ESTIMATION AND EXPLOITATION OF LINKAGE DISEQUILIBRIUM IN PIGS

By

Yvonne Martina Badke

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Animal Science - Doctor of Philosophy

2013

ABSTRACT

ESTIMATION AND EXPLOITATION OF LINKAGE DISEQUILIBRIUM IN PIGS

By

Yvonne Martina Badke

The United States Pork Industry is an important source of income in rural America, and its continued profitability and success can be facilitated through genetic improvement for a variety of production and health traits. Prediction of genomic breeding values (GEBV) based on high density genotypes has the potential to increase genetic progress. The overall objective of this dissertation was to describe the structure of linkage disequilibrium (LD) across the pig genome, assess the potential of genotype imputation from low to high density genotypes, and estimate accuracy of genomic prediction in pure-bred pig populations using either observed or imputed high density genotypes.

The first study focused on the estimation of LD and pairwise persistence of phase across the genome of four US pig populations. Observed LD was high between adjacent SNP (0.36-0.46) and persisted at high levels as pairwise distance between SNP increased to 1 Mb (0.20-0.25). Persistence of phase is a measure of prediction reliability of markers in one population by those in another and ranged between 0.87 and 0.92 for pairwise SNP distance <10 kb. We concluded that high estimates of LD between adjacent SNP in this study are promising for the implementation of genomic selection, especially in conjunction with genotype imputation to increase cost efficiency. However, persistence of phase appears to be too low to indicate that the use of combined training panels would be advantageous for accuracy of genomic prediction at the current marker density.

The second study focused on the accuracy of genotype imputation and variables affecting

imputation accuracy. Using a commercially available 10K tagSNP panel and a small reference panel of 128 haplotypes average accuracy of imputation was 0.95. Increasing the size of the haplotype reference panel led to an overall increase in imputation accuracy ($IA = 0.97$ with 512 haplotypes), but was especially useful in increasing imputation accuracy of SNP with MAF below 0.1 and for SNP located in the chromosomal extremes. In addition, our results show that randomly sampling individuals to genotype for the construction of a reference haplotype panel is more cost efficient than specifically sampling older animals or trios with no observed loss in imputation accuracy. From these results, we expected that losses in accuracy of genomic prediction using imputed genotypes would be minimal.

In the third study we assessed the loss of prediction accuracy of GEBV obtained for Yorkshire pigs using imputed instead of observed genotypes. Accuracy of genomic evaluation using observed genotypes was high for three traits (0.65-0.68). Using genotypes imputed with high accuracy ($R^2 = 0.95$) for genomic evaluation did not significantly decrease accuracy of prediction. The decrease in accuracy of genomic evaluation was significant when imputation accuracy dropped to $R^2 = 0.88$. Genomic evaluation based on imputed genotypes in selection candidates is a cost efficient alternative for implementation of genomic selection in pigs. Furthermore, genotyping animals at lower cost and low density, followed by imputation, can result in increased accuracy by allowing more animals into the training panel.

In conclusion, we showed that accurate prediction of GEBV in a US Yorkshire population is possible, and cost efficiency can be increased through the use of genotype imputation in selection candidates. Furthermore, our results of LD for three other US pig populations indicate that similar or high accuracy of prediction can be expected within each of these populations. In addition, we briefly discuss how our results can be extended to prediction of breed composition, and GEBV prediction and GWAS using whole genome sequence.

To my parents, Jutta and Ruediger for their love and support, and my grandmother for her persistent insistence that I finish my education.

Meinen Eltern, Jutta und Rueder, fuer ihre Liebe und Unterstuetzung, und meiner Oma fuer Ihre durchgehende Motivation meine Ausbildung abzuschliessen.

ACKNOWLEDGMENTS

First I would like to acknowledge and express my deepest gratitude to my family, especially my parents and my brother Alexander. Without your support, encouragement, understanding, and humor the completion of this program would not have been possible. You were there for me when I had good news and more importantly you were always there for the bad news. Knowing that you trusted me to finish this program was an incredible source of strength for me and a huge motivation to not disappoint you.

Secondly, I would like to thank my advisor Juan P Steibel for his encouragement, guidance, and support during the last 4 years. Coming from a background with little knowledge of animal breeding and very limited recollection of Mendelian genetics learned in high-school I would have been lost in this field without his continued assistance and academic mentoring. Thanks to his support and patience in getting through the trials and triumphs of my PhD program I feel that I have grown as a scientist and as a person.

I would also like to thank the members of my guidance committee Drs Cathy Ernst, Ron Bates, Rob Tempelman, and Yuehua Cui for their patience and encouragement, helping me to navigate my dissertation research, each offering their unique perspective and knowledge on the topics I worked on. Especially Drs Cathy Ernst and Ron Bates for their insights, suggestions, and kind revisions of my publications as well as their constant availability to answer any questions I had. I would also like to thank Drs. Rob Tempelman and Dennis Banks for giving me the opportunity to teach in their classes. It has been a unique experience and I realized that I really enjoy teaching and I hope I will be able to integrate some of teaching into my future career.

