marker densities was also computed for use as a reference value. The tested marker densities corresponded approximately to 10, 25, 50, 75 and 100% of the SNPs present in the current SNP chip panel. For the smallest panel (6,000 SNPs), the average and predicted LD from both models were similar (Table 3.1). However, for higher marker densities the LD predicted using nonlinear regression was larger than both the average and the predicted LD using loess regression. The LD predicted using loess was always smaller than the empirical average, but much closer to that average value than was the prediction from the nonlinear model (Table 3.1).

Table 3.1 Predicted and average linkage disequilibrium (r^2) for different marker densities.

Number of SNPs	6,000	15,000	30,000	45,000	60,000
Density (Kb/SNP)	467	187	93	62	47
LD Average	0.21	0.27	0.31	0.34	0.36
LD Nonlinear	0.18	0.35	0.52	0.62	0.69
LD Loess	0.22	0.27	0.29	0.30	0.30

Considering LD >0.3 between markers as a suitable LD level for GS and GWAS, we estimated the required number of SNPs for GWAS and GS based on the average LD and predicted LD using the nonlinear and loess regression models. This predicted number of required markers differed considerably across models, with 37,333 required SNPs based on the average LD, 11,667 SNPs based on the nonlinear regression model and 62,222 SNPs based on the loess regression model. Using the nonlinear model, 3 times fewer markers were required compared to the number of markers required when the average LD was applied.

3.3 Discussion

The comparison between nonlinear and loess regression to describe LD decay in two distinct populations showed that the loess regression model provides a better fit to the LD decay than the nonlinear model. However, when the LD decay of the two populations was compared the same result was found using a test for equality of nonlinear curves and ANOVA-type statistic for nonparametric curve comparison.

The great interest in GS and GWAS has resulted in more studies that describe (Khatkar et al., 2008; Bohmanova et al., 2010; García-Gámez et al., 2012), compare (Amaral et al., 2008; Uimari and Tapio, 2011; Badke et al., 2012; Alhaddad et al., 2013; Veroneze et al., 2013) or use LD as auxiliary information to explain the results of genetic studies (Duijvesteijn et al., 2010); however, none of these studies has evaluated how well the nonlinear and loess models fitted the data.

The real observations are widely scattered around the curves generated using the nonlinear and loess regression models for both lines. This is mainly true for short distances where the LD has large variability; in turn, this variability explains the small values of R^2.

Our residual graphs revealed superiority of the loess model compared to the nonlinear model because the lack of residual independence, normality and homogeneity of variance are more noticeable for the nonlinear model than for the loess model. Furthermore, the better fit of the loess regression model may be because this model is more flexible, without the need for parameters to give shape to the curve (Schmidt et al., 2013) while at the same time requiring weaker assumptions (Andersen, 2009b).

Loess regression has been used successfully to describe nonlinear relationships between variables in genetics and animal breeding. Gulisija et al. (2007) evaluated the nonlinear patterns of inbreeding depression and found that loess improved the fit over that of first-order regression on inbreeding for milk yield traits.

The observed differences between the nonlinear and loess models may have been influenced by the design of the porcine SNP chip. When two SNPs were close together on the genome and they exhibited high LD only one of them may have been included in the SNP chip. This selection process may have lead to artificially lower averages of LD at short distances.

The equation proposed by Sved (1971) assumes that the value of LD at the intercept (when the distance between markers is zero) is equal to one; the impact of this assumption was evaluated by Corbin et al. (2010), who found that fixing the intercept at one resulted in approximate doubling of the parameter β, thereby impacting predictions of LD and effective population size. Furthermore, using modified equations that include a parameter to estimate the intercept Corbin et al. (2010) found values above two for the intercept.

The LD pattern is controlled by a range of factors that are not completely understood, such as genetic drift, population structure and growth, migration, natural selection, mutation and recombination (Ardlie et al., 2002). These elements make the LD adjustment challenging, especially for short distances where the LD has large variability. Therefore, the prediction of LD over short distances or prediction of Ne over many generations cannot be precisely estimated using deterministic equations that force the LD to assume a certain value for short distances.

We compared the nonlinear and loess regression models to describe LD decay and evaluated the lack of fit of both models. The results revealed a rationale for the use of loess regression estimates to improve our understanding of the LD. For markers that are close together, the LD is much smaller than unity and the loess regression model predicts a slower LD decay in comparison to the nonlinear model. Moreover, the ANOVA-type test for the comparison of nonparametric curves and the equality of curves test for nonlinear models made it possible to perform a statistical comparison of the LD decay in different breeds or species (or generations).

The large variability of LD over short distances attributed to gene conversion (Wall, 2001) limits the capacity of its prediction over short distances; however, the impact of this variability is less pronounced in the loess model than in nonlinear regression models. In complex situations where assumptions of residual normality, independence and homogenous variance do not hold, the loess regression model can perform better. Lack of fit analysis showed that the nonlinear regression model violates many more assumptions than the loess regression model.

3.3.1 Implications for selecting SNP

The nonlinear and loess regression models allow predictions of LD for marker distances for which no data are available. However, this sort of prediction will typically be from larger to smaller distances given the typical progression from low to medium and high density SNP panels. We tested the impact of these models on the estimation of LD for the selection of a SNP panel with marker densities that corresponded to approximately 10, 25, 50, 75 and 100% of the SNPs present in the current SNP chip panel. For low marker densities (467 Kb/SNP), the methods provided similar predictions of the LD; however, when the distance of SNPs decreased the nonlinear model tended to overestimate the LD level. Furthermore, when predicting the number of SNPs required to attain a certain LD level, the nonlinear model clearly underestimated the number of markers given unrealistic estimates.

The selection of SNP markers based on the predicted LD using the nonlinear model resulted in a LD that was smaller than the predicted for panels with more than 15,000 markers in pigs, which can have consequences for the analysis that will be performed with this set of markers.

3.3 Conclusions

The loess regression model is less influenced by the lack of residual normality, independence and homogeneity of variance than is the nonlinear regression model and results in a better fit to LD decay curve. The loess regression model results in more reliable LD predictions and is therefore more appropriate for the design of customized SNP chips than nonlinear models. Both statistical approaches demonstrated can be used to formally compare the LD decay curves between the two populations evaluated in this study, which showed non-significant differences.

3.4 Acknowledgements

We thank the CAPES/Nuffic program, CNPq and FAPEMIG for financial support, and the TOPIGS Norsvin for providing data and assisting in manuscript preparation.

4

Accuracy of genome-enabled prediction exploring purebred and crossbred pig populations

Renata Veroneze[1,2], Marcos S Lopes[2,3], André M Hidalgo[2,4], Simone EF Guimarães[1], Fabyano F Silva[1], Barbara Harlizius[3], Paulo S Lopes[1], Egbert F Knol[3], Johan AM van Arendonk[2], John WM Bastiaansen[2]

[1] Universidade Federal de Viçosa, Departamento de Zootecnia, Viçosa, Brazil; [2] Wageningen University, Animal Breeding and Genomics Centre, Wageningen, the Netherlands; [3] Topigs Norsvin Research Center, Beuningen, the Netherlands; [4] Swedish University of Agricultural Sciences, Departament of Animal Breeding and Genetics, Uppsala, Sweden

Journal of Animal Science (Accepted)

Abstract

Pig breeding companies keep relatively small populations of pure sire and dam lines that are selected to improve the performance of crossbred animals. This design of the pig breeding industry presents challenges to the implementation of genomic selection (GS) which requires large datasets to obtain high accurate genomic breeding values. The objective of this study was to evaluate the impact of different reference sets (across- and multi-population) on the accuracy of genomic breeding value in three purebred pig populations and to assess the potential of using crossbred performance in genomic prediction. Data consisted of phenotypes and genotypes on animals from three purebred populations (sire lines SL1, n=1146; SL2, n=682; and SL3, n=1264) and three crossbred pig populations (TER1, n=183; TER2, n=106 and TER3, n=177). Animals were genotyped using the Illumina Porcine SNP60 Beadchip. For each purebred population, within-, across- and multi-population predictions were considered. In addition, data from the paternal purebred populations were used as reference set to predict the performance of crossbred animals. Backfat thickness phenotypes were pre-corrected for fixed effects and subsequently included in the GBLUP model. A genomic relationship matrix that accounted for the differences in allele frequencies between lines was implemented. Accuracies of GEBVs obtained within the three different sire lines varied considerably. For within-population prediction, SL1 showed higher values (0.80) than SL2 (0.61) and SL3 (0.67). Multi-population predictions had similar accuracies to within-population for the validation in SL1. For SL2 and SL3 the accuracies of multi-population prediction were similar to the within-population prediction when the reference set was composed by 900 animals (600 of the target line plus 300 of other line). For across-population predictions, the accuracy was mostly close to zero. The accuracies of predicting crossbred performance were similar for the three different crossbred populations (ranging from 0.25 to 0.29). In summary, the differences in accuracy of the within-population scenarios may be due to line divergences in heritability and genetic architecture of the trait. Within- and multi-population predictions yield similar accuracies. Across-population prediction accuracy was negligible. The moderate accuracy of prediction of crossbred performance appears to be a result of the relationship between the crossbreds and its parental lines.

Key words: backfat thickness, within-population, crossbred, multi-population, genetic architecture

4.1 Introduction

The advantages from using genomic information in breeding, such as reduction in the generation interval and/or an increase in the prediction accuracy of young animals, have led to industry-wide application of genomic selection (GS) especially in dairy cattle (VanRaden et al., 2009). Pig breeding companies keep a range of pure sire and dam lines that are selected to improve the performance of crossbred animals (the final “finisher” product of the pig industry). This design of the pig industry is posing specific challenges to the implementation of GS because phenotypes of interest are expressed by crossbreds and individual pure lines are relatively small. Data on crossbreds could, therefore, provide a powerful addition to the reference set because these animals have close relationships with their pure lines ancestors. However, a large number of crossbred animals with genotype as well as phenotype information is typically not (yet) available in pig breeding programs. Therefore, the size of training datasets may be increased by adding data from other populations (multi-population GS) or using a reference set completely composed of animals from a different, unrelated population (across-population GS). Multi- and across-population reference sets have been tested in simulation studies (Toosi et al., 2010; Ibáñez-Escriche et al., 2009; Zeng et al., 2013) and with real data from cattle (Hayes et al., 2009), sheep (Legarra et al., 2014), pigs (Hidalgo et al., 2015) and chicken (Simeone et al., 2012). Simulation has pointed to significant gains in accuracy mainly from multi-population reference sets, whereas studies that use real data have shown favorable as well as unfavorable outcomes. The objective of this study was to evaluate the prediction accuracy of purebred performance using different reference sets (within-, across- and multi-populations) and the prediction accuracy of crossbred performance using the purebred sire line population as the reference set.

4.2 Material and methods

Data recording and sample collection were conducted strictly in line with the Dutch law on the protection of animals.

4.2.1 Data

Data for this study consisted of phenotypes and genotypes on animals from three purebred pig populations (sire line SL1, n=1146; SL2, n=682 and SL3, n=1264) and three crossbred finishing pig populations (TER1, n=183; TER2, n=106 and TER3, n=177). SL1 is a Duroc-based population, SL2 is a population based on a combination of Large White and Pietrain populations, and SL3 is a Pietrain population. TER1, TER2 and TER3 are commercial finishing pigs resulting from a cross between an F_1 dam (Large White x Landrace) and a sire from SL1, SL2, or SL3, respectively.

Animals were genotyped using the Illumina Porcine SNP60 Beadchip (Ramos et al., 2009). The package GenABEL (Aulchenko et al., 2007) implemented in R (http://cran.r-project.org/) was used for sample and Single Nucleotide Polymorphism (SNP) quality control. Individuals with call rates <95% and markers with call rates <95% and/or minor allele frequency (MAF) <0.01 within each population were excluded. For the purebred populations, SNPs that deviated from Hardy-Weinberg equilibrium (HWE) ($P < 10^{-7}$) were also removed. Genotypes from crossbred animals were not tested for HWE, because the assumptions are not applicable. Single Nucleotide Polymorphisms located on sex chromosomes were also excluded. Missing genotypes of SNPs that were retained after quality control were imputed using the software BEAGLE 3.3.2 (Browning and Browning, 2009) assuming the default parameters.

The trait evaluated in this study was backfat thickness (BF) and a summary of the genotype and phenotype data is presented in Table 4.1. The response variables used in the genomic predictions were phenotypes pre-corrected for fixed effects instead of the original observations. In order to more accurately account for the contemporary group effects, they were estimated in a larger data set (706,023 animals) that included all contemporaneous animals of the genotyped animals in the pre-correction of the phenotypes. The estimates of the fixed effects used for the pre-corrections of the phenotypes were obtained fitting a single trait pedigree-based linear model using ASReml v3.0 (Gilmour et al. 2009). The model consisted of sex, herd-year-month, and the covariate weight at the time of measuring BF as fixed effects and the animal additive genetic, common litter and residual as random effects.

Table 4.1 Genotypic and phenotypic (backfat thickness) data description.

Line	SL1	SL2	SL3	TER1	TER2	TER3
Genotyped animals	1,405	842	1,475	254	280	233
Phenotyped animals	1,146	682	1,264	183	106	177
SNPs[1]	45,151	41,713	45,225	48,751	47,091	48,532
Average Backfat	10.41	10.23	7.83	11.54	10.60	11.23
Heritability[2]	0.50	0.39	0.33	0.44	0.25	0.35

[1] The initial number of SNPs was 64,232

[2] Computed using only pedigree information

4.2.2 Multidimensional scaling (MDS)

To evaluate the relationships between and within breeds, a multidimensional scaling was applied to the genomic relationship matrix (G) that was computed as described by Van Raden (2008): $G = ZZ' \big/ 2\sum p_i q_i$, where Z is a matrix of centered genotypes and p_i and q_i are the allelic frequencies of the i^{th} SNP based on observed genotypes. The SNP genotypes were coded as 0, 1 and 2. MDS maps a high dimensional space to a low-dimensional projection of the data while preserving, as closely as possible, the pairwise distances between data points (Bishop, 2006). The method rests on the eigenvalue decomposition of the distance matrix to find a configuration of points in a space where each point represents one of the objects or individuals (Cox and Cox, 2008). The analysis was done using the function *cmdscale* implemented in R (http://www.R-project.org/).

4.2.3 Scenarios

For each sire line, predictions were made within- (scenario 1), across- (scenarios 2 and 3) and with multiple populations (scenarios 4 to 7). The size of the reference set was kept constant at 600 animals, except for the multi-population scenarios. In the multi-population scenarios, an additional situation was tested, supplementing the within-population reference set with 300 animals from a different line (scenarios 5 and 7). In scenarios 8 - 10, the potential to predict crossbred performance with the paternal line as reference set was investigated.

Animals were randomly assigned to the reference and validation sets. The predictions were repeated 20 times with different reference and validation sets (20-fold cross validation).

4.2.3 Statistical analysis

The GBLUP model was used for the prediction of GEBVs in all evaluated scenarios with the model: $y = 1\mu + Zg + Wc + e$, where y is the phenotype corrected for fixed effects; μ is the overall mean; g is the vector of breeding values, $g \sim N(0, \sigma_g^2 G)$; c is the vector of random litter effect, $c \sim N(0, \sigma_c^2 I)$; e is the vector of residuals, $e \sim N(0, \sigma_e^2 I)$. Z and W are the incidence matrices for *g* and *c,* respectively.

In the multi-population scenarios, a fixed effect of population was added to the model. The genomic relationship matrix was built according to Chen et al. (2013), accounting for differences in allele frequencies between populations. Summarizing, X was a matrix with genotype values coded as -1, 0 and 1 for the three SNP genotypes and with dimension n x m (number of animals x number of SNPs). Matrix X included all animals from both the reference and validation sets. The matrix X was organized into two blocks: $X=\begin{bmatrix}X_1 & X_2\end{bmatrix}'$ where X_1 represented the genotypes of line 1 and X_2 the genotypes of line 2. P was a matrix of allele frequencies $P=\begin{bmatrix}P_1 & P_2\end{bmatrix}'$ corresponding to X, each row in P1 (or P2) was a replicated row vector p1 (or p2) with the frequency of allele A for SNP *k* in line 1 (or line 2). The matrix Z was computed to set mean values of the allele effects to 0: $Z=\begin{bmatrix}Z_1 & Z_2\end{bmatrix}'=X-2P+1$ where 1 represents a matrix of ones.

A 2-population genomic relationship matrix was constructed as (Chen et al., 2013):

$$G^*=\begin{bmatrix} Z_1Z_1^{'}\big/2\sum p_{1k}(1-p_{1k}) & Z_1Z_2^{'}\big/2\sum[p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})]^{1/2} \\ Z_2Z_1^{'}\big/2\sum[p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})]^{1/2} & Z_2Z_2^{'}\big/2\sum p_{2k}(1-p_{2k}) \end{bmatrix}$$

The software ASREML (Gilmour et al., 2009) was used to predict the genomic breeding values with G entered as a user defined matrix (*grm* option). Animals assigned to the validation set had their phenotypes removed before predicting GEBV.

The accuracy of the breeding values was computed as the Pearson correlation between the predicted genomic breeding value and the corrected phenotype divided by the square root of the heritability. To measure the bias of the GEBV, the slope coefficient for the regression of the corrected phenotypes on GEBVs was calculated for each scenario. Values of slope different from 1 indicate a prediction bias.

4.3 Results

4.3.1 Genomic relationships between populations

In the multidimensional scaling the first two eigenvalues of G explained a high proportion of the covariance across individuals (96%) and distinguished the six evaluated populations (SL1, SL2, SL3, TER1, TER2 and TER3) (Figure 4.1). In addition, the crossbred lines (TER1, TER2 and TER3) were projected near their parental line. Although the crossbred populations share 50% of their genetic origin, because they are descendents of the same F1 cross, they could be distinguished using the two dimensions.

The heat map of individual animal relationships (Figure 4.2) showed that the relationships within line (darker blocks along the diagonal) are higher than relationships between lines and a higher relationship between SL2 and SL3 is evident in comparison with the relationship of SL1 with these two lines. As expected, each terminal crossbred population had the highest relationship with its paternal line.

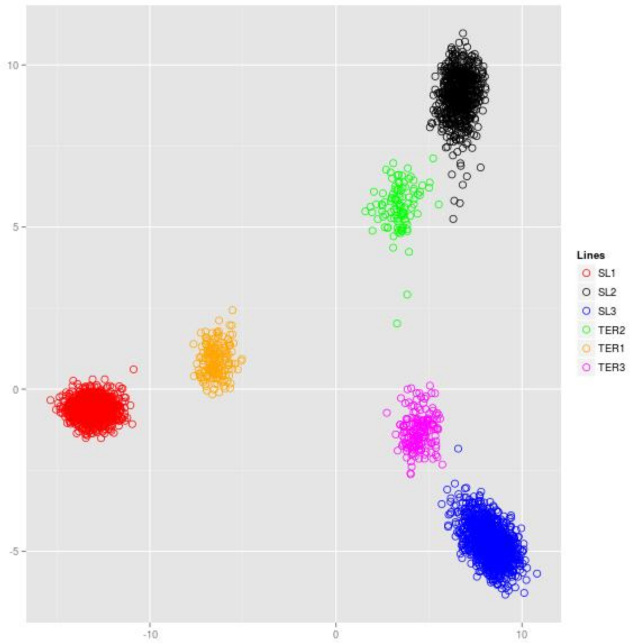

Figure 4.1 Multidimensional scaling showing a two dimensional projection of the populations distances.

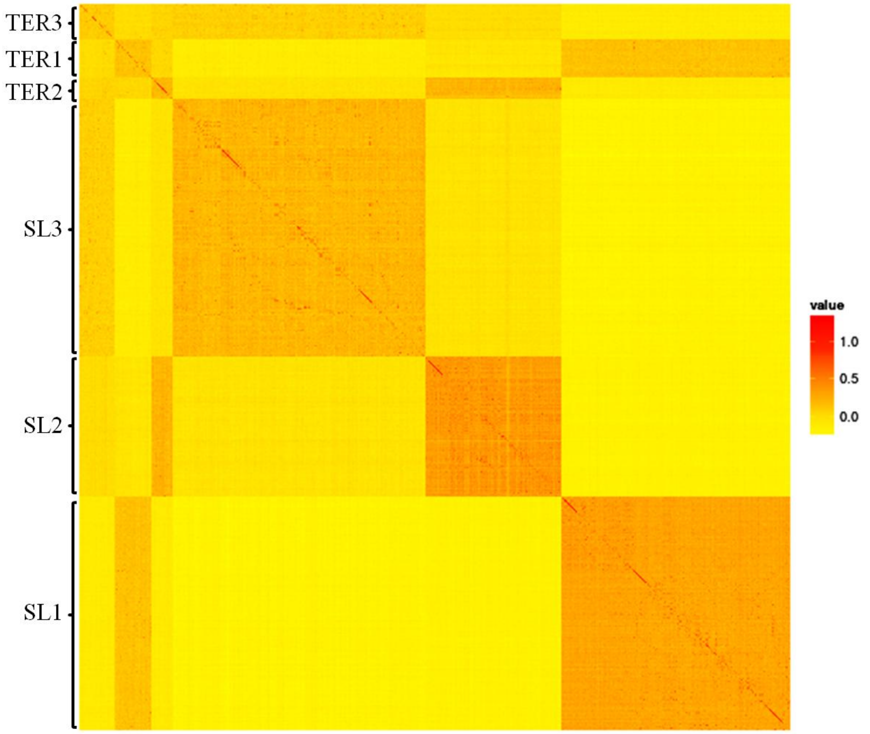

Figure 4.2 Genomic relationships heat map.

Although the crossbred populations share 50% of their genetic origin, because they are descendents of the same F1 cross, they could be distinguished using the two dimensions.

4.3.2 Prediction accuracies

The within-population accuracies were 0.80, 0.61 and 0.67 for SL1, SL2 and SL3, respectively (Table 4.2). In general, the across-populations predictions (scenarios 2 and 3) resulted in low accuracies that were not significantly different from zero. Only when predicting SL1 performance with SL3 animals as reference set, a somewhat higher accuracy was observed (0.27) (Table 4.2). In the across-population scenarios, the relationship between animals in the reference and validation set were low with no individuals having a relationship>0.1 between populations (Table 4.2).

Table 4.2 Accuracy and bias (slope) of the GEBVs and average relationship between animals in the reference and validation set for the different scenarios evaluated.

Validation population	Scenarios	Reference population	Reference population	Validation population	Accuracy backfat[1,2]	Slope	Average number of relationship 0.1 - 0.3	Average number of relationship >0.3
SL1	1	SL1	600	546	0.8	1.02	21.6	2.4
	2	SL2	600	1146	0.01	0.08	0	0
	3	SL3	600	1146	0.27	1.63	0	0
	4	SL1+SL2	600 (300+300)	846	0.82	1.2	10.8	1.2
	5	SL1+SL2	900(600+300)	546	0.86	1.07	14.4	1.8
	6	SL1+SL3	600 (300+300)	846	0.76	1.09	10.8	1.2
	7	SL1+SL3	900(600+300)	546	0.84	1.06	14.4	1.8
SL2	1	SL2	600	82	0.61	1.13	37.8	3
	2	SL1	600	682	0.07	0.26	0	0
	3	SL3	600	682	0.09	0.42	0	0
	4	SL2+SL1	600 (300+300)	382	0.35	0.7	18.6	1.2
	5	SL2+SL1	900(600+300)	682	0.48	0.91	24.6	1.8
	6	SL2+SL3	600 (300+300)	382	0.44	0.91	19.2	1.2
	7	SL2+SL3	900(600+300)	682	0.57	1.04	25.2	1.8
SL3	1	SL3	600	664	0.67	1.08	24.6	3
	2	SL1	600	1264	0.12	0.38	0	0
	3	SL2	600	1264	0.03	0.1	0	0
	4	SL3+SL1	600 (300+300)	964	0.46	0.83	12	1.2
	5	SL3+SL1	900(600+300)	664	0.56	0.88	16.2	1.8
	6	SL3+SL2	600 (300+300)	964	0.59	1.13	12.6	1.2
	7	SL3+SL2	900(600+300)	664	0.66	1.07	16.8	1.8
TER2	8	SL2	600	107	0.29	0.67	9.6	0.6
TER3	9	SL3	600	178	0.28	1.17	9.6	0.6
TER1	10	SL1	600	184	0.25	0.51	6.6	0.6

[1] Pearson correlation between GBV and corrected phenotypes divided by the square root of the heritability

[2] The average standard deviation across scenarios was 0.06

In the multi-population scenarios with 600 animals in the reference set (scenarios 4 and 6), the outcome depended on the sire line used as validation set. For the validation in SL1, the accuracies were similar to the within-population scenarios (0.76 - 0.82), whereas for SL2 and SL3, the accuracies were lower than the accuracies of the within-population scenarios. For scenarios 5 and 7, where a multi-population data set with 900 animals (600 + 300) was used, the accuracies for SL1 increased slightly, from 0.84 to 0.86, compared to the within-population scenarios. Whereas for SL2 and SL3, the accuracies from a multi-population data set were similar to within-population scenarios. In the scenarios 8-10, we used the parental sire line to predict the crossbreds. For these scenarios the accuracies were similar (0.25 – 0.29) for all three crossbred populations (Table 4.2).

The slope coefficients showed that the within-population prediction was generally unbiased whereas across-population prediction showed some evidence of bias, resulting in under or overestimated breeding values. For multi-population scenarios the slope ranged from 0.7 to 1.2 and in crossbred prediction from 0.51 to 1.17.

In the scenarios where SL1 was used as the reference set, the genomic heritability was always higher than in the scenarios where the reference set was formed by SL2 and/or SL3 (Figure 3).

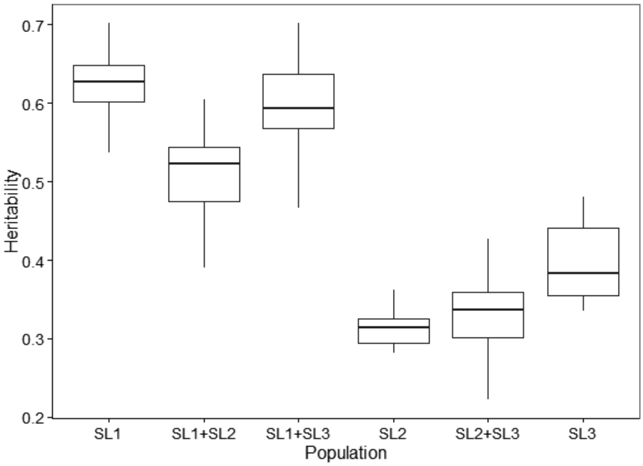

Figure 4.3 Box plot represent the genomic heritability for the different reference sets evaluated.

4.4 Discussion

The investigated scenarios (within-, across- and multi-population) were repeated with three sire lines in the validation set. Within- and multi-population results were similar and across-population accuracies were very low (ranging from 0.01 to 0.27).

4.4.1 Within-population prediction

Between sire lines the level of accuracy for within-population predictions varied considerably, with SL1 having higher accuracies than SL2 and SL3. A number of factors can influence the accuracy of genomic predictions, including the relationship between reference and validation sets (Wientjes et al., 2013), levels of linkage disequilibrium (LD), the size of the reference population, the heritability of the trait and the genetic architecture of the trait (number of QTL and distribution of QTL variance) (Hayes et al., 2009). The LD in these populations was evaluated in Veroneze et al. (2014) and the highest LD was observed in SL2, while SL1 and SL3 showed similar but lower LD. Consistent with the higher LD, the heat map (Figure 2) shows a higher level of overall relationships within SL2. The number of animals with relationship >0.1 between reference and validation sets was also higher for SL2 in comparison with SL1 and SL3.

Based on the higher LD and higher relationships, high accuracy of genomic prediction was expected for SL2 which was, however, not observed. The pedigree-based heritabilities were similar for SL2 and SL3, while SL1 presented high heritability (Table 1). Thus, because we kept the size of the reference constant, the variation in accuracy between the sire lines was attributed to differences in heritability and genetic architecture (number of QTL controlling the trait and the distribution of the trait variance in different QTL) of BF between the populations.

The high pedigree-based heritability in SL1 also reflected in higher genomic heritability in this line than in SL2 and SL3 (figure 3). The trait heritability have a direct effect on the accuracy (Hayes et al., 2009), thus the high heritability for BF of SL1 may be a explanation for the highest prediction accuracy in this line. Although the same trait was analyzed and therefore a similar genetic architecture might have been expected, these populations do have a different history of selection, differences in their genetic backgrounds as well as demography. These differences may result in a genetic architecture of the trait that is specific of each population which also affect the accuracies of the breeding values of each line and reflects on the genomic heritabilities.

Two other studies also reported quite different accuracies for prediction of BF while their reference populations were similar size. Using a population of 983 Yorkshire pigs with deregressed EBVs, Badke et al. (2014) found accuracies of 0.68

for BF when the accuracy was computed as correlation of GBV and EBV and of 0.80 when the correlation was adjusted by reliability of EBV. Studying a population of Duroc with 1,047 animals in the reference set and pre-corrected phenotypes, Jiao et al. (2014) found an accuracy of 0.36 for backfat thickness. In addition to investigating different breeds, these two studies also applied different methodologies, which may have affected the differences in accuracy for the GBV of the same trait but the different breeds also are expected to play a role. In our study, the same method was applied to all three populations and the impact of the breed (divergences in the trait variability and genetic architecture) is the only explanation for differences in accuracy.

4.4.2 Across-population prediction

For most of the across-population predictions the accuracies were near to zero, which can be explained by (i) the low relationship between lines (Table 2), (ii) the fact that the number of SNPs used was not sufficient to provide the same LD phase across the purebred lines (Veroneze et al., 2014) and (iii) different functional mutations may be segregating in the populations. The patterns of relatedness found using the G matrix (Figure 2), matched the persistence of phase results reported in Veroneze et al. (2014). The differences in accuracy observed in across- and multi-population scenarios can be at least in part be explained from these differences in linkage disequilibrium phase. Evaluating across-population prediction in two dam lines, Hidalgo et. al. (2015) found accuracies of zero for age at first insemination and for total number of piglets born whereas accuracies for litter birth weight were between 0.17 and 0.26 and for litter variation between 0.12 and 0.18. Evaluating Jersey and Holstein breeds Hayes et al. (2009b), showed that limited relationship between the breeds exist and that the genomic breeding value accuracies were low for across-population prediction. Studying Angus and Charolais, Chen et al. (2014) found accuracies between 0.10 and 0.22 for across-breed genomic prediction. In the current study we observed that even for lines that have the same breeds in their genetic background (SL2 and SL3) the accuracy from across-population prediction is very limited. This is an important outcome for the pig breeders because in several cases lines in a breeding program have common genetic backgrounds, but across line prediction does not appear to be successful given the number of samples and SNPs that are currently used.

4.4.2 Multi-population prediction

In multi-population predictions, accuracies were similar to the within-population predictions for the validation in SL1, but considerably lower for the validation in SL2 and SL3 for a same size reference set. For the validation in SL2 and SL3, the accuracies increased by 17% or more when SL2 and SL3 were combined in comparison to combinations with SL1 (Table 4.2, scenarios 4 vs 6 and 5 vs 7). These results are consistent with the higher genomic relationships (Figure 2) and higher persistence of phase (Veroneze et al., 2014) between SL2 and SL3 than between these two lines and SL1.

When the same reference set was used in multi-population scenarios (scenario 4 for example for lines SL1 and SL2), the accuracies were always found to depend quite strongly on which population was used as the validation set (0.82 for SL1 and 0.35 for SL2). As the reference set was the same, we can conclude that the effects identified by the SNPs were equal for prediction in both populations. Then, this result may be explained by differences in the true effect of the QTL, in the LD between marker and QTL and in the allele frequencies of the SNP that is tracking the QTL in the two different validation populations.

To evaluate the effect of expanding a reference set with individuals from another line, we investigated scenarios where 300 extra animals from another line were added to the reference set of 600 from the target line (scenarios 5 and 7). For SL1, the accuracy with these additional 300 animals was slightly higher than without them, whereas for SL2 and SL3 the accuracies were lower or similar when adding animals from a different line. Evaluating multi-population prediction in pig dam lines, Hidalgo et. al. (2015) also found that adding animals from another population did not increase the accuracies (personal communication). Using a data set with six breeds of sheep, Legarra et al. (2014) concluded that the addition of animals from another population increased the accuracy marginally just for a couple of breeds and that pooling populations did not increase the accuracy of genomic evaluations in dairy sheep. However, studying multi-population reference set for Jersey and Holstein breeds, Hayes et al. (2009b) found comparable or higher accuracies when multi-population reference sets were used in comparison to a purebred reference set. The drop in the relationships between reference and validation sets in the multi-population scenarios in comparison to within-population scenarios may be preventing an increase in accuracy even though a larger reference population was used. It has been shown that the accuracy of genomic prediction can be improved by reducing distances between the reference and validation animals and by increasing distances between animals within the reference population (Pszczola et al., 2012). In the current study and in the studies mentioned above, animals that were added to the multi-population reference set were selected at random from the other populations. An increase in the accuracy may be observed if animals with higher relatedness to the target population were added to the reference set (M. P. L. Calus, personal communication).

In addition to careful selection of animals to be combined in the reference, predictions using multi-population reference might improve with denser genotyping, because the LD phase between the populations studied is not persistent using the present 60K SNP panel (Veroneze et al. 2014). Moreover, we expect that models that consider the allele origin could also improve the predictions because the genetic architecture of the trait appears to be different between the lines.

4.4.2 Crossbred prediction

Simulation studies have suggested that data from crossbreds could be used to successfully select purebreds for crossbred performance (Ibánez-Escriche et al., 2009; Toosi et al., 2010; Zeng et al., 2013). The results of predicting crossbred performance using purebred data give an indication of what could be expected from such a scenario. Accuracies in the crossbred scenarios were generally higher than in the across-population predictions. Presumably this is because of the higher relationships between the animals from a sire line and the crossbreds produced with that same sire line. Accuracies were similar for the three crosses evaluated (0.25 - 0.29) which may reflect the fact that the genome sharing (50%) of the sire line with the crossbred is the same for the three scenarios evaluated.

Despite the existence of relatively high relationships between crossbreds and their parental line, the accuracies were moderate which might be an indication of differences in the genetic architecture of the traits between the purebred and crossbred populations. The inclusion of the maternal line (F1 cross) in the reference set may result in an increase in accuracy because it might raise the relationships between reference and validation population. The maternal line (F1) and crossbred exhibited high persistence of LD (Veroneze et al., 2014), which reflect the relationship between them.

Although, the most interesting scenario for the pig industry should be the use of crossbred information to select purebred, the scenario evaluated in this study is useful to illustrate what we could expected when predicting breeding values for crossbred animals. This is done in the pig industry in particular occasions i.e., when populations are crossed for the development of a new line.

4.5 Conclusions

Within- or multi-population predictions yield similar accuracies; across-population prediction accuracy was negligible even when the lines had common breeds in their genetic background. Backfat thickness appears to have a different genetic architecture in these different populations, which can influence the level of accuracy attainable from genomic predictions. Differences between validation populations in the true effect of the QTL, in the LD between marker and QTL in the validation population and in the allele frequencies of the SNP that is tracking the QTL may impact the prediction accuracies in multi-population genomic selection. The moderate accuracy of prediction of crossbred performance appears to be a result of the differences in genetic architecture between purebred and crossbred animals.

5

Accounting for genetic architecture in single- and multi-population genomic prediction using weights from GWAS

Renata Veroneze[1,2], Paulo S Lopes[1], Marcos S Lopes[2,3], André M Hidalgo[2,4], Simone EF Guimarães[1], Barbara Harlizius[3], Egbert F Knol[3], Johan AM van Arendonk[2], Fabyano F Silva[1], John WM Bastiaansen[2]

[1] Universidade Federal de Viçosa, Departamento de Zootecnia, Viçosa, Brazil; [2] Wageningen University, Animal Breeding and Genomics Centre, Wageningen, the Netherlands; [3] Topigs Norsvin Research Center, Beuningen, the Netherlands; [4] Swedish University of Agricultural Sciences, Departament of Animal Breeding and Genetics, Uppsala, Sweden

Journal of animal breeding and genetics (Submitted)

Abstract

Genome-wide association studies (GWAS) have opened the possibility of exploiting differences in genetic architecture to improve genomic predictions. We studied the effect of using GWAS results on the accuracy of single- and multi-population genomic predictions. Phenotypes (backfat thickness) and genotypes of animals from two purebred sire lines (SL1, n=1146 and SL3, n=1264) were used in the analyses. First, GWAS were conducted for each line individually and for the combined dataset (both lines together) to estimate the variance of each SNP. These estimates were used to build a matrix of weights (D), which was incorporated into a genomic best linear unbiased prediction (GBLUP) method. The single nucleotide polymorphism (SNP) effects showed correlations close to zero between the pig lines, which indicated that the lines had different genetic architectures. Single population scenarios evaluated with a traditional GBLUP had accuracies of 0.30 for SL1 and 0.31 for SL3. When the GBLUP employed weights, which had been estimated in a combined GWAS, the accuracies for both lines were higher (0.32 for SL1 and 0.34 for SL3) than those obtained with the traditional GBLUP. When unrelated animals were added to create a multi-population reference set, the accuracies were higher than those obtained with the single-population reference set and a traditional GBLUP (0.36 for SL1 and 0.32 for SL3). In addition, putting together the multi-population reference set and the weights from the combined GWAS provided even higher accuracies (0.37 for SL1, and 0.34 for SL3). The results of this study showed that the use of multi-population predictions and weights estimated from a combined GWAS could increase the accuracy of genomic predictions.