Finally I would like to thank many of my good friends and fellow graduate students: Maria

Arceo, Pablo Parraga, Elodie Hablot, Oscar Arreola, Emily McKinney, Marcos Oliveira, Jacqueline Reit, and many others. Coming to MSU has given me the opportunity to meet so many different and amazing people and I am grateful for every one of you and I hope I will keep in touch with all of you. Your kindness, support, encouragement, and example have been a vital part of my strength and motivation over the last years. We have shared many of the unique experience of aspiring to a PhD degree, and comforted each other through failed experiments and long weekends with the knowledge that it will work out in the end, and it will be worth it. I knew I could come to you when I needed someone to lift me up and tell me that will get better, and you were always right about that, and you were there to celebrate whenever I had something to celebrate.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Pigs are an important livestock species in the United States producing total sales of more than $97 billion and directly employing almost 35.000 people around the country according to 2011 estimates from the National Pork Producers Council. Genetic improvement through breeding for meat quality, reproductive performance, and health traits is an important tool to assure the pork industries competitiveness and continued success. There has been an increase in lean percentage of roughly 0.1% per year since 2001 (USDA, 2009a) and the number of piglets per litter has increased from 8.8 to 9.5 (USDA, 2009b). This progress in animal performance and productivity can be attributed to a number of factors such as improvements in overall herd health and housing, improved feeding, and genetic improvement through traditional breeding schemes. Traditional breeding describes the estimation of breeding values (EBV) for all animals based on the performance of their relatives. Linear mixed models are fitted to obtain predictions of performance of animals through their expected relationship matrix combining information across all available relatives (Henderson, 1984). Favorably ranking animals are further evaluated based on the information of their offspring to obtain highly accurate EBV to be used in predictions for future generations. It has been conservatively estimated that improvement from traditional genetic selection programs increases revenue to the pork industry by more than $42 million annually (C. Schwab, personal communication). A large number of studies to date have aimed to identify quantitative trait loci (QTL) associated with economically important traits, such that the pig QTL database currently documents 8402 QTL from 356 publications (http://www.animalgenome.org/pigs/). Certain traits, i. e. halothane sensitivity, could be traced back to the gene and the corresponding mutation directly affecting the phenotype, allowing these loci to be directly used in gene assisted selection (Meuwissen *et al.*, 2013). However, the majority of phenotypes appears to be associated with a large number of loci with comparably small individual effects

across the genome (Meuwissen *et al.*, 2001) making selection based on single loci impossible for most economic traits. A related attempt to utilize genetic information, mostly single nucleotide polymorphisms (SNP), for breeding was marker assisted selection, where SNP were tested for association with the phenotype and subsequent selection was based significantly associated SNP (Meuwissen *et al.*, 2013). However, due to imperfect linkage disequilibrium (LD) between the SNP used for selection and the underlying QTL (Meuwissen *et al.*, 2013) and ascertainment bias in selecting significantly related SNP (Meuwissen *et al.*, 2001), accuracy from this approach was generally unsatisfactory.

An alternative to traditional breeding schemes, and gene- or marker-assisted selection has been proposed by Meuwissen *et al.* (2001). Based on the assumption that phenotypes are associated with a large number of loci with small effects, many of which will fail to reach statistical significance when QTL are mapped, they proposed a method of genomic evaluation fitting a prediction equation to thousands of genetic markers in parallel to obtain predicted genomic breeding values (GEBV; Meuwissen *et al.*, 2001). As initially proposed, genomic selection models were based on a two-stage design. In the first stage, a prediction equation was fit to a training panel of animals with highly accurate EBV to estimate SNP effects for all markers. In the second stage, the estimated SNP effects and observed genotypes of selection candidates were used to calculate GEBV. GEBV, especially for young animals, are expected to be more accurate than EBV due to the inclusion of both close and distantly related animals, as well as the use of actual proportions of identity shared by descent (IBD) between animals vs. the expected proportion of IBD that is used to estimate EBV. In addition, GEBV are expected to increase the accuracy of selection for traits that are difficult to measure such as health traits, life-time productivity, or those only available post-mortem. Also, since this method allows selection of animals at a very young age it is expected to

shorten generation intervals (Meuwissen *et al.*, 2001; VanRaden *et al.*, 2009).

Research and implementation of this method has been facilitated in several livestock species through the recent availability of high density genotyping platforms for bovine (Matukumalli *et al.*, 2009), ovine (Archibald *et al.*, 2010), chicken (Groenen *et al.*, 2009), and pig (Ramos *et al.*, 2009). An impressive body of research including simulation experiments and studies on real populations over the last few years have furthered our understanding of variables affecting the accuracy of genomic evaluation and possible designs which can be used to implement GEBV in the livestock industry (Daetwyler *et al.*, 2013).

Among variables affecting the accuracy of genomic selection are the level of LD across the genome, the composition and size of the training population, the statistical models used for prediction (Meuwissen *et al.*, 2013), the genetic architecture and heritability of the trait (Hayes *et al.*, 2009a), and in case genotype imputation is utilized the accuracy of imputation (Weigel *et al.*, 2010a). Interactions between these variables, like the genetic architecture of the trait defining the optimal statistical model (Meuwissen *et al.*, 2013) increase the difficulty of choosing an optimal design.