5.1 Introduction

Genome-wide association studies (GWAS) have been conducted to disclose the genetic architecture of complex traits. These studies have generated a considerable amount of information for many traits in livestock species, but this information has not been extensively exploited in genomic prediction.

For single-population genomic prediction, Zhang et al. (2010) proposed a trait-specific, marker-derived relationship matrix (TA-matrix), which had a greater predictive ability than the traditional genomic best linear unbiased prediction (GBLUP) method. Those authors attributed the improvement to the fact that the TA-matrix emphasized markers that contributed to the genetic variance of the trait.

Multi-population genomic prediction emerged as an alternative for implementing genomic selection in small populations. Combining populations from different breeds or lines increases the number of animals available, which might contribute to a more accurate prediction of genomic breeding values (GEBVs) than a single-population prediction.

Multi-population genomic predictions have been studied in cattle (Olson et al., 2012; Chen, et al., 2013), pigs (Hidalgo et al., 2015 and Veroneze et al., 2015), chickens (Simeone et al., 2012), and sheep (Legarra et al., 2014). However, the results showed that multiple populations sometimes increased and other times decreased the accuracy of genomic predictions. The variability in results reflected the fact that predictions in multi-population analyses were more complex than single-population predictions. First, increasing the reference population by adding unrelated animals could decrease the average relationship between animals in the reference and validation sets. Second, differences in the population history, like demography, inbreeding, genetic background, and selection, could lead to divergences in linkage disequilibrium (LD), allele frequencies, and genetic architecture. Genetic differences between populations depend on the number of generations since their last common ancestor, the size of the populations, and the degree of exchange of genetic material between populations.

According to Harris and Johnson (2010), differences in allele frequency between populations should be considered in a multi-population prediction. In a study on beef cattle, Chen et al. (2013) proposed a two-population genomic relationship matrix, which considered differences in allele frequencies between populations. However, they did not find increased accuracy in the multi-population prediction compared to the single-population prediction. Based on the same approach with experimental data from several different pig populations, Veroneze et al. (2015) found similar or lower accuracies for the multi-population scenarios compared to single-population scenarios. Those studies suggested that considering differences in allele frequency was insufficient to improve the accuracy of multi-population predictions.

Another potential improvement that addressed differences between populations was the incorporation of GWAS results from these populations. This approach could benefit multi-population genomic predictions, because including GWAS results could emphasize markers that explain genetic variance in the target population. In addition, differences in allele frequency between populations could be accounted for simultaneously in the two-population genomic relationship matrix. However, the value of using GWAS results in multi-population predictions has not been studied.

The objective of this study was to use results from GWAS to create weighted genomic relationships and to apply them to single- and multi-population genomic predictions.

5.2 Material and methods

Data recording and sample collection were conducted strictly in line with the Dutch law on the protection of animals.

5.2.1 Data

Data used in this study consisted of phenotypes (backfat thickness) and genotypes of animals from two purebred pig populations (SL1, n=1146; SL3, n=1264). SL1 is a combination of Duroc (mostly) and Belgian Landrace breeds, and SL3 is a Pietrain-based sire line. Animals were genotyped with the Illumina Porcine SNP60 Beadchip, and all single nucleotide polymorphisms (SNP) on both sex chromosomes were excluded. The GenABEL package, implemented in R software (Aulchenko et al., 2007), was used to perform individual sample and SNP quality control. Animals with call rates <95% and SNPs with call rates <95%, with minor allele frequencies <0.01, or with deviations from Hardy-Weinberg equilibrium ($P < 10^{-7}$) were excluded. After quality control, missing genotypes of SNPs were imputed with BEAGLE 3.3.2 software (Browning and Browning, 2013), with the default parameters.

Estimates of the fixed effects used for pre-correcting the phenotypes were obtained by fitting a single trait, pedigree-based linear model to a dataset of 706,023 animals, with ASReml v3.0 (Gilmour et al*.,* 2009). The model included fixed effects of sex, herd-year-month, and the covariate body weight at the time of measuring backfat. The animal additive genetic, litter and residual were included as random effects.

5.2.2 Model and genomic relationship matrix

The genomic best linear unbiased prediction (GBLUP) method was used for genomic prediction. The general model was: $y = 1\mu + Zg + Wc + e$, where y is the phenotype corrected for fixed effects; μ is the overall mean; g is the vector of breeding values, $g \sim N(0, \sigma_g^2 G)$; c is the vector of random litter effect, $c \sim N(0, \sigma_c^2 I)$; e is the vector of residuals, $e \sim N(0, \sigma_e^2 I)$. Z and W are the incidence matrices for *g* and *c,* respectively. In the multi-population scenarios, a fixed effect of population was added to the model.

In the genomic relationship matrix, differences in allele frequencies between populations were accounted for with the method described by Chen et al. (2013). Summarizing, X was a matrix with genotype values coded as -1, 0, and 1 for the three SNP genotypes; the matrix had dimensions, n × m, where n was the number of animals and m was the number of SNPs. Matrix X included all animals, from both the reference and validation sets. Matrix X was organized into two blocks: $X = [X_1 \;\; X_2]'$ where X$_1$ represents the genotypes of line 1, and X$_2$ represents the genotypes of line 2. P was a matrix of allele frequencies $P = [P_1 \;\; P_2]'$ that corresponded to X; each row in P1 (or P2) was a replicate row vector, p1 (or p2), with the frequency of allele A for SNP *k* in line 1 (or line 2). The matrix M was computed to set the mean values of the allele effects to 0: $M = [M_1 \;\; M_2]' = X - 2P + 1$, where 1 represents a matrix of ones. The matrix, G, was computed as follows:

$$G = \begin{bmatrix} M_1 D M_1' \big/ 2\sum p_{1k}(1-p_{1k}) & M_1 D M_2' \big/ 2\sum [p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})]^{1/2} \\ M_2 D M_1' \big/ 2\sum [p_{1k}(1-p_{1k})p_{2k}(1-p_{2k})]^{1/2} & M_2 D M_2' \big/ 2\sum p_{2k}(1-p_{2k}) \end{bmatrix}$$

Here, D is a diagonal matrix of weights for the SNPs, which will be described in detail in the next section. In the traditional GBLUP, D is an identity matrix.

5.2.3 Diagonal matrices

First, a single-population and multi-population GBLUP analysis was carried out with a G matrix, computed as described by Van Raden (2008), which included all animals:

$$G = ZZ' / 2\sum p_i q_i$$

This G matrix was entered as a user defined matrix (*grm* option) in the software ASREML (Gilmour et al., 2009) to predict the GEBVs. The predicted GEBVs $(\hat{g})$ were used to compute the diagonal elements of D, according to the method proposed by Wang et al. (2012). In that method, SNP effects $(\hat{u})$ were estimated with the equation: $\hat{u} = \lambda Z' G^{-1} \hat{g}$, where $\lambda = 1 \big/ (\sum_{i=1}^{m} 2p_i(1-p_i))$; in the latter equation, m is the number of SNPs, and p_i is the allele frequency of the second allele of the i^{th} SNP. Then, the variance of each SNP effect was estimated as described by Falconer and Mackay (1996): $\hat{\sigma}^2_{u_i} = \hat{u}_i^2 2p_i(1-p_i)$. These variances were used to build the diagonal elements of the D_0 matrix. The D_0 matrix was normalized with $D = (tr(I)/tr(D_0)) * D_0$, where *I* is an identity matrix.

Four different D matrices were used in this study: one was an identity matrix (traditional GBLUP), and three were D matrices obtained with the three datasets used to estimate the weights (Figure 5.1). These diagonal D matrices contained weights for the SNPs; these weights were included in the genomic relationship matrix, which resulted in four different G matrices (G^{identity}, $G^{\text{D\_comb}}$, G^{D1}, and G^{D3}). Each G matrix was used to predict GEBVs for animals from SL1 and SL3, with both single- and multi-population reference sets.

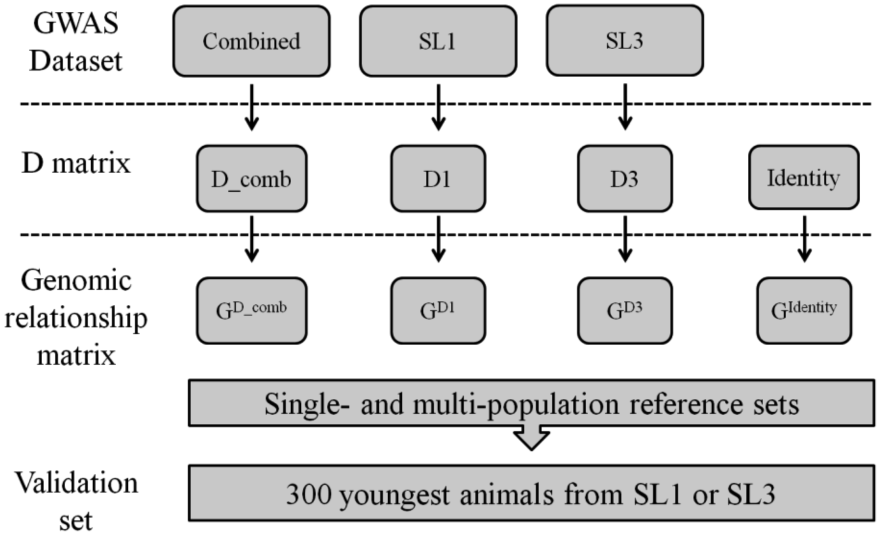
GWAS Dataset
Combined
SL1
SL3
D matrix
D_comb
D1
D3
Identity
Genomic relationship matrix
G^D_comb
G^D1
G^D3
G^Identity
Single- and multi-population reference sets
Validation set
300 youngest animals from SL1 or SL3

Figure 5.1 Schematic representation of the scenarios evaluated. Combined refers to a dataset composed by the combination of SL1 and SL3.

The predictions were conducted according to four strategies: First, the traditional GBLUP, where all markers had the same weight (scenarios 1, 5, and 9 in Table 1);

second, markers were weighted, and the weights were from the same population that will be predicted (scenarios 2, 6, and 10); third, the markers were weighted, but the weights were from an unrelated population (scenarios 3, 7, and 11); and fourth, the markers were weighted, and the weights were from a combined GWAS, which used both pig lines together (scenarios 4, 8, and 12). In scenarios 1-4, the predictions were performed with a single-population reference set. In scenarios 5-8, half of the animals were substituted with individuals from another population to calculate multi-population predictions. In scenarios 9-12, two unrelated populations were combined, which doubled the number of animals, to calculate multi-population predictions.

The validation sets consisted of the 300 youngest animals of each population. The reference sets consisted of 800 animals for single-population predictions and 800 (400 animals of each line) or 1600 (800 animals of each line) animals for multi-population predictions.

The accuracy of the GEBVs was computed as the Pearson correlation between the predicted GEBV and the corrected phenotype. To measure the bias of the GEBV, the slope coefficient of the regression of the corrected phenotypes on GEBVs was calculated for each scenario.

5.3 Results

5.3.1 SNP effects

Our aim was to improve genomic predictions, based on single- and multi-population reference sets. The first step of our methodology was to estimate the effects of each SNP on backfat thickness that were used to compute the weights for our genomic markers. Manhattan plots (Figure 5.2) show these SNP effects for the SL1 and SL3 lines, separately, and for the combined SL1+SL3 dataset. The estimated SNP effects in the SL3 dataset were, on average, lower and less variable than the SNP effects in the SL1 dataset. When the two pig lines were combined (SL1+SL3), the average effects became larger than those observed for a single-population analysis. In addition to the Manhattan plots, the dispersion plots (Figure 5.3) clearly showed differences in the genetic architecture between these pig lines. Although most SNPs had effects near zero in both lines, the correlation between the SNP effects in the two lines was only 0.02. The dispersion plots of SNP effects estimated for the combined dataset and the SNP effects for the single populations showed higher correlations between the effects (0.36 for SL1 and 0.38 for SL3).

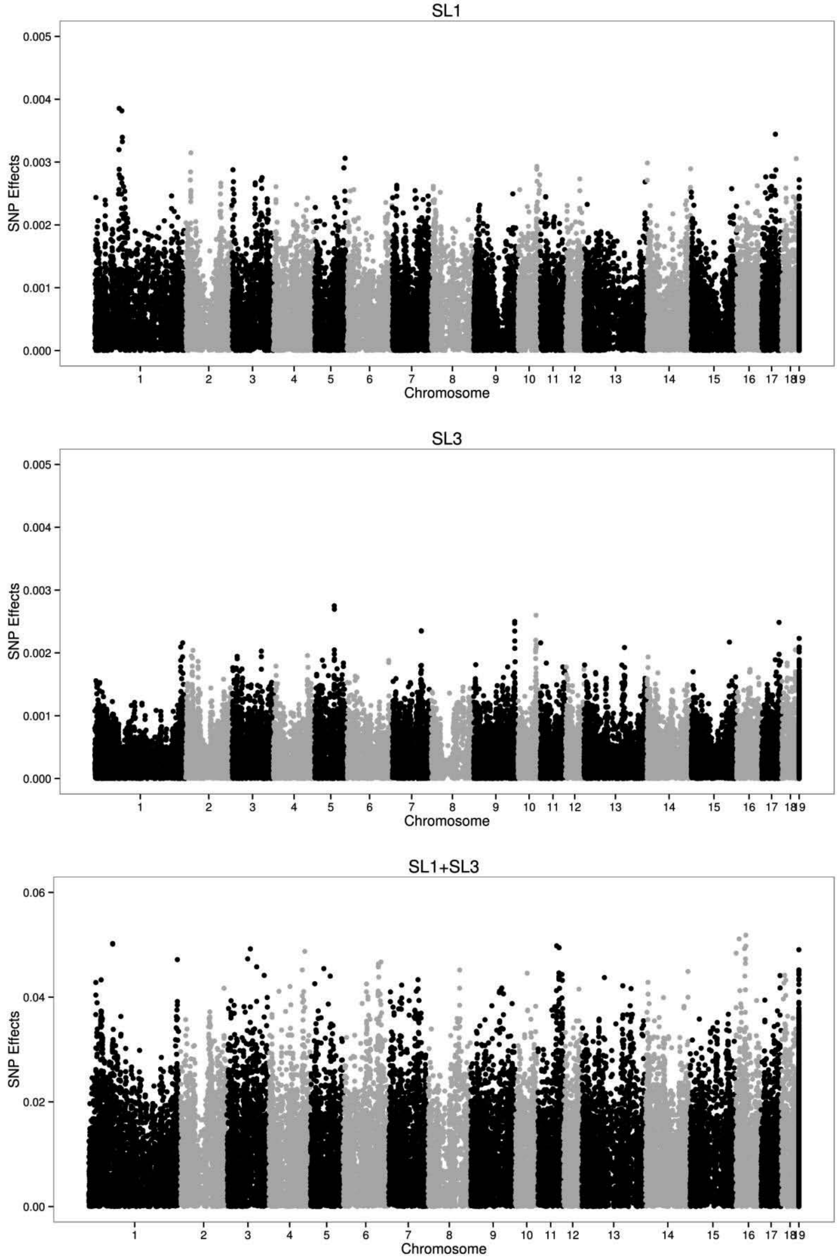

Figure 5.2 Manhattan plots of the SNPs effects for backfat thickness for SL1, SL3 and the combined data set (SL1+SL3).

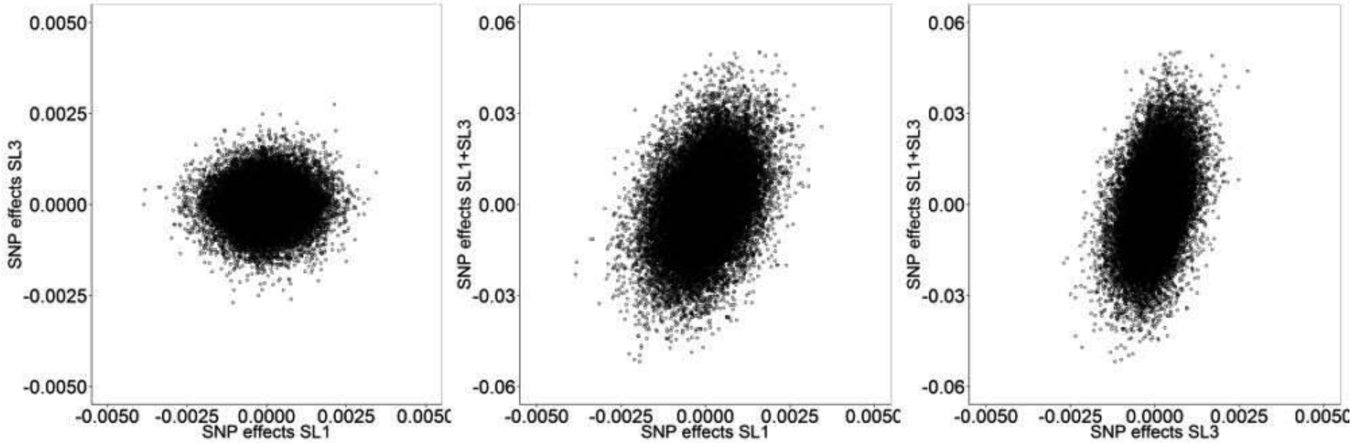

Figure 5.3 Dispersion plot for backfat thickness for SL1, SL3 and the combined data set (SL1+SL3).

5.3.2 Genomic relationships

The effects of the different weights for the SNPs in the G matrix were visualized in multidimensional scaling plots of the populations (Figure 5.4). In both lines, the use of weights that were estimated from the dataset of the other line (scenario 3) only slightly modified the projection of reference and validation populations. With weights that were estimated from the dataset of the same line (scenario 2) or weights estimated from the combined dataset (scenario 4), the projection of the animals became more dispersed.

5.3.2 GEBVs accuracy

In the first 4 scenarios, the reference and validation populations were from the same pig line. For these *within-population* scenarios, in a traditional GBLUP ($G^{Identity}$), where all markers had the same weight, the accuracies (Table 1) were 0.30 for SL1 and 0.31 for SL3. When weighted markers were used in the genomic relationship matrix, the accuracy for SL1 increased (0.32) when the weights were obtained from the dataset of the same line (G^{D1}) or from the combined dataset (G^{D_comb}), but the accuracy decreased when the weights were obtained from the dataset of the other line (G^{D3}). For SL3, a different pattern was observed. The accuracy increased when the weights were obtained from the dataset of the other line (G^{D1}) or from the combined dataset (G^{D_comb}), but the accuracy decreased when the weights were obtained from the dataset of the same line (G^{D3}).

In scenarios 5 to 8, half the animals in the reference set were replaced with animals from the other pig line. Thus, this *multi-population* prediction was conducted with 400 animals from each line. In scenarios 5 to 8, the reference population was always the same, and only the genomic relationship matrix was changed (Table 1). For SL1, adding weights to the markers increased the accuracy, and the highest accuracy was obtained with G^{D_comb} (0.32). This accuracy was equal to that obtained in scenario 4, where 800 animals from the same population were included, and the G^{D_comb} weighted matrix was used. For SL3, scenarios 6 (G^{D3}) and 8 (G^{D_comb}) showed increased accuracy compared to scenario 5 ($G^{Identity}$), but the accuracy was lower than the single-population predictions obtained in scenarios 1 to 4.

Scenarios 9 to 12 were also *multi-population* predictions, but the reference sets included an extra 800 animals from a different line, which doubled the reference set to 1600 animals (Table 5.1). For SL1, this increase in the reference population resulted in greater accuracies with all the G matrices compared to all the single-population predictions (scenarios 1-4).

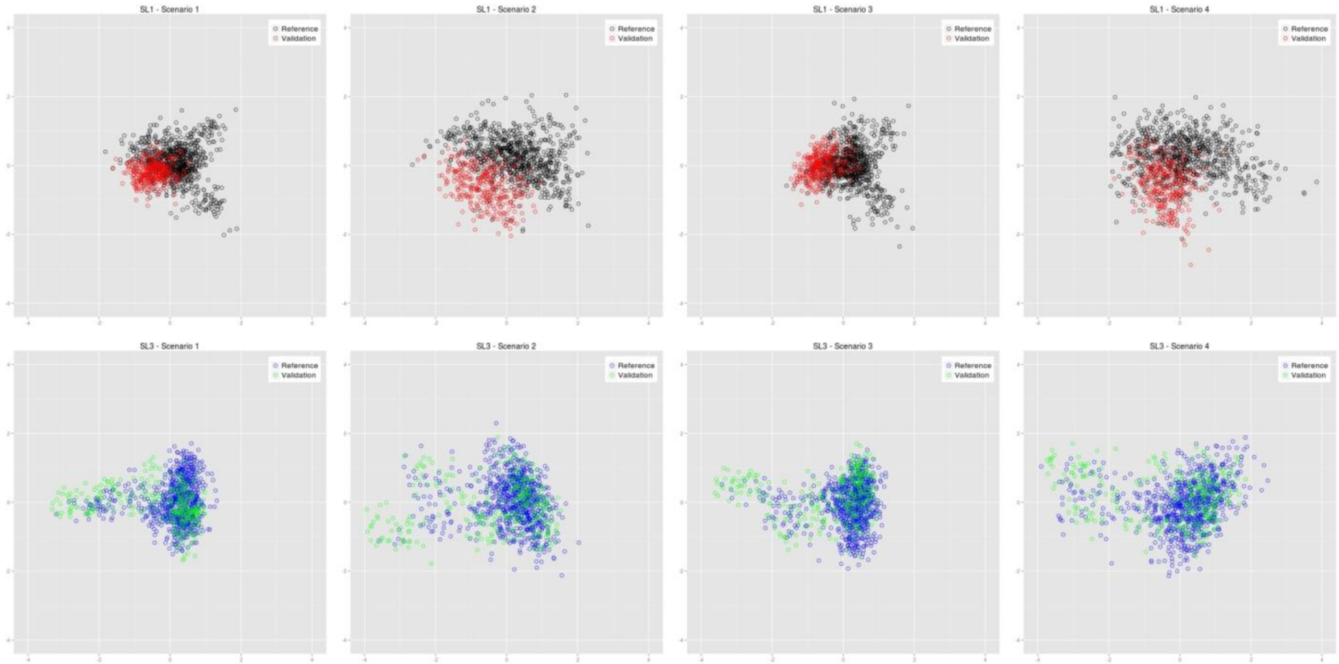

Figure 5.4 Multidimensional scaling plot showing a two dimensional projection of the populations SL1 (top row) and SL2 (bottom row) using different weighted relationships. Scenario 1 (column 1) correspond to traditional GBLUP, scenario 2 (column 2) the weights were obtained using the same population that was predicted, in scenario 3 (column 3) the weights were computed using a unrelated population and in scenario 4 (column 4) the weights were computed combining the two populations.

The use of G^{D_comb} (scenario 12) resulted in the highest accuracy (0.37); the next highest accuracy was achieved with $G^{identity}$ (scenario 9; 0.36). For SL3, the increase in the reference population only increased the accuracy compared to the single-population predictions with $G^{identity}$ (0.32 vs. 0.31) and G^{D3} (0.29 vs. 0.24), but the accuracy was reduced with G^{D1} (0.23 vs. 0.33) and it remained the same with G^{D_comb} (0.34 vs. 0.34).

Weighted markers affected the estimates of genomic heritability (h^2). The size of this effect was found to depend on which population was used to obtain the weights. For both lines, the h^2 was inflated in scenario 2, where a single population was used, and the weights were obtained from the dataset of the reference population. The lowest h^2 was observed when G^{D_comb} was used (Table 1). The slope coefficient of the regression of the corrected phenotypes on GEBVs was in most of the cases away from 1, indicating bias in the GEBVs, mainly when using G^{D1} and G^{D3}. The use of G^{D_comb} resulted in less biased predictions, in most scenarios, compared to predictions calculated with the other G matrices.

Table 5.1 Accuracy, genomic heritability (h²) and slope for backfat thickness in the scenarios evaluated.

Validation population	Scenarios	Reference population	Genomic relationship matrix	Reference population size	Accuracy	h²	Slope
SL1	1	SL1	$G^{Identity}$	800	0.30	0.49	0.89
	2	SL1	G^{D1}	800	0.32	0.76	0.47
	3	SL1	G^{D3}	800	0.29	0.45	0.87
	4	SL1	G^{D_comb}	800	0.32	0.40	0.88
	5	SL1+SL3	$G^{Identity}$	800	0.20	0.43	0.78
	6	SL1+SL3	G^{D1}	800	0.29	0.57	0.62
	7	SL1+SL3	G^{D3}	800	0.25	0.52	0.73
	8	SL1+SL3	G^{D_comb}	800	0.32	0.36	1.07
	9	SL1+SL3	$G^{Identity}$	1600	0.36	0.38	1.16
	10	SL1+SL3	G^{D1}	1600	0.33	0.53	0.58
	11	SL1+SL3	G^{D3}	1600	0.36	0.42	0.98
	12	SL1+SL3	G^{D_comb}	1600	0.37	0.30	1.09
SL3	1	SL3	$G^{Identity}$	800	0.31	0.40	1.22
	2	SL3	G^{D3}	800	0.24	0.65	0.47
	3	SL3	G^{D1}	800	0.33	0.42	1.27
	4	SL3	G^{D_comb}	800	0.34	0.37	1.12
	5	SL1+SL3	$G^{Identity}$	800	0.08^{ns}	0.43	0.30
	6	SL1+SL3	G^{D3}	800	0.15	0.52	0.32
	7	SL1+SL3	G^{D1}	800	0.04^{ns}	0.57	0.11
	8	SL1+SL3	G^{D_comb}	800	0.24	0.36	0.69
	9	SL1+SL3	$G^{Identity}$	1600	0.32	0.38	1.12
	10	SL1+SL3	G^{D3}	1600	0.29	0.42	0.64
	11	SL1+SL3	G^{D1}	1600	0.23	0.53	0.63
	12	SL1+SL3	G^{D_comb}	1600	0.34	0.30	1.10

5.4 Discussion

The Manhattan plots revealed differences in the SNP effects on backfat between the two populations evaluated. Differences observed in both the peak sizes and in the distributions indicated that the two lines had different genetic architectures. We explored whether these different genetic architectures affected genomic predictions. Extending on the methodology proposed by Wang et al. (2012), we computed weights for the SNPs, based on GWAS analyses. These weights were subsequently used to build different G matrices for the GBLUP analyses. GBLUP was then applied with either single- or multi-population reference sets. We found that, when the G matrix was built with weights based on GWAS information of the two populations combined (G^{D_comb}), the prediction accuracy was increased with both single and multiple reference populations. Moreover, in addition to using G^{D_comb}, doubling the multi-population datasets resulted in even higher prediction accuracies than when a single population was used.

5.4.1 SNP effects

The plots of the SNP effects on backfat thickness for SL1 and SL3 pigs showed that, despite the fact that the same trait was evaluated, the populations differed in the distribution and size of the effects. These divergences can be explained by differences between the two lines in LD, population history (initial variants, bottlenecks, and allele frequencies), and gene interactions. Previously, Veroneze et al. (2014) showed differences in the LD patterns of SL1 and SL3.

Results of a GWAS are difficult to replicate in different human populations, when there is heterogeneity in the LD structure across those populations (Li and Keating, 2014). Differences in allele frequency also represent a major challenge in replicating a GWAS result. When a variant that is common in one population is rare in another population, a larger sample size is needed to detect a significant association (Li and Keating, 2014). The challenges of the reproducibility of GWAS findings related to human obesity were evaluated by Li et al. (2013). They indicated that gene-gene or gene-environment interactions may introduce a challenge in replicating GWAS results in different populations.

In the present study, when SL1 and SL3 data were combined into one dataset, the SNP effects were, on average, greater than the effects observed in the single-population estimation. In a study on dairy cattle, Raven et al. (2014) suggested that a multi-breed GWAS resulted in more precise mapping of the QTL, due to the lower level of LD between markers, in comparison to single-breed LD. In a study on German Holstein cattle, Liu et al. (2011) showed that, when the number of reference bulls increased from 735 to 5025, the SNP with the largest effect showed a 4.13-fold increase in effect size. In addition to increasing the precision of finding the QTL peaks, the use of a larger number of animals in the multi-population GWAS might increase the statistical power of the analysis (Stranger et al., 2011).

5.4.2 Accuracy

The traditional GBLUP assumes that quantitative traits are controlled by a large number of genes that contribute equally to the trait (infinitesimal model); thus, the same variance is, *a priori*, attributed to all markers (Goddard, 2009). Nevertheless, it has been shown that a finite number of genes control quantitative traits (Hayes and Goddard, 2001); therefore, models that represent the true underlying genetic architecture of the trait may have higher accuracy than the GBLUP.

Two populations can exhibit distinct genetic architectures (QTL number, distribution, and effects), due to divergences in breeding goals or in the initial allele frequency. For example, allele substitution in the *DGAT1* locus caused different effects in Jersey and Holstein-Friesian populations in New Zealand (Spelman et al., 2002) and between Fleckvieh and Holstein-Friesian populations in Germany (Thaller et al., 2004). Hence, differences in genetic architecture between populations could be exploited by emphasizing the markers that explain more genetic variance in the target population. This notion introduces the possibility of using prior knowledge of genetic architecture for making single- and multi-population predictions.

In the present study, we evaluated three strategies for computing weights of markers to be used in the GBLUP: (*i*) the weights were obtained using the same population that was predicted, (*ii*) the weights were obtained in an unrelated population, and (*iii*) the weights were obtained in the two populations combined.

When the weights were obtained using the same population as the reference set (scenario 2), the prediction accuracy increased for SL1, but decreased for SL3. It has been shown that the accuracy of genomic prediction can be improved by reducing distances between the reference and validation animals and by increasing distances between animals within the reference population (Pszczola et al., 2012). In the present study, for both pig lines, including the marker weights resulted in increasing the distances between animals within the reference population (Figure 4). This effect was more pronounced for SL1 than for SL3 pigs. However, before the inclusion of weights, compared to SL1, the SL3 pigs showed greater distances between individuals within the reference set and smaller distance between the validation and reference animals. These initial differences in the distances between the SL1 and SL3 groups may be one cause for the different changes in prediction accuracies when marker weights were included in the matrix. The use of weights obtained in an unrelated population (scenario 3) increased the accuracy of predicted GEBVs for SL3. This finding is important, because it suggests that, if a large number of genotype animals is available in the other line, conducting a GWAS in that line will have greater power and result in better SNP effects estimates which leads to better marker weights for the smaller population.

Indeed, we found that using the G^{D_comb} resulted in higher prediction accuracy than the traditional GBLUP ($G^{Identity}$) for both pig lines, in single- and multi-population scenarios. This improvement could be attributed to a better estimation of the marker effects, due to the large number of animals. Moreover, when two lines are pooled, the LD is reduced, and this increases the QTL mapping resolution. Thus, in the combined dataset, the SNPs closest to the QTL would be identified, and consequently, they would be better able to track the effects of interest, both within and across populations.

Previous studies (Zhang et al., 2010; Zhang et al., 2014; Tiezzi and Maltecca, 2015) have indicated that the accuracy gained by using weighted relationships matrices depended on how the trait was controlled; they showed that more efficiency was likely to be gained for traits controlled by small number of QTLs. The trait we evaluated was backfat thickness, which is apparently controlled by a large number of QTLs. Therefore, our results may not have been favored by the trait analyzed.

With dairy cattle and rice data, Zhang et al. (2014) also incorporated GWAS results in genomic predictions by adding weights for the markers. This strategy increased the prediction accuracy for two of three traits in dairy cattle and for nine of 11 traits in rice. In a study on Holstein cattle, Tiezzi and Maltecca (2015) concluded that weighted relationship matrices yielded more accuracy and less bias in predictions for traits regulated by a few QTLs.

In scenarios 9-12, we added extra animals from an unrelated population to the reference set. With this approach, even when all markers were equally weighted (scenario 9), the accuracy was increased compared to that of the single-population prediction; the highest accuracies were obtained with G^{D_comb} (scenario 12) for both populations. Chen et al. (2014) found that, with pooled data, the accuracy of genomic prediction may be reduced when the analysis used weakly correlated QTL effects or a relatively low-density SNP panel. In our study, the SNP effects of both lines were very weakly correlated; therefore, for a different trait with highly correlated SNP effects across lines, a large increase in prediction accuracy might be found with this approach. Alternatively, the prediction accuracy might be further improved for the same trait by using a higher-density SNP panel. The SNP panel used in the present study did not show a consistent LD phase across the evaluated lines (Veroneze et al., 2014).

This was the first study to compute weights based on GWAS results from combined cohorts, and then, to use them in multi-population predictions. The methodology presented here should be evaluated with additional traits and populations. Nevertheless, our approach made it possible to include genotyped and phenotyped individuals from multiple populations, and it emphasized the similarities between the populations. In addition, the method for computing weights in the multi-population approach addressed two important differences between groups that can affect the accuracy of genomic predictions; (i) allele frequency and (ii) genetic architecture.

5.5 Conclusion

We found that the highest accuracy of genomic predictions was achieved by using weighted markers, where the weights were based on GWAS results from a combined population analysis. We also found that using a multi-population reference set resulted in greater prediction accuracies than using single-population reference set, both when performing the GBLUP without weights and when performing the GBLUP with weights based on the GWAS results from the combined population.

6

General discussion

6.1 Introduction

Inheritance, genetic control, how genes produce phenotypes, and genetic variation are some of the subjects in genetics that have raised our curiosity, and humans have endeavored to understand these genetic phenomena throughout the ages. Great effort has also been made in determining how to apply this accumulated knowledge to medicine and food production. Animal breeding is an example of the successful combination of applied knowledge about genetics and statistical analysis. This success has resulted in a large increase in livestock productivity. However, many scientists have pursued the goal of opening the "black box" of genetics to examine factors that underlie the results from practical animal breeding. A better comprehension of these genetic mechanisms might contribute to improvements in breeding programs.

The rapid development of molecular technologies has generated databases of genomic sequences and a large number of molecular markers. These types of data have opened a new window that can shed light onto the comprehension of genetic selection and breeding, and thus, scientists are closer to opening the "black box". The initial challenge was to combine basic molecular genetics knowledge and statistics and incorporate them into a practical application, as envisioned by Meuwissen et al. (2001). The latter group introduced the concept of using dense molecular marker panels to predict breeding values in an approach now referred to as genomic selection. More recently, this concept has also contributed to human genetics studies, and it has been used in genome wide association studies (GWAS). However, several challenges remain to be overcome, and knowledge gaps must be bridged to create practical, feasible applications of this new technology. In this thesis, I studied some of these challenges, starting with the levels and measures of linkage disequilibrium (LD; Chapters 2 and 3). LD is a basic genetic concept that affects the efficiency of genomic selection. The results from chapters 2 and 3 were used in subsequent chapters to propose practical strategies and methods for implementing genomic selection in small populations. I then applied and tested the effectiveness of these methodologies in data from a pig breeding program (Chapters 4 and 5).

In this chapter, I discuss the results of my thesis in a broader context. I will discuss practical and theoretical aspects of LD and its measures in the computation of effective population size, in GWAS, and in breeding. Next, I will discuss the application of genomic selection in small populations and in pig breeding. Finally, I discuss the utility and prospects of using whole genome sequence (WGS) data in genomic prediction.

6.2 Linkage disequilibrium in breeding and population effective size computation

The availability of dense marker panels for livestock resulted in a large number of studies that explored molecular markers in diverse branches of genetics. Among other factors, marker panels have contributed to a better comprehension of the demographic history of different species, the genetic architecture of traits, and the different aspects that influence genetic selection. LD comes into play in these three contexts at different levels.

6.2.1 Population effective size

The population history, breeding system, and pattern of geographic subdivision are reflected in the LD throughout the genome (Slatkin, 2008). Thus, LD can be used to investigate the demographic history and dynamics of a given population (Li and Merilä, 2010). According to Hill (1981), LD between markers that are in close proximity reflects the ancient population history, and LD between markers separated by large distances reflects recent history. Demographic events, such as population bottlenecks, migration, and expansions, are reflected in the historical effective population size (Ne).

Effective population size is a crucial measure in population management, and it is frequently used in animal breeding. Effective population size affects the genetic variability, the effectiveness of selection relative to drift, inbreeding, and the long-term survival of conserved populations. When studying a single population, the Ne can be estimated from the LD (r^2) without pedigree information. This approach is advantageous, because pedigree information is not always available or may not be accurate. Sved (1971) derived an equation based on the theory of genetic drift and recombination that has become the most popular method for estimating Ne from LD. The equation assumes an isolated population with random mating and constant population size.

Most models that predict LD assume that all loci pairs in close proximity show complete LD, due to the lack of recombination over short intervals. However, previous studies (Ardlie et al., 2001), and our work described in Chapter 3 of this thesis, have shown that a significant number of marker pairs that are close together show incomplete LD. This lack of complete LD between markers that are close together has been attributed to gene conversion (Ardlie et al., 2001; Wall, 2001). Gene conversion is defined as the "recombinational transfer of information between alleles or loci without reciprocal exchange of DNA (i.e., without crossover)" (Jeffreys and May, 2004). Studying the hot spot DNA3 located in the major histocompatibility complex (MHC), Jeffreys and May (2004) suggested that gene conversion was 4 to 15 times more frequent than crossover.

Gene conversion is important for breaking up the LD between closely linked sites, but it has a negligible effect on more distant sites (Ardlie et al., 2001). This principle is consistent with the results described in Chapter 3. There, the deterministic model, which assumed non-recombination for closely-located sites, showed a poor fit for sites short distances apart, but good predictions of LD for distant sites. Within short distances, LD cannot be predicted precisely with deterministic equations based only on crossovers, because those equations force the LD to assume a certain (high) value at short distances.

Figure 1 of Chapter 3 shows the observed LD fitted with a deterministic equation. For short marker separations, the observed LD values ranged from 0 to 1, but the equation always predicted higher values than those observed. A consequence of this mismatch between observed and predicted LD was that the Ne of the population many generations in the past cannot be precisely estimated based on LD of markers separated by short distances. Frisse et al. (2001) showed that LD data was not compatible with the assumption that gene conversion was absent; when gene conversion was included in the modeling, the estimation of Ne was more consistent.

In Chapter 3, I proposed a new statistical approach (LOESS regression) for modeling the LD. This approach resulted in better predictions of LD for markers separated by short distances. The model allows the functional form of dependent and independent variables to be determined by the data, without requiring parameters to give shape to the curve. Although the LD predictions were improved with this method, they were not always precise for individual pairs of SNPs separated by short distances. Thus, for a given distance between marker pairs, the LD might be either complete or incomplete for different pairs, even when they were in the same region of the genome. Moreover, although the method proposed in Chapter 3 improved the prediction of LD, the absence of equation parameters precluded its use for predicting Ne. Thus, further study is necessary to develop a method that combines the flexibility of the LOESS regression and the ability to estimate the Ne based on LD.

6.2.2 Linkage disequilibrium and genome wide association

Identifying key genes, knowing the contribution of each allele to the desired trait, and quantifying gene-gene interactions would enhance our understanding of complex traits and also open the possibility to directly select animals for the desirable (combination of) alleles (Meuwissen and Goddard, 1996). GWAS have been widely used to search for genes that give rise to variation in complex traits (Goddard and Hayes, 2009). The GWAS approach is indirect, because the causal variants for a given trait are either unknown, or they are generally not included in the genotyping panel. Therefore, GWAS identify markers that track these variants through association. This approach is only successful when the marker genotypes are highly correlated (in LD) with the causal variants (Meuwissen et al., 2001). As the LD is reduced between markers and the causal variants, the chances are diminished for detecting a marker that tracks these variants. Thus, the existing level of LD in a population provides a basis for determining the number of markers that should be used to detect the associations of interest. Determining the number of markers to use in a panel is a first, simple, practical application of knowledge about the LD in a given population.

In Chapter 2, I studied the LD levels of eight pig populations (pure lines and crossbreds). I found that all the lines exhibited an average LD (r^2) greater than 0.3 for markers that were 100 to 150 kb apart. This level of LD is considered sufficient for association studies (Ardlie et al., 2002; Du et al., 2007). Thus, the available marker panel (Porcine SNP60 BeadChip), which contains approximately 65,000 markers, should be sufficiently dense for whole genome studies in these lines and in the progeny from crossbreeding these lines.

In Chapters 2 and 3, we showed that pigs exhibited high LD levels, and also that the LD extended over long distances (Veroneze et al., 2013) but there I did not discuss the implications of these findings for genome wide association. When the population exhibits LD over large distances, it is possible to detect associations between markers and QTLs and/or carry out genomic selection, even with a relatively low-density SNP panel. However, the mapping resolution is lower when the LD extends over large distances (Fu et al., 2015). In addition, the identification of causal mutations may become challenging, because multiple markers that span a large region of the genome may be associated with the causal mutation. Thus, in livestock, the use of high-density marker panels may not benefit the identification of causal mutations, due to high LD extension. The extent of LD between populations and the sizes of haplotype segments that are shared between populations, on average, are much shorter than the same measures evaluated in a single population. Therefore, it should be possible to increase the mapping resolution in livestock by combining more than one population in the same analysis (multi-population) to reduce the extent of LD.

6.2.3 Linkage disequilibrium and breeding

In general, more than one breed is used for production in livestock. Thus, a marker that tracks a large effect in one population would be doubly valuable if it could be used similarly in another population. However, even when a marker segregates in both populations, the effect that it tracks may not be equal between the populations. Differences in QTL effects that are tracked by a marker across different populations is an issue in GWAS, and also, in genomic selection. In genomic selection, a lack of consistent marker-trait associations may result in limited accuracy in cross- or multi-population approaches.

In Chapter 2, I showed that the persistence of LD phase was high between the pig crossbreds and the parental lines studied. This high persistence indicated that the marker effects might be expected to be similar across these populations. Across the pure pig lines, the persistence of phase was low; thus, high-density panels should be used for adequate marker-QTL associations across these lines. Based on the findings in Chapter 2, we predicted that it would be necessary to genotype between 56,000 and 280,000 segregating markers to achieve a similar marker-QTL association across different pairs of lines.

Although the 60K marker panel resulted in high LD phase persistence between crossbred and parental lines, the accuracy of using pure lines to predict a trait in the crossbred population was much lower than the accuracy of within-population predictions. In addition, I concluded that the 60K marker panel was not sufficiently dense to maintain the same phase across pure lines. Nevertheless, I expected that pairs of lines that exhibited high phase persistence (Chapter 2) would produce high accuracies in genomic predictions based on either a across population reference set (composed of animals unrelated to the validation set of animals) or a multi-population reference set (more than one population in the reference set). However, the accuracies of the genomic predictions obtained in Chapter 4 for across- and multi-populations did not fulfill this expectation.

Those results suggested that the LD phase may not be as important for genomic prediction accuracy as previously thought. Thus, the interplay between LD, genetic architecture, and allele frequencies must be better understood to improve the accuracies of predictions, when more than one population is involved. Disentangling this complex interplay could give insights into the factors that challenge the accuracy of genomic predictions performed using across- or multi-populations strategies. These same challenging factors may play a role, albeit smaller, across generations of the same population. To study these factors, it is necessary to have access to a database with a large number of genotypes and phenotypes from multiple populations and traits. The first step of such a study would be to conduct a GWAS for each population to identify similarities and differences in the significant markers. The second step would be to analyze the frequencies of these markers and the LD in the region. Third, cross-population genomic predictions should be conducted to check the impact of the different factors.

6.3 Challenges of genomic selection in small populations

The use of genomic information increases the genetic gain by reducing the generation interval and increasing the prediction accuracy for young animals (Hayes et al., 2009). The first description of genomic selection (Meuwissen et al., 2001) rapidly led to applications in Holstein cattle, and currently, most major breeding companies have implemented genomic selection. Although genomic selection is used in practice, applying genomic selection has been challenging in many livestock populations. One difficulty is the size of the reference population, because among other factors, the accuracy of genomic prediction depends on the number of animals that are available for the reference set (Goddard and Hayes, 2007).

Estimated GEBVs were reported by Interbull (http://www.interbull.org/), based on a multiple-trait, international evaluation of 200,285 bulls of six breeds in April, 2014. The evaluation showed that there was a preponderance of the Holstein Friesian breed and the other breeds were considerably smaller. The Holstein Friesian breed comprised 135,646 bulls (67.7% of the total), and the Guernsey breed comprised 1,060 bulls, which was the lowest number evaluated (0.53% of the total). Moreover, the number of breeds used for milk production worldwide is much larger than the six breeds evaluated by Interbull, and many of these breeds comprise small populations.

In pig and chicken breeding programs, selection is carried in nucleus populations. The companies maintain specialized dam and sire lines to produce crosses for commercial production. Thus, the purebred populations are small, and the companies do not exchange data on populations, even when they are derived from the same breed.

The concept of a “small” population is relative to the to the cost:benefit of the genotyping. Among dairy cattle, bulls are typically used in the reference population, which comprises animals with highly accurate breeding values based on offspring performance. Among pigs and chickens, the reference animals typically have own performance and limited information is available on relatives. In addition, the generation interval for pigs and chickens are relatively short; thus, genomic selection provides a much smaller increase in genetic gain per year than the increase achieved in selecting cattle. The shorter generation interval of pigs implies that the reference population may require more frequent renewal.

Even in populations with large numbers of animals, the size of a reference population may be restricted by the number of phenotypes available for some traits. For example, some traits are expensive to measure (e.g., residual feed intake), and others can only be measured in small trials (e.g., disease resistance).

Therefore, a small reference population can potentially be encountered in all livestock species, and it constitutes a challenge for genomic prediction. This challenge calls for effective reference population design. One possibility is to include across- and/or multi-populations strategies to improve prediction accuracy.

6.3.1 Reference population design

Genomic selection is expected to provide higher genetic gain than traditional selection; thus, genomic selection is an important methodology for breeding companies. For small populations, their competitive position and market share might be affected by the implementation of new technologies over the long term. Recently, a number of studies have developed methodologies and strategies for implementing genomic selection in small populations (Gaspa et al., 2014; Hozé et al., 2014; Riggio et al., 2014; Thomasen et al., 2014).

In dairy cattle, the genotyped animals are typically the sires. However, when the available number of sires is insufficient, adding females to the genomic evaluation may a good option for achieving higher accuracy and lower inbreeding rates (Jiménez-Montero et al., 2012; Thomasen et al., 2014). Jiménez-Montero et al., (2012) showed that the prediction accuracies increased when the additional females in the reference set exhibited the upper and lower extremes of the trait distribution.

When animals have phenotypes, but genotyping is cost prohibitive, a low-cost alternative is the imputation of genotypes in non-genotyped animals. The animals to be genotyped should be strategically chosen to facilitate accurate imputations. For example, imputation accuracy can be considerably increased when the non-genotyped individuals have genotyped offspring; in that case, accuracies above 0.90 can be achieved, when four offspring are available (Bouwman et al., 2014). However, the drawback of genotyping multiple offspring from the same animal (siblings) is that many individuals among the group of genotyped animals will be highly related. Close relationships in the reference set might not be a good strategy, because higher accuracy is achieved when the relationships within the reference set are minimized, and relationships between reference and selection candidates are maximized (Pszczola et al., 2012).

Among pigs and poultry, parents can be selected to optimize the accuracy of genomic prediction. Instead of choosing the top sires and dams, the selection intensity could be somewhat reduced to gain genomic prediction accuracy by minimizing relationships in the reference set and maximizing relationships between reference and validation animals. Selection intensity and prediction accuracy both contribute to genetic gain. These strategies must be evaluated and optimized to attain the maximum level of genetic gain.

6.3.2 Across-population prediction

Originally, it was thought that a large population (like the Holstein breed in dairy cattle) could be used to estimate marker effects, and that subsequently, these effects could be used to select candidates in a smaller population, such as the Guernsey breed. In this thesis, this procedure is called across-population prediction. Due to the lack of pedigree relationships between animals of different populations, these predictions were based on LD between markers and QTL(Daetwyler et al., 2012). Several studies have since shown that across-population predictions led to accuracies near zero (Hayes et al., 2009; Kachman et al., 2013; Riggio et al., 2014).

Pig breeding organizations typically maintain more than one line, and some lines have a common genetic background. In general, these pig lines have diverged more recently than the time since breeds were formed. Therefore, the haplotype sharing across lines derived from the same original breed was expected to be higher than across breeds, due to the smaller number of recombinations since the last common ancestor. LD was shown to extend over long genomic distances in commercial pig lines (Veroneze et al., 2013). In Chapter 2, I showed that the persistence of LD phase across purebred populations was higher between populations with a common genetic background than between populations with divergent backgrounds. Therefore, the accuracy of across-population predictions was expected to be higher in pigs than in cattle. In Chapter 4, I evaluated the accuracy of across-population predictions with the sire lines studied in Chapter 2. However, we did not observe the expected higher accuracy with across-population predictions between populations with high LD phase persistence.

Based on high (700 K) and medium (50 K) density SNP panels, Erbe et al. (2012) concluded that a small, non-significant advantage in the predictions was observed when the high panel was used for across-population predictions in Holstein and Jersey breeds. In my results in pigs, and in the dairy cattle results from Erbe et al. (2012), the persistence of LD was presumably high, but that did not lead to the expected improvement in accuracy. Therefore, increasing the density of genotypes was not the solution for improving across-population predictions. In other words, not having the same LD phase in both populations was only one factor that hampered the accuracy of genomic prediction; other factors include differences in genetic architecture and allelic frequencies.

6.3.3 Multi-population prediction

In addition to across-population prediction, another approach for genomic selection in small populations could be to combine more than one population in a single reference set. In this thesis, this approach is called multi-population prediction. This strategy was previously evaluated, again mainly in cattle, and the general conclusion was that the accuracy was highly dependent on the genetic distance between the populations (Lund et al., 2014). Combining populations will decrease the relationship between the reference and validation sets; thus, the prediction will rely more on the LD instead of on family relationships. In theory, the accuracy of multi-population prediction is expected to increase when the populations exhibit the same LD phase (de Roos et al., 2009). This condition is achieved by using a high-density panel.

In Chapter 2, the study on phase persistence showed that different degrees of relatedness exist between different pairs of lines, and as expected, lines that shared common genetic background were more related to each other. A previous cattle study that showed that the accuracy of multi-population predictions depended on the genetic distance between populations. Based on those results, we expected that combining pig lines that were more related would produce the highest prediction accuracy. However, when we evaluated the multi-population predictions in Chapter 4, accuracy was only increased for one of the three evaluated sire lines when animals from another line were added to the reference data. This difference in prediction accuracies for the different lines could not be explained by different levels of relationship or persistence of phase between populations. Although the SNP panel used in Chapter 4 was not sufficiently dense to provide a consistent LD phase across the populations evaluated (Chapter 2), I expected that the accuracies for a multi-population prediction would reflect the degree of relatedness between lines that was observed in the SNP data.

The accuracy of multi-population predictions in Holstein and Jersey breeds did not significantly increase with the use of a high-density SNP panel (700K) (Erbe et al., 2012). This lack of improved accuracy was probably due to the fact that the LD phase is only one of multiple factors that influence the accuracy of multi-population predictions. Multi-population genomic prediction is also influenced by the allele frequencies of the SNP and QTL, and by genetic architecture. In addition, across-, and multi-population predictions are more prone to be affected by environmental interactions with the genotype (QTL) than single-population predictions, because individuals from different populations are more likely to experience different environments, such as feeding levels and climate conditions.

In Chapter 4, we showed differences in the genetic architecture of backfat thickness across populations, and we suggested that these differences might influence the effectiveness of multi-population predictions. This hypothesis was tested in Chapter 5, where marker weights were computed, based on the proportion of variance explained by the SNPs. These weights were then used in GBLUP analyses for multi-population predictions. With this strategy, we expected that the marker weights would improve the model representation of the specific genetic architecture of the trait.

We found that prediction accuracies were either increased or unaffected by the use of marker weights obtained from a combined GWAS (both lines combined). Those results indicated that the use of weights could benefit predictions based on multi-population datasets. They also showed that weights obtained from a different GWAS analysis (a separate line) could result in divergent outcomes, depending on which lines were analyzed when the weights were applied. Only weights derived from the combined GWAS were beneficial in all prediction scenarios. This benefit could be attributed to the use of a larger number of animals to estimate the SNP effects. However, more importantly, the consistent improvement observed in both lines could be due to the enhanced ability of the GWAS model to capture the genetic architecture of the trait in the combined dataset. This was consistent with the finding that QTL mapping precision was improved in multi-population datasets, due to the reduction in LD. As explained in Chapter 5, Raven et al. (2014) showed that multi-breed genome wide association analyses could more accurately pinpoint the locations of well-described mutations that affected milk production, such as DGAT1. Therefore, the marker weights estimated with multi-population GWAS also appeared to benefit from better identification of the regions that affected the trait.

In Chapter 5, I implemented an approach that considered the differences in genetic architecture together with differences in allele frequency in multi-population predictions. This methodology, which included weights for markers, could also be used to incorporate other information on the genetic architecture of traits, such as results from a post-GWAS analysis or the proximity of candidate genes. The methodology may also be extended, in a straightforward manner, by combining more than two populations in the GWAS analysis and in the genomic prediction.

6.4 Genomic selection in pig breeding

The production of pork is based on a pyramidal structure, which comprises three tiers: the nucleus, the multipliers, and the commercial producers. The breeding and selection are conducted in the nucleus. Multiple populations of sire and dam lines are maintained with high health status in the nucleus. The genetic improvements achieved in the nucleus are passed to the commercial level, generally through a multiplier phase that introduces a genetic lag of 3 to 5 years. In the commercial herds, heterosis is exploited by crossing F1 sows with boars, typically to produce three-way crosses.

6.4.1 The use of crossbred data

The correlation between performance of the pure lines in the nucleus and that of the crossbreds in commercial herds may be reduced below 1.0, due to genotype interacting with the environment and genotype-genotype interactions. In a study on two pig lines, Lutaaya et al. (2001) showed that the correlation between purebred and crossbred performance was population-specific; the correlation varied from 0.62 to 0.99 for lifetime daily gain and from 0.32 to 0.70 for backfat thickness. A low correlation between pure breed and crossbred performance implies that only part of the genetic improvement achieved in the nucleus persisted as an increase in performance at the commercial level. To improve selection for the expression of traits at the commercial level, phenotypic data must be collected in commercial farms. In particular, traits related to disease resistance or tolerance are not available at the nucleus level.

As an alternative to selection based on performance in the pure lines, it is possible to use data from crossbred offspring to predict the genetic value of pure breeds for crossbred performance. Using the traditional BLUP, this approach was predicted to produce greater performance in crossbreds (Bijma and van Arendonk, 1998) and to lead to higher inbreeding (Bijma et al., 2001). Dekkers (2007) showed that, with markers, the selection of pure breeds for crossbred performance would result in greater selection success and less inbreeding, without the need to collect pedigree information.

Simulation studies have demonstrated that data from crossbred animals could be used to select purebred animals with a genomic selection approach (Ibánez-Escriche et al., 2009; Toosi et al., 2010). In addition, Zeng et al.(2013) showed that the dominance model was superior to the additive model and the breed-specific allele model for selecting pure breeds for crossbred performance. There is a lack of studies evaluating these simulation-based findings in real data because a large number of crossbred animals with genotype as well as phenotype information is typically not (yet) available. Genomic selection of purebreds for crossbred performance is expected to be successful, because crossbred animals are typically closely related to their pure line ancestors.

In Chapter 2, high LD phase persistence was shown between crossbreds and their pure line ancestors for the SNPs on the current pig 60K SNP panel. In Chapter 4, I showed that the relationships between purebreds and crossbreds were higher than between different purebreds.

However, although the amount of crossbred data available was sufficient for the study conducted in Chapter 2, it was insufficient to be used as a reference set for genomic prediction. To test predictions across purebred and crossbred animals, I used purebred animals as the reference set to predict the GEBVs of crossbreds. I found similar prediction accuracies for the three different crossbred populations analyzed (range: 0.25 to 0.29), which may reflect the fact that the relationships of the sire line with the crossbreds were the same for the three combinations evaluated.

Unfortunately, it was not clear to what extent the accuracies found in Chapter 4 reflected the efficiency of the selection of purebreds for crossbred performance. This uncertainty arose from the analysis in Chapter 4, which showed the puzzling result that the accuracy achieved by using purebred population A to predict purebred population B was different from the accuracy achieved with the reverse prediction. We expect similar challenges to arise when crossbreds are used to predict purebred performance. I suggest that these divergences may be caused by differences in the allele frequencies of the SNPs that tracked the QTL of interest. When marker frequency is low in population A, but high in population B, the marker will not show the same power of prediction in both directions; thus, the results may show asymmetrical accuracy in genomic predictions.

Because prediction accuracies observed in real data do not always follow expectations that are based on LD consistency or relatedness between populations, it is crucial for pig breeders to evaluate the genomic selection methodologies proposed for use with crossbred data. The amount of gain achieved in crossbred performance will dictate how genomic selection might be implemented in pig breeding. Phenotype data are not commonly collected in farms with crossbred animals, and collection might be challenging to implement. Before implementing a data collection scheme, the cost:benefit ratio must be somewhat clear. Most likely, this cost:benefit ratio will depend on the trait evaluated; the most valuable ratios are likely to be found for traits with low correlations between purebred and crossbred performance and for traits that cannot be measured in the nucleus, but that have a big impact in pig production (i.e., disease resistance).

6.4.2 Optimization of genomic selection

The implementation of genomic selection in cattle breeding had the advantages of substantially reducing the generation interval and of the availability of highly informative individuals (progeny-tested bulls). These advantages makes the relative cost of implementing genomic selection in cattle lower than in pig breeding, because the generation intervals are inherently short, and an individual animal is not highly informative (own performance).

In addition, the challenge of applying genomic selection to pig breeding is compounded by the fact that pig breeders maintain a range of purebred populations. Therefore, a large number of animals must be genotyped and phenotyped for genomic selection implementation in each population separately. Consequently, the cost:benefit of genomic selection must be optimized in pigs. Optimization is possible by using low-density SNP panels, selective genotyping (Van Eenennaam et al., 2014), multi-population reference sets, and implementing the single-step GBLUP (ssGBLUP), which combines data from genotyped and non-genotyped animals (Legarra et al., 2009).

In commercial pig lines, LD extends over long distances (Veroneze et al., 2013), which is beneficial for imputation. Consequently, the use of low-density panels is highly efficient in these lines. Imputation accuracy in pigs has been tested with panels of 384, 450, 768, 3000, and 6000 markers (Huang et al., 2012; Cleveland and Hickey, 2013; Wellmann et al., 2013). Genotyped ancestors were shown to be important for obtaining low imputation errors with very low-density panels (384 and 450 markers) (Huang et al., 2012; Cleveland and Hickey, 2013). With 450 markers, Cleveland and Hickey (2013) obtained imputation accuracies of 0.963 with data from only 359 individuals that were parents and grandparents of the test individuals. However, they also showed that the imputation accuracy was not constant across the genome, and the GEBVs captured less information with this low-density panel compared to a higher-density panel. Huang et al. (2012) recommended an optimized, low-cost genotyping strategy, where male parents were genotyped at high-density, female parents were genotyped at low-density (3000 markers), and selection candidates were genotyped at very low-density (384 markers). However, the authors did not evaluate the impact of their strategy on GEBV accuracy.

The design of the low-density panel (determining which SNPs to include) requires careful consideration, because the LD between markers is not constant across the genome (Chapter 3); this inconsistency could cause differences in imputation accuracy in different parts of the genome. The distribution of haplotype blocks across several pig chromosomes revealed that regions with low and high LD existed across the genome (Veroneze et al., 2013), and these patterns were different across the six pig lines evaluated. Thus, ideally, the low-density panel should be designed for each population separately to reduce fluctuations in imputation accuracy across the genome. Regions important for the trait under selection with low imputation accuracy impact the GEBV prediction. Therefore, markers that are recognized to impact the important traits in pig breeding should always be included in the panel.

As discussed for small populations, selecting appropriate individuals for genotyping is also important for maximizing the gain that can be achieved by implementing genomic selection. Lillehammer et al. (2011) showed that, for the selection of maternal traits, females should be genotyped in addition to the boars tested. Another strategy expected to produce better accuracies is to maximize the relationships between the reference population and the selection candidates, and simultaneously, minimize the relationships within the reference set (Pszczola et al., 2012). Although genomic selection has been implemented in pigs, the best genotyping strategy is not immediately clear, because the published studies have focused on different specific questions, such as imputation accuracy or relatedness between animals. Studies should be performed that integrate the use of low-density panels, genotyping strategies, and multi-population genomic predictions in sire and dam lines, because for practical implementations, all these strategies must be combined. It is necessary to understand how these strategies interact to optimize their combined use in genomic selection.

Currently, as an intermediate solution, ssGLBUP (Legarra et al., 2009) seems to be the first choice for most breeding companies, because it provides increases in prediction accuracy by including both genotyped and non-genotyped animals. In addition, this method is easily integrated into the traditional pig breeding workflow. Because ssGBLUP allows the use of both genotyped and non-genotyped animals at the same time, it can be implemented during the time needed to acquire a sufficient number of genotyped animals. To date, the ssGBLUP has been applied in dairy cattle, pigs, and chickens (Misztal et al., 2013). The use of ssGBLUP predictions in pigs improved the prediction accuracy for all the animals evaluated (genotyped and non-genotyped) (Forni et al., 2011; Ibáñez-Escriche et al., 2014). Another advantage of ssGBLUP is that it can be easily implemented in more complicated models, such as the multi-trait or threshold models, because it uses the BLUP methodology, but with a different relationship matrix.

In the near future, advances in the model will be essential, because the number of relevant markers is increasing. Therefore, the assumed infinitesimal model might not be the best method for exploiting the information that can be captured with the new high-density panels and genome sequence data. In addition, dominance and epistasis are also components of genetic variation, and the use of markers in genomic selection is opening up new possibilities for exploring and incorporating these effects. Previously, these effects could not be implemented in ssGLUP. Thus, the ssGBLUP must also be improved if it is to remain the methodology of choice for implementing genomic selection.

6.5 Whole genome sequence data in genomic prediction

The ultimate SNP density will be achievable when the whole genome sequence (WGS) becomes available for all animals. This density would not rely on capturing all QTL through LD, but would include all QTL that are based on SNP variants within the dataset. Currently, experience with WGS datasets is limited, but Ober et al. (2012) reported a study performed with the genomic sequence data from 157 inbred lines of *Drosophila melanogaster*. They found little or no gain in accuracy when the number of SNPs was increased above 14.6 *NeL* (equivalent to 43,000 markers in Holstein cattle). Although the sample size in their study was small, and the population structure (unrelated inbred lines and highly accurate phenotypes) was different from that used in livestock, their results may indicate the utility of sequenced-based prediction in non-model organisms. In chickens, Heidaritabar et al. (2015) found that the use of WGS data increased the prediction accuracy by only 1% in a GBLUP. When data from Jersey and Holstein cattle were used to predict the GEBV of a Holstein population, Hayes et al. (2014) found that using the WGS increased the prediction accuracy by 2% compared to predictions based on 800 K genotypes.

To date, using the WGS for genomic prediction has resulted in only small increases in the accuracy of single- and multi-population predictions. In single populations, the prediction relies largely on family structure (Clark et al., 2012); consequently, increasing the LD by using the WGS would not be expected to increase the prediction accuracy. However, in across- and multi-population predictions, LD plays a major role. Due to the complex interplay of the factors that control prediction accuracy, the problem of predicting across populations cannot be resolved by merely increasing the number of markers, without changing the models. Consideration must be given to factors that might hamper the accuracy in multi-population strategies, such as differences in genetic architecture and differences in allelic frequencies.

Using WGS data with the currently available models (GBLUP, BayesR, BayesRC) does not seem to be a promising approach. Moreover, it is not clear how to exploit WGS data in a practical breeding application. Nevertheless, WGS data represent a valuable information source for attaining a better comprehension of complex trait architecture. This knowledge can be used to select the best markers to include in SNP panels, based on their effects on important traits. Another advantage of WGS data is that they facilitate studies on other sources of genetic diversity, such as copy number variations (CNV), which have been reported to affect important traits in cattle (Xu et al., 2014; Yue et al., 2014).

In pigs, breeding values are estimated every week for multiple traits and lines with thousands of animals. The incorporation of WGS data in this system would require more complex models, more computation capacity, and more time. Given these difficulties in implementation, combined with the low benefit derived from using WGS for genomic predictions, it is not likely that WGS will be used directly in pig breeding value estimations in the near future.

References

References

Abasht, B., E. Sandford, J. Arango, P. Settar, J. E. Fulton, N. P. O. Sullivan, A. Hassen, D. Habier, R. L. Fernando, J. C. M. Dekkers, and S. J. Lamont. 2009. Extent and consistency of linkage disequilibrium and identification of DNA markers for production and egg quality traits in commercial layer chicken populations. *BMC Genomics* **10**(Suppl 2):S2.

Ai, H., L. Huang, and J. Ren. 2013. Genetic diversity, linkage disequilibrium and selection signatures in chinese and Western pigs revealed by genome-wide SNP markers. *PLoS One* **8**:e56001.

Alhaddad, H., R. Khan, R. a Grahn, B. Gandolfi, J. C. Mullikin, S. a Cole, T. J. Gruffydd-Jones, J. Häggström, H. Lohi, M. Longeri, and L. a Lyons. 2013. Extent of linkage disequilibrium in the domestic cat, Felis silvestris catus, and its breeds. *PLoS One* **8**:e53537.

Amaral, A. J., H.-J. Megens, R. P. M. A. Crooijmans, H. C. M. Heuven, and M. A. M. Groenen. 2008. Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* **179**:569–579.

Amuzu-Aweh, E. N., H. Bovenhuis, D.-J. de Koning, and P. Bijma. 2014. Proceedings, 10. In: 10th World Congress of Genetics Applied to Livestock Production Prediction. Vancouver.

Andersen, R. 2009. Nonparametric Methods for Modeling Nonlinearity in Regression Analysis. *Annu. Rev. Sociol.* **35**:67–85.

Andreescu, C., S. Avendano, S. R. Brown, A. Hassen, S. J. Lamont, and J. C. M. Dekkers. 2007. Linkage disequilibrium in related breeding lines of chickens. *Genetics* **177**:2161–9.

Ardlie, K. G., L. Kruglyak, and M. Seielstad. 2002. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3**:299–309.

Ardlie, K., S. N. Liu-Cordero, M. a Eberle, M. Daly, J. Barrett, E. Winchester, E. S. Lander, and L. Kruglyak. 2001. Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.* **69**:582–9.

Aulchenko, Y. S., S. Ripke, A. Isaacs, and C. M. van Duijn. 2007. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**:1294–6.

Badke, Y. M., R. O. Bates, C. W. Ernst, J. Fix, and J. P. Steibel. 2014. Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. *G3:Genes|Genomes|Genetics* **4**:623–31.

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

Bates, D. M., and D. G. Watts. 1988. Nonlinear Regression Analysis and Its Applications. Wiley, New York.

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

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

Bishop, C. M. 2006. Pattern recognition and machine learning. Springer, New York, NY.

Bohmanova, J., M. Sargolzaei, and F. S. Schenkel. 2010. Characteristics of linkage disequilibrium in North American Holsteins. *BMC Genomics* **11**:421.

Bouwman, A. C., J. M. Hickey, M. P. Calus, and R. F. Veerkamp. 2014. Imputation of non-genotyped individuals based on genotyped relatives: assessing the imputation accuracy of a real case scenario in dairy cattle. *Genet. Sel. Evol.* **46**:6.

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. *Am. J. Hum. Genet.* **84**:210–23.

Carlson, C. S., M. a Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. a Nickerson. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**:106–20.

Chen, L., C. Li, S. Miller, and F. Schenkel. 2014. Multi-population genomic prediction using a multi-task Bayesian learning model. *BMC Genet.* **15**:53.

Chen, L., F. Schenkel, M. Vinsky, D. H. Crews Jr, and C. Li. 2013. Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. *J. Anim. Sci.* **91**:4669–4678.

Clark, S. a, J. M. Hickey, H. D. Daetwyler, and J. H. J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genet. Sel. Evol.* **44**:4.

Cleveland, W. S. 1979. Robust locally eighted regression and smoothing catterplots. *J. Am. Stat. Assoc.* **74**:368.

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

Corbin, L. J., S. C. Blott, J. E. Swinburne, M. Vaudin, S. C. Bishop, and J. a Woolliams. 2010. Linkage disequilibrium and historical effective population size in the Thoroughbred horse. *Anim. Genet.* **41** Suppl 2:8–15.

Daetwyler, H. D., K. E. Kemper, J. H. J. van der Werf, and B. J. Hayes. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. *J. Anim. Sci.* **90**:3375–84.

Daetwyler, H. D., A. A. Swan, J. H. J. van der Werf, and B. J. Hayes. 2012. Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genet. Sel. Evol.* **44**:33.

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* **3**:e3395.

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

Dekkers, J. C. M. 2007. Marker-assisted selection for commercial crossbred performance. *J. Anim. Sci.* **85**:2104–14.

Dette, H., and N. Neumeyer. 2001. Nonparametric analysis of covariance. *Ann. Stat.* **29**:1361–1400.

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

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

Van Eenennaam, A. L., K. a. Weigel, A. E. Young, M. a. Cleveland, and J. C. M. Dekkers. 2014. Applied Animal Genomics: Results from the Field. *Annu. Rev. Anim. Biosci.* **2**:105–139.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. a Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* **95**:4114–29.

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

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

Frisse, L., R. R. Hudson, a Bartoszewicz, J. D. Wall, J. Donfack, and a Di Rienzo. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet*. **69**:831–43.

Fu, W., J. C. Dekkers, W. R. Lee, and B. Abasht. 2015. Linkage disequilibrium in crossbred and pure line chickens. *Genet. Sel. Evol.* **47**:1–12.

García-Gámez, E., G. Sahana, B. Gutiérrez-Gil, and J.-J. Arranz. 2012. Linkage disequilibrium and inbreeding estimation in Spanish Churra sheep. *BMC Genet.* **13**:43.

Gaspa, G., H. Jorjani, C. Dimauro, M. Cellesi, P. Ajmone-Marsan, a Stella, and N. P. P. Macciotta. 2014. Multiple-breed genomic evaluation by principal component analysis in small size populations. *Animal* 1–12.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2009. ASReml User Guide Release 3.0.

Goddard, M. E., B. J. Hayes, H. McPartlan, and A. J. Chamberlain. 2006. Can the same genetic markers be used in multiple breeds? In: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production. Belo Horizonte. p. 22–16.

Goddard, M. E., and B. J. Hayes. 2007. Genomic selection. *J. Anim. Breed. Genet.* **124**:323–30.

Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat. Rev. Genet.* **10**:381–91.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**:245–57.

Groenen, M. A. M., A. L. Archibald, H. Uenishi, C. K. Tuggle, Y. Takeuchi, M. F. Rothschild, C. Rogel-Gaillard, C. Park, D. Milan, H.-J. Megens, S. Li, D. M. Larkin, H. Kim, L. A. F. Frantz, M. Caccamo, H. Ahn, B. L. Aken, A. Anselmo, C. Anthon, L. Auvil, B. Badaoui, C. W. Beattie, C. Bendixen, D. Berman, F. Blecha, J. Blomberg, L. Bolund, M. Bosse, S. Botti, Z. Bujie, M. Bystrom, B. Capitanu, D. Carvalho-Silva, P. Chardon, C. Chen, R. Cheng, S.-H. Choi, W. Chow, R. C. Clark, C. Clee, R. P. M. A. Crooijmans, H. D. Dawson, P. Dehais, F. De Sapio, B. Dibbits, N. Drou, Z.-Q. Du, K. Eversole, J. Fadista, S. Fairley, T. Faraut, G. J. Faulkner, K. E. Fowler, M. Fredholm, E. Fritz, J. G. R. Gilbert, E. Giuffra, J. Gorodkin, D. K. Griffin, J. L. Harrow, A. Hayward, K. Howe, Z.-L. Hu, S. J. Humphray, T. Hunt, H. Hornshøj, J.-T. Jeon, P. Jern, M. Jones, J. Jurka, H. Kanamori, R. Kapetanovic, J. Kim, J.-H. Kim, K.-W. Kim, T.-H. Kim, G. Larson, K. Lee, K.-T. Lee, R. Leggett, H. a Lewin, Y. Li, W. Liu, J. E. Loveland, Y. Lu, J. K. Lunney, J. Ma, O. Madsen, K. Mann, L. Matthews, S. McLaren, T. Morozumi, M. P. Murtaugh, J. Narayan, D. T. Nguyen, P. Ni, S.-J. Oh, S. Onteru, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* **491**:393–8.

Gulisija, D., D. Gianola, and K. A. Weigel. 2007. Nonparametric Analysis of the Impact of Inbreeding on Production in Jersey Cows. *J. Dairy Sci.* **90**:493–500.

Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J. Dairy Sci.* **93**:1243–52.

Hayes, B., and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* **33**:209–229.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009b. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet. Sel. Evol.* **41**:51.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* **92**:433–43.

Hayes, B. J., I. M. MacLeod, H. D. Daetwyler, P. J. Bowman, A. J. Chamberlain, C. J. vander Jagt, A. Capitan, H. Pausch, P. Stothard, X. Liao, C. Schrooten, E. Mullaart, R. Fries, B. Guldbrandtsen, M. S. Lund, D. Boichard, R. F. Veerkamp, C. Van Tassell, B. Gredler, T. Druet, A. Bagnato, J. Vilkki, D.-J. de Koning, E. Santus, and Goddard, M. E. 2014. Genomic Prediction from Whole Genome Sequence in Livestock: the 1000 Bull Genomes Project. In: Proceedings, 10th World Congress of Genetics Applied to Livestock Production Genomic. p. 1–3.

Heifetz, E. M., J. E. Fulton, N. O'Sullivan, H. Zhao, J. C. M. Dekkers, and M. Soller. 2005. Extent and consistency across generations of linkage disequilibrium in commercial layer chicken breeding populations. *Genetics* **171**:1173–81.

Heidaritabar, M., M. P. L. Calus, H-J Megens, A Vereijken, MAM Groenen, and JWM Bastiaansen. 2015. Accuracy of genomic prediction using whole genome sequence data in White egg layer chickens. Manuscript in preparation

Herrero-Medrano, J. M., H.-J. Megens, M. A. Groenen, G. Ramis, M. Bosse, M. Pérez-Enciso, and R. P. Crooijmans. 2013. Conservation genomic analysis of domestic and wild pig populations from the Iberian Peninsula. *BMC Genet.* **14**:106.

Hidalgo, A. M., J. W. M. Bastiaansen, M. S. Lopes, B. Harlizius, M. A. M. Groenen and D-J de Koning. 2015. Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3 in press.

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**:226–231.

Hill, W. G. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genet. Res.* **38**:209–216

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, and P. Croiseau. 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. *J. Dairy Sci.* **97**:1–12.

Huang, Y., J. M. Hickey, M. a Cleveland, and C. Maltecca. 2012. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *Genet. Sel. Evol.* **44**:25.

Hurvich, C. M., J. S. Simonoff, and C. Tsai. 1998. Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion. *J. R. Stat. Soc.* **60**:271–293.

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

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

Ibáñez-Escriche, N., J. Reixach, N. Lleonart, and J. L. Noguera. 2011. Genetic evaluation combining purebred and crossbred data in a pig breeding scheme. *J. Anim. Sci.* **89**:3881–9.

Jeffreys, a J., L. Kauppi, and R. Neumann. 2001. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**:217–222.

Jeffreys, A. J., and C. a May. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat. Genet.* **36**:151–156.

Jiao, S., C. Maltecca, K. A. Gray, and J. P. Cassady. 2014. Feed intake , average daily gain , feed efficiency , and real-time ultrasound traits in Duroc pigs : I . Genetic parameter estimation and accuracy of genomic prediction. *J. Anim. Sci.* **92**:2377–2386.

Jiménez-Montero, J. A., O. González-Recio, and R. Alenda. 2012. Genotyping strategies for genomic selection in small dairy cattle populations. *Animal* **6**:1216–1224.

Kachman, S. D., M. L. Spanger, G. L. Bennett, K. J. Hanford, L. a Kuehn, W. M. Snelling, R. M. Thallman, M. Saatchi, D. J. Garrick, R. D. Schnabel, J. F. Taylor, and E. J. Pollak. 2013. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. *Genet. Sel. Evol.* **45**:30.

Khatkar, M. S., F. W. Nicholas, A. R. Collins, K. R. Zenger, J. a L. Cavanagh, W. Barris, R. D. Schnabel, J. F. Taylor, and H. W. Raadsma. 2008. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC Genomics* **9**:187.

Larmer, S. G., M. Sargolzaei, and F. S. Schenkel. 2014. Extent of linkage disequilibrium, consistency of gametic phase, and imputation accuracy within and across Canadian dairy breeds. *J. Dairy Sci.* **97**:3128–41.

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

Legarra, A., G. Baloche, F. Barillet, J. M. Astruc, C. Soulas, X. Aguerre, F. Arrese, L. Mintegi, M. Lasarte, F. Maeztu, I. Beltrán de Heredia, and E. Ugarte. 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**:1–13.

Li, A., D. Meyre, L. Of, R. Due, T. O. False, P. Result, R. In, and O. Study. 2013. Challenges in reproducibility of genetic association studies : lessons learned from the obesity field. *Int. J. obesity* **37**:559–567.

Li, M. H., and J. Merilä. 2010. Extensive linkage disequilibrium in a wild bird population. *Heredity* **104**:600–610.

Li, Y. R., and B. J. Keating. 2014. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* **6**:91.

Lillehammer, M., T. H. E. Meuwissen, and A. K. Sonesson. 2011. Genomic selection for two traits in a maternal pig breeding scheme. *J. Anim. Sci.* **91**:3079–3087.

Liu, Z., F. R. Seefried, F. Reinhardt, S. Rensing, G. Thaller, and R. Reents. 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genet. Sel. Evol.* **43**:19.

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

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

Marigorta, U. M., and A. Navarro. 2013. High trans-ethnic replicability of GWAS results implies common causal variants. *PLoS Genet.* **9**:e1003566.

Megens, H.-J., R. P. M. A. Crooijmans, J. W. M. Bastiaansen, H. H. D. Kerstens, A. Coster, R. Jalving, A. Vereijken, P. Silva, W. M. Muir, H. H. Cheng, O. Hanotte, and M. A. M. Groenen. 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. *BMC Genet.* **10**:86.

Meuwissen, the, and M. Goddard. 1996. The use of marker haplotypes in animal breeding schemes. *Genet. Sel. Evol.* **28**:161–176.

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

Misztal, I., S. E. Aggrey, and W. M. Muir. 2013. Experiences with a single-step genome evaluation. *Poult. Sci.* **92**:2530–4.

Ober, U., J. F. Ayroles, E. a Stone, S. Richards, D. Zhu, R. a Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. C. Mackay, and H. Simianer. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. *PLoS Genet.* **8**:e1002685.

Olson, K. M., P. M. VanRaden, and M. E. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *J. Dairy Sci.* **95**:5378–83.

Pei, Y. F., J. Li, L. Zhang, C. J. Papasian, and H. W. Deng. 2008. Analyses and comparison of accuracy of different genotype imputation methods. *PLoS One* **3**.

Pszczola, M., T. Strabel, H. a Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* **95**:389–400.

Qanbari, S., E. C. G. Pimentel, J. Tetens, G. Thaller, P. Lichtner, a R. Sharifi, and H. Simianer. 2010. The pattern of linkage disequilibrium in German Holstein cattle. *Anim. Genet.* **41**:346–56.

R Development Core Team: R: A language and environment for statistical computing. 2013.

Ramos, A. M., R. P. M. A. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald, J. E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M. S. Hansen, J. Hedegaard, Z.-L. Hu, H. H. Kerstens, A. S. Law, H.-J. Megens, D. Milan, D. J. Nonneman, G. A. Rohrer, M. F. Rothschild, T. P. L. Smith, R. D. Schnabel, C. P. Van Tassell, J. F. Taylor, R. T. Wiedmann, L. B. Schook, and M. A. M. Groenen. 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PLoS One* **4**:e6524.

Rao, C. R. 1973. Linear statistical inference and its applications. John Wiley, New York.

Raven, L.-A., B. G. Cocks, and B. J. Hayes. 2014. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics* **15**:62.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. 2001. Linkage disequilibrium in the human genome. *Nature* **411**:199–204.

Riggio, V., M. Abdel-Aziz, O. Matika, C. R. Moreno, a Carta, and S. C. Bishop. 2014. Accuracy of genomic prediction within and across populations for nematode resistance and body weight traits in sheep. *Animal* **8**:520–8.

De Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. *Genetics* **183**:1545–53.

De Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics* **179**:1503–12.

Schaffner, S. F. 2004. The X chromosome in population genetics. *Nat. Rev. Genet.* **5**:43–51.

Schmidt, C. O., T. Ittermann, A. Schulz, H. J. Grabe, and S. E. Baumeister. 2013. Linear, nonlinear or categorical: how to treat complex associations? Splines and nonparametric approaches. *Int. J. Public Health* **58**:161–5.

Simeone, R., I. Misztal, I. Aguilar, and Z. G. Vitezica. 2012. Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. *J. Anim. Breed. Genet.* **129**:3–10.

Slatkin, M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nat. Rev. Genet.* **9**:477–85.

Spelman, R. J., C. a Ford, P. McElhinney, G. C. Gregory, and R. G. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. *J. Dairy Sci.* **85**:3514–3517.

Stranger, B. E., E. a Stahl, and T. Raj. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**:367–83.

Sved, J. a. 2009. Linkage disequilibrium and its expectation in human populations. *Twin Res. Hum. Genet.* **12**:35–43.

Sved, J. A. 1971. Linkage Disequilibrium and Homozygosity of Chromosome Segments in Finite Populations. *Theor. Popul. Biol.* **2**:125–141.

Thaller, G., W. Kramer, A. Winter, B. Kaupe, G. Erhardt, and R. Fries. 2004. Effects of DGAT1 variants on milk production traits in Jersey cattle. *J. Anim. Sci.* **81**:1911–1918.

Tiezzi, F., and C. Maltecca. 2015. Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. *Genet. Sel. Evol.* **47**:24.

Thomasen, J. R., a C. Sørensen, M. S. Lund, and B. Guldbrandtsen. 2014. Adding cows to the reference population makes a small dairy population competitive. *J. Dairy Sci*. **97**:1–11.

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

Uimari, P., and M. Tapio. 2011. Extent of linkage disequilibrium and effective population size in Finnish Landrace and Finnish Yorkshire pig breeds. *J. Anim. Sci.* **89**:609–14.

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

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

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

Vicente, a a, M. I. Carolino, M. C. O. Sousa, C. Ginja, F. S. Silva, a M. Martinez, J. L. Vega-Pla, N. Carolino, and L. T. Gama. 2008. Genetic diversity in native and commercial breeds of pigs in Portugal assessed by microsatellites. *J. Anim. Sci.* **86**:2496–507.

Wall, J. D. 2001. Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. *Curr. Opin. Genet. Dev.* **11**:647–51.

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.

Wang, X. 2010. Package " fANCOVA ."

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

Warnes, G., and F. Leisch. 2005. genetics: Population Genetics.

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

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

Xu, L., J. B. Cole, D. M. Bickhart, Y. Hou, J. Song, P. M. VanRaden, T. S. Sonstegard, C. P. Van Tassell, and G. E. Liu. 2014. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC Genomics* **15**:683.

Yue, X.-P., C. Dechow, T.-C. Chang, J. M. Dejarnette, C. E. Marshall, C.-Z. Lei, and W.-S. Liu. 2014. Copy number variations of the extensively amplified Y-linked genes, HSFY and ZNF280BY, in cattle and their association with male reproductive traits in Holstein bulls. *BMC Genomics* **15**:113.

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

Zhang, Z., J. Liu, X. Ding, P. Bijma, D.-J. de Koning, and Q. Zhang. 2010. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. *PLoS One* **5**:1–8.

Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao, J. He, J. Li, and H. Simianer. 2014. Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. *PLoS One* **9**: e93017.

Summary

Summary

Securing a sufficiently large set of genotypes and phenotypes can be a limiting factor when implementing genomic selection. This limitation may be overcome by combining data from multiple populations or by using information of crossbred animals. The research described in this thesis characterized linkage disequilibrium (LD) patterns in different pig populations and evaluated whether the consistency of LD between populations allows us to make predictions about the performance of genomic selection when multiple populations are included in the prediction and/or validation datasets.

In chapter 2 I evaluated the persistence of LD and patterns of LD decay of pure and crossbred pig populations using real data that was representative of the crossbreeding structure of pig production. The persistence of phase between the crosses and their parental populations was high, indicating that similar marker effects might be expected across these populations. Across the purebred populations the persistence of phase was low therefore higher density panels should be used to have the same marker-QTL associations across these populations.

In chapter 3, the well-known nonlinear model developed by Sved (1971) was compared against an alternative, loess regression, to describe LD decay. The loess regression model was found to be less influenced by the lack of normality, independence and homogeneity of residual variance than the nonlinear regression model. The loess regression model resulted in more reliable LD predictions and can be used to formally compare the LD decay curves between populations.

Chapter 4 showed the utility of different reference sets (across- and multi-population) for the prediction of genomic breeding values, as well as the potential of using crossbred performance in genomic prediction. None of the accuracies obtained using across-population, or multi-population genomic prediction, nor the accuracies obtained using crossbred data, followed the expectations based on LD that was described in chapter 2. I showed that across-population prediction accuracy was negligible even when the populations had common breeds in their genetic background. The variable accuracies of multi-population prediction and moderate accuracy of prediction of crossbred performance appeared to be a result of the differences in genetic architecture between pure populations and between purebred and crossbred animals.

In chapter 5, a methodology that uses information from genome wide association analyses in the genomic predictions was developed and evaluated. The aim in chapter 5 was to let the genomic prediction model use information from the genetic architecture in single- and multi-population genomic prediction. I showed

that using weights based on GWAS results from a combined population did result in higher accuracies of GBLUP in single- as well as in multi-population predictions.

In chapter 6 I placed my results in a broader context. I discussed about the theoretical and practical aspects of linkage disequilibrium in breeding and in the estimation of effective population size. I also discussed the application of genomic selection in a small population and in practical pig breeding, including the prospects of using whole genome sequence for genomic prediction.

Samenvatting

Samenvatting

Het verzamelen van dataset met voldoende genotypes en fenotypes kan een limiterende factor zijn in de uitvoering van genomic selection. Deze limit kan worden opgeheven door het combineren van data afkomstig van verschillende populaties, of door het gebruik van data afkomstig van gekruiste dieren. Het onderzoek beschreven in dit proefschrift karakteriseert de patronen van Linkage Disequilibrium (LD) in verschillende varkens populaties en onderzoekt of de consistentie van het LD tussen populaties een voorspeller is voor de prestatie van genomic selection wanneer er meerdere populaties opgenomen zijn in de training en/of de validatie datasets.

In hoofdstuk 2 gebruik ik data uit van populaties die representatief zijn voor de kruisings structuur zoals gebruikt in de varkens fokkerij om de persistentie en het verval van LD in zuivere lijnen en ook in gekruiste dieren te onderzoeken. De persistentie van de LD tussen de kruisingen en hun ouder populaties was hoog. Hieruit kan worden afgeleid dat de merker effecten over deze populaties heen naar alle waarschijnlijkheid vergelijkbaar zijn. Tussen de verschillende zuivere lijnen was de persistentie van LD laag. Om vergelijkbare merker effecten te vinden over zuivere lijnen heen zou daarom een merker panel met een hogere merker dichtheid nodig zijn dan het panel dat hier gebruikt is.

In hoofdstuk 3 wordt voor het beschrijven van het verval van LD het bekende niet-lineaire model zoals afgeleid door Sved (1972) vergeleken met een alternatief, loess regressie model. Het loess regressie model bleek minder gevoelig voor de afwezigheid van normaliteit, en voor het niet onafhankelijk en homogeen verdeeld zijn van de residuele variantie, in vergelijking met het niet-lineaire regressie model van Sved. Voorspelling van LD met het loess regressie model was meer betrouwbaar. Op basis van de resultaten wordt geconcludeerd dat het loess regressie model kan worden gebruikt voor het vergelijken van verschillen in verval van LD tussen populaties.

Hoofdstuk 4 beschrijft de effecten van verschillende training datasets (across- en multi-populatie) op de betrouwbaarheid van genomische fokwaarden, en de mogelijkheden voor het gebruik van kruisings data voor het schatten van genomische fokwaarden. De betrouwbaarheid van de genomische fokwaarden voor de scenarios across-populatie, multi-populatie, en ook van de voorspellingen met kruisingsdata voldeed in geen geval aan de verwachtingen op basis van LD resultaten beschreven in hoofdstuk 2. Ik laat zien dat de betrouwbaarheid verwaarloosbaar is wanneer data van een andere populatie wordt gebruikt als training dataset, zelfs wanneer er sprake is van een gemeenschappelijke genetische achtergrond tussen beide populaties. De betrouwbaarheid van multi-populatie

voorspellingen was variabel, en de betrouwbaarheid van fokwaarden van gekruiste populaties was matig. Deze resultaten lijken het gevolg van verschillen in genetische architectuur tussen zowel de zuivere lijnen als ook tussen zuivere lijnen en gekruiste populaties.

In hoofdstuk 5 heb ik een methode uitgewerkt en getest waarmee informatie van associatie studies gebruikt wordt in de genomische voorspelling van fokwaarden. Het doel van hoofdstuk 5 was om de meerwaarde vast te stellen van het gebruik van informatie over de genetische architectuur in genomische voorspelling van fokwaarden wanneer één of meerdere populaties gebruikt worden in de training dataset. Ik laat zien dat het gebruik van wegingsfactoren op basis van een associatie studie in een gecombineerde dataset resulteert in hogere betrouwbaarheid van genomische fokwaarden, zowel binnen een populatie als ook in multi-populatie voorspellingen.

In hoofdstuk 6 plaats ik mijn resultaten in een bredere context. Ik bespreek de theoretische en praktische aspecten van LD in de fokkerij en in het schatten van de effectieve populatie grootte. Ik bespreek ook de toepassingen van genomic selection in de kleine populaties en in de varkensfokkerij, inclusief de mogelijkheden voor het gebruik van complete genoom sequenties in het voorspellen van genomische fokwaarden.

Acknowledgements

Acknowledgements

Most of the gains we have achieved in our lives would not have been possible without the direct or indirect contributions of many hands. I feel myself a very lucky person because in this journey I had the opportunity to learn with a lot of people.

Paulo Sávio, I am grateful for you to be my advisor since the bachelor and for your guidance in science. I always admire your professionalism and rectitude. Thank you for the great professional experiences that would not have been possible without the support of you and Simone.

Simone, thank you for giving me lessons that extend beyond academia and for always making me believe in myself. Your strength and capacity to build scientific networks are admirable.

Fabyano, your enthusiasm for scientific research is amazing and contagious. Thank you, for all the great ideas, discussions and scripts. I am grateful for your friendship and for always trying to help me inside and outside the university.

Dear Johan, your scientific view enhanced my papers and this thesis . Thank you for your kindness and for your support in pursuing my PhD. Without your help with all the formalities and agreements the PhD at Wageningen University would not have been possible.

It was not imaginable for me how much I developed my scientific skills during my PhD at Wageningen and I have no doubts that John was the biggest contributor to make it happen. I am really grateful to you John, because your daily supervision taught me a lot about being a researcher, advisor and professor. Thank you for your dedication and for your invaluable help with this thesis.

Dear Egbert and Barbara thank you for challenging me to go deep into the practical and biological insights of my research. It helped me to improve the thesis and to broaden my scientific vision.

I feel really proud of being your friend Marcos! You have been taking such a successful and beautiful path. Thank you for our friendship, for the constructive help while working on this thesis and for being my paranymph.

André, thank you for your contribution to this thesis and our friendship.

Claudia thank you for the scientific discussions and for being my paranymph.

Thanks to all members of ABGC. The time that I spent at WUR was short but rich because of all of you.

Ada and Lisette thank you for the fundamental help with all my documents. Lisette I am really grateful for your help in handling my thesis.

I would like to thank the Topigs Norsvin Research Center for the data and for all the support that they provided to my research.

Thank you to CNPq, FAPEMIG and CAPES for their financial support.

Thank you to the Universidade Federal de Viçosa and Graduate Program of Breeding and Genetics for the support and for making possible de PhD double-degree.

It is almost 10 years already at the Animal breeding and genetic group of the Universidade Federal de Viçosa and I am really happy and grateful to have the opportunity to work, learn, grow and have fun in this team. In this journey a lot of people passed through the group and I thank you all for your contributions to my growth as a person and as a researcher.

André, Darlene, Vanessa, Carolina, Walmir, Rafael, Cândida and Luiz, the times that we had are unforgettable, mainly the funny ones. Thank you for making my life happier and easier. For being my partners in beers, barbecues and pizza.

Thank you Lucas, Katiene and Priscila for our scientific discussions and friendship.

Vanessa and Carolina thank you for the great times that we had in Wageningen.

Saturday is always the best day of the week because of MEIMEI. Thank you to all the members and kids of this institution.

My dear friend Thaís thanks a lot for all these years of friendship. Your advice is always a good guide.

Rosana and Junior, thank you for being my full-time friends and for us being always united even when apart.

I thank my mother (Irene) for letting me be dreamer, for always believing in me, for being an example of a courageous and strong woman, who has always been my greatest motivator and source of admiration and inspiration.

I am grateful to you father (Rosivaldo) for teaching me the importance of working hard and being dedicated.

Curriculum Vitae

About the author

Renata Veroneze is born on July 27th in Capivari, Brazil. She grew up on a farm, which was the inspiration to take the BSc in Animal Science at Universidade Federal de Viçosa (UFV, Viçosa, Brazil). In the beginning of her bachelor she became interested in genetics and statistics and in 2006 became a Junior researcher at the Animal Breeding group. During her bachelor she did an internship in the Netherlands at the Institute for Pig genetics and in January 2009 she graduated with a BSc in Animal Science. The following February, Renata started her MSc in Animal Science at Universidade Federal de Viçosa. In February 2011, she defended her MSc thesis entitled: "Linkage disequilibrium and haplotype block structure in six commercial pig lines". In the same month, she started her PhD in Genetics and Breeding at UFV. The PhD project was built in a partnership with the Animal breeding and Genetic group of Universidade Federal de Viçosa, and the Animal Breeding and Genomics Centre of Wageningen University. Renata was accepted as a PhD candidate at the Animal Breeding and Genomics Center at Wageningen University in 2014. During her PhD, she worked on the project "Linkage disequilibrium and genomic selection in pigs", and focused on the characterization of LD patterns in different pig populations and evaluated whether the consistency of LD between populations could be used as an indicator for the performance of genomic predictions when multiple populations were included in the prediction and/or validation datasets. The results of the project are presented in this thesis. Currently Renata is writing research proposals to continue her scientific career in Brazil..

Peer reviewed publications

1. Campos CF, Lopes MS, Silva FF, **Veroneze R**, Knol EF, Lopes PS, Guimarães SEF (2015) Genomic selection for boar taint compounds and carcass traits in a commercial pig population. Livestock Science 174:10-17.
2. Hidalgo AM, Bastiaansen JWM, Lopes MS, **Veroneze R**, Groenen MAM, Koning D-J (2015) Accuracy of genomic prediction using deregressed breeding values estimated from purebred and crossbred offspring phenotypes in pigs. Journal of Animal Science 94:8899.
3. Costa EV, Diniz DB, **Veroneze R**, de Resende MDV, Azevedo CF, Guimarães SEF, Silva FF, Lopes PS (2015) Estimating additive and dominance variances for complex traits in pigs combining genomic and pedigree information. Genetics and Molecular Research 4:6303-6311.
4. **Veroneze R**, Bastiaansen JWM, Knol EF, Guimarães SEF, Silva FF, Harlizius B, Lopes MS, Lopes PS (2014) Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (*Sus scrofa*) populations. BMC genetics 15:126.
5. **Veroneze R**, Lopes PS, Guimarães SEF, Guimarães JD, Costa EV, Faria VR, Costa KA (2014) Using pedigree analysis to monitor the local Piau pig breed conservation program. Archivos de Zootecnia 63:45-54.
6. **Veroneze R**, Lopes PS, Guimarães SEF, Silva FF, Lopes MS, Harlizius B, Knol EF (2013) Linkage disequilibrium and haplotype block structure in six commercial pig lines. Journal of Animal Science 91:3493 - 3501.
7. De Almeida MP, Nascimento CS, Périssé IV, Souza MD, **Veroneze R,** Guimarães SEF (2013) Treatment of long term stored DNA - comparison between different methods to obtain high quality material. Electrophoresis 34.
8. Mendonça PT, Lopes PS, Braccini Neto J, Carneiro PLS, Torres RA, **Veroneze R**, Guimarães SEF (2012) Estimation of genetic parameters in a F2 pig population. Brazilian Journal of Animal Health and Production 13:330-343.
9. Serão NVL, **Veroneze R**, Ribeiro AMF, Verardo LL, Braccini Neto J, Gasparino E, Campos CF, Lopes PS, Guimarães SEF (2011) Candidate gene expression and intramuscular fat content in pigs. Journal of Animal Breeding and Genetics 128:28-34.
10. Sousa KRS, Ribeiro AMF, Goes PRN, Guimarães SEF, **Veroneze R**, Gasparino E (2011) Toll-Like Receptor 6 differential expression in two pig genetic groups vaccinated against Mycoplasma hyopneumoniae. BMC Proceedings 5:S9.

11. Barbosa L, **Veroneze R**, Lopes PS, Regazzi AJ, Torres RA, Santana Junior ML (2010) Estimation of variance components, genetic parameters and genetic trends for litter size of swines. Brazilian Journal of Animal Science 39:2155-2159.
12. Barbosa L, **Veroneze R**, Lopes PS, Regazzi AJ, Torres RA, Santana Junior ML (2008) Estimation of genetic parameters in pigs using Gibbs Sampler. Brazilian Journal of Animal Science 37:1200-1206.
13. Barbosa L, Lopes PS, Regazzi AJ, Torres RA, Santana Junior M L, **Veroneze R** (2008) Estimation of genetic parameters for litter size in pigs using multi-trait analyses. Brazilian Journal of Animal Science 37:1947-1952.

Training and Supervision Plan

The Basic Package (3 ECTS)	year	credits
WIAS Introduction Course	2014	1.5
Course on philosophy of science and/or ethics	2015	1.5
Scientific Exposure (9 ECTS)		
International conferences		
III International Symposium of breeding and genetics, Viçosa - Brazil, 07 - 11 november	2011	1.5
IV International Symposium of breeding and genetics, Viçosa - Brazil, 26 november	2013	0.3
XXVI Genetics Days, Prague - Czech Republic, 03 - 04 September	2014	0.6
V International Symposium of breeding and genetics, Viçosa - Brazil, 06 - 08 november	2014	0.9
Seminars and workshops		
Academic Integration Symposium, Viçosa - Brazil, 20 - 25 october	2014	1.8
Presentations		
Persistence of linkage disequilibrium phase in purebred and crossbred pigs, Wageningen - Netherlands, 30 April, Oral	2014	1.0
Within-, across- and multi-populations reference sets for genomic selection in pigs, Prague - Czech Replubic, 3 September, Poster	2014	1.0
Genomic selection for backfat thickness in pig sire lines, Viçosa - Brazil, 21 October, Poster	2014	1.0
Across- and multi-population genomic selection in pigs, Viçosa - Brazil, 08 November, Poster	2014	1.0
In-Depth Studies (18 ECTS)		
Disciplinary and interdisciplinary courses		
Statistical genomics	2011	2.1

Biometric models applied to genetic improvement	2011	2.1
Quantitative genetics	2012	2.1
Data analysis in animal genetic improvement	2012	2.1
Statistical methods in genomic selection	2012	1.6
Special topics - Use of Genomics in Animal Breeding Programs: goals and current accomplishments	2013	0.5
Genetics meets genomics:Livestock applications of genomic analisys	2013	0.5

Advanced statistics courses

Bayesian inference	2011	1.6
Statistical methods II	2012	2.1

PhD students' discussion groups

Seminars in Genetic improvement	2011	2.1
Seminars in Genetic improvement	2013	1.0
Quantitative Genetics Discussion Group (QDG)	2014	0.5

Professional Skills Support Courses (4 ECTS)

How to elaborate a paper	2013	0.6
Writing a scientific paper in one week	2013	1.4
Course in teaching assistant	2011	2.1

Research Skills Training (2 ECTS)

Trainee at Institute for pigs genetics (IPG), Beuningen, the Netherlands	2012	2.0

Didactic Skills Training (9 ECTS)

Lecturing

Pig genetic improvement, Viçosa - Brazil	2013	2.0

Tutorship

Teaching assistant in Animal breeding and genetics	2013	7.3

Supplementary material

Supplementary material

Chapter 2

Table S2.1 SNP data description according to the quality control criteria.

	SL1	SL2	SL3	DL1	DL2	DLF1	TER1	TER2
MAF < 0.05	7,969	9,550	6,627	6,399	7,271	4,262	2,705	4,685
HWE P-value <0.0001	478	2,500	1,535	4,999	1,526	2,837	5,239	3,493
SNP call rate < 90%	581	1,543	4,075	1,366	1,629	2,334	1,050	1,279
SNPs utilized	38,769	35,505	36,136	35,392	38,058	38,529	38,752	38,583
Number of animals	1,307	643	276	626	1,013	186	286	330

Table S2.2 Grouping of lines and the number of SNPs for persistence of phase estimation.

Lines	Group 1	Group 2	Group 3	Group 4
DL1	X	X	X	X
DL2	X	X	X	X
SL1		X		X
SL2			X	X
SL3				X
DLF1	X	X	X	
TER1		X		
TER2			X	
Number of SNPs	28,153	22,272	22,794	20,435

Table S2.3 Coefficient of variation (CV) for the number of SNP pairs in bins of 10, 30, 50, 70 and 100 Kb.

Distance intervals (Kb)	Group1	Group2	Group3	Group4
10	0.066	0.078	0.077	0.075
30	0.056	0.068	0.067	0.064
50	0.055	0.068	0.066	0.063
70	0.055	0.067	0.066	0.063
100	0.055	0.068	0.066	0.063

Table S2.4 Values of $\chi^2_{computed}$(below the diagonal) and p-values for the pairwise comparison.

	$\hat{\beta}_{SL1}$	$\hat{\beta}_{SL2}$	$\hat{\beta}_{SL3}$	$\hat{\beta}_{DL1}$	$\hat{\beta}_{DL2}$	$\hat{\beta}_{DLF1}$	$\hat{\beta}_{TER1}$	$\hat{\beta}_{TER2}$
$\hat{\beta}_{SL1}$	-	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$
$\hat{\beta}_{SL2}$	6346.2	-	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$
$\hat{\beta}_{SL3}$	130.96	4663.36	-	$<10^{-6}$	0.0117*	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$
$\hat{\beta}_{DL1}$	1380.75	13782.9	2374.35	-	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$
$\hat{\beta}_{DL2}$	81.72	5108.87	6.35	2179.95	-	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$
$\hat{\beta}_{DLF1}$	4561.1	22155.14	6290.32	902.39	6017.03	-	$<10^{-6}$	$<10^{-6}$
$\hat{\beta}_{TER1}$	10848.53	12284.15	13452.27	4481.85	13143.96	1412.85	-	$<10^{-6}$
$\hat{\beta}_{TER2}$	845.92	12284.15	1670.97	79.69	1498.69	1585.25	6031.08	-

Using Bonferroni correction: $\alpha^* = 0.0018$

Chapter 3

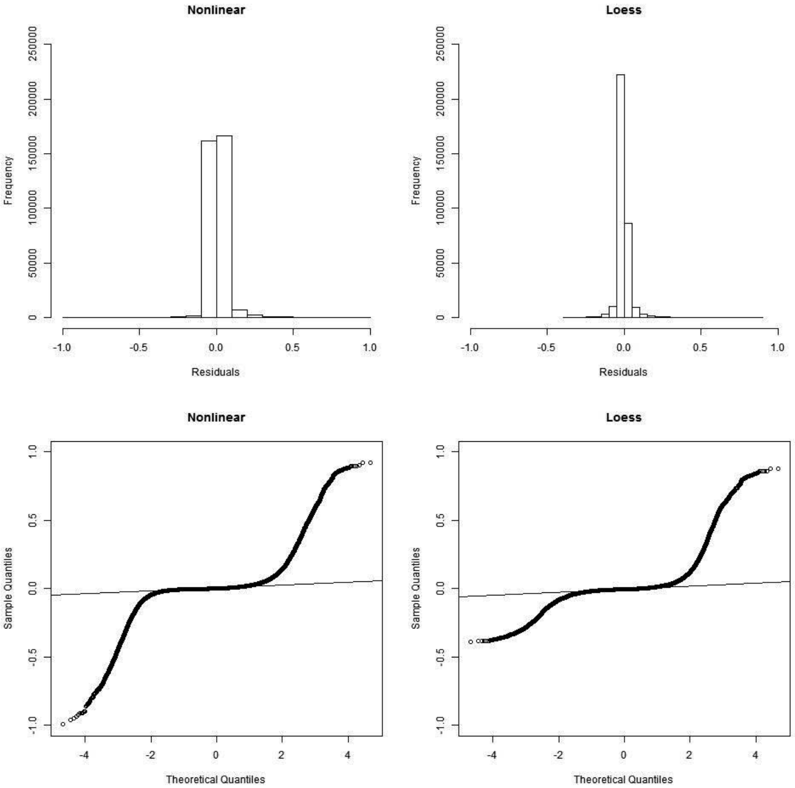

Figure S3.1 Residual frequency and QQ plot for nonlinear and loess regression in DL2.

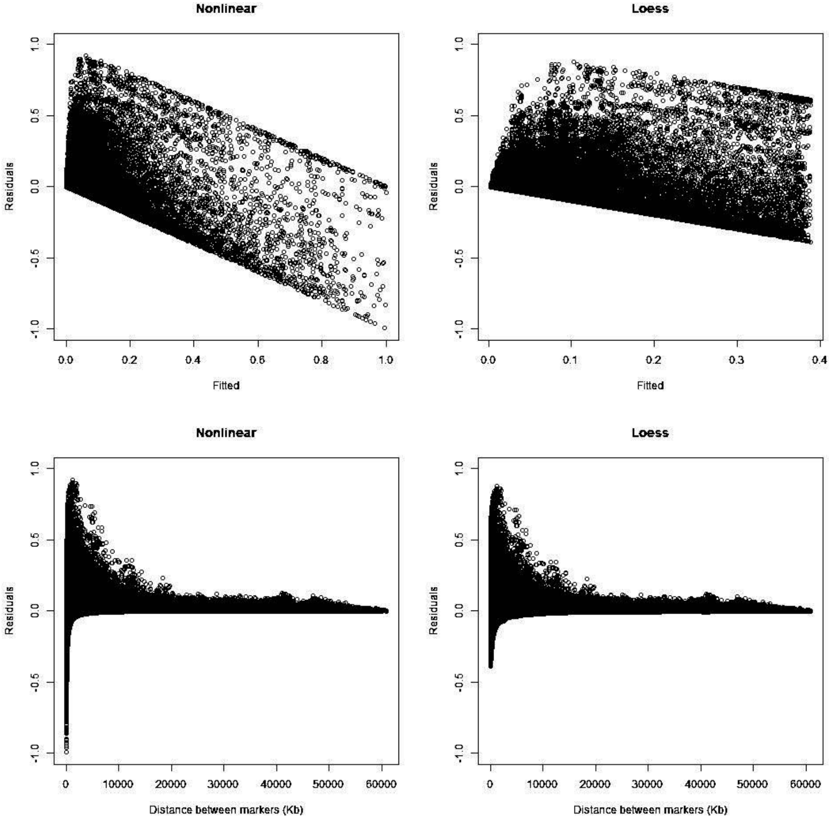

Figure S3.2 Plots of the residuals against fitted values (top) and against the distance between markers (bottom) for nonlinear (left) and loess regression (right) in line DL2.

Colophon

Colophon

The research described in this thesis was financially supported by the Brazilian Government through CAPES/Nuffic program, CNPq and FAPEMIG.

The data used in this thesis were provided by Topigs Norsvin Research Center (Beuningen, The Netherlands).

The cover of this thesis was created and designed by Renata Veroneze and Ágatha Kretli (www.agathakretli.com.br).