When the Illumina PorcineSNP60 Genotyping BeadChip (Illumina Inc.; Ramos *et al.*, 2009) was first released in 2008 it introduced the possibility of enabling genomic selection in swine breeding. It was the goal of this dissertation research to use this newly available platform to assess the potential of genomic selection to be implemented into commercial swine breeding.

Our initial step was to assess the extent of LD across the genome in four breeds of pigs and estimate persistence of phase between the breeds. Extent of LD is an important precursor for genomic selection (Hayes *et al.*, 2009a), since the average LD between SNP is indicative of the LD that can be expected between a QTL and neighboring SNP. We compared aver-

age LD between both neighboring SNP and also SNP at an increasing distance, to assess whether or not reducing the number of markers necessary to implement genomic selection could potentially be reduced to increase cost efficiency (Badke *et al.*, 2012). Persistence of phase between populations is informative to assess the usability of training panels combining animals across populations for genomic selection (Goddard *et al.*, 2006; de Roos *et al.*, 2009). If phase is conserved between breeds to a large extent, then combining animals across breeds to form a training panel will likely lead to an increase in selection accuracy (de Roos *et al.*, 2009). However, if persistence of phase between two populations is low then combining them to form a training panel would have a negative impact on the resulting accuracy of genomic selection (de Roos *et al.*, 2009). Due to the relatively small size of pure-bred swine populations combining animals across populations to form a training panel for genomic selection could decrease the initial investment necessary.

After assessing that LD in swine is comparably high between both neighboring SNP and SNP at an increasing distance we considered options to implement low density genotyping and genotype imputation to potentially increase cost efficiency of genomic selection (Badke *et al.*, 2012). Cost efficiency is critical in swine breeding to allow a widespread implementation of genomic methods, but accuracy of genomic prediction based on imputed instead of observed genotypes depends on the accuracy of genotype imputation Weigel *et al.* (2010a). Though overall cost of genotyping has decreased dramatically over the last few years it is still not feasible for pig breeding operations to genotype young animals or selection candidates using high density genotypes on the PorcineSNP60. We proposed to select a small panel of maximally informative tagSNP (Qin *et al.*, 2006) that could subsequently be used to impute genotypes from low density back to the original high density array. Genotype imputation from low density panels using a reference panel of haplotypes is computationally

efficient and imputed genotypes can reach almost perfect accuracy at a much reduced cost (Browning and Browning, 2009; Dassonneville *et al.*, 2011; Gualdrón Duarte *et al.*, 2013; Huang *et al.*, 2012; Wiggans *et al.*, 2012). Accuracy of imputation mainly depends on the amount of information available for imputation, such that a large reference panel of haplotypes (Dassonneville *et al.*, 2011) and the inclusion of linkage and LD information positively impact the resulting accuracy (Gualdrón Duarte *et al.*, 2013; Huang *et al.*, 2012). It was our second objective to select a maximally informative panel of low density SNP in four US pig breeds using several selection strategies such as the physical distance between SNP, a minimum threshold of LD between tagSNP (Qin *et al.*, 2006), or the ability of each SNP to accurately predict the genotypes of markers not included in the panel (Badke *et al.*, 2013). Since low density to high density genotype imputation necessitates the use of a reference panel of haplotypes we also sought to investigate the effect of reference panel composition. Parallel to our work on designing low density SNP panels the GeneSeek Genomic Profiler for Porcine LD (GGP Porcine), a low density platform comprising roughly 10K, assembled using the SNPspace software (C.P. Van Tassell, unpublished data) was released by GeneSeek (Lincoln, NE). Recognizing the potential of a non breed-specific general low density panel for industrial application in pig breeding we changed our focus from the design of breed-specific low density platforms to perform a detailed assessment of genotype imputation accuracy using the GGP Porcine to impute high density genotypes in Yorkshire pigs.

Finally, we combined the results obtained in our previous publications (Badke *et al.*, 2013, 2012) to investigate the accuracy of genomic evaluation in a US Yorkshire population using both observed and imputed genotypes. Imputation accuracy observed in this Yorkshire population was comparable to previous reports in dairy cattle (Dassonneville *et al.*, 2011; Wiggans *et al.*, 2012) and we expect it can be further increased through the use of algorithms

that combine information on linkage and LD to obtain imputed genotypes (Gualdrón Duarte *et al.*, 2013; Huang *et al.*, 2012; Wiggans *et al.*, 2012). Therefore, the loss of accuracy in genomic evaluation when imputed instead of observed genotypes are used in selection candidates is expected to be minimal, which would concur with previous results in dairy cattle breeding (Dassonneville *et al.*, 2011; Wiggans *et al.*, 2012). We used a cross validation design to implement genomic evaluation using a computationally efficient model and assessed the accuracy of this evaluation for both observed genotypes, and genotypes imputed using different size reference panels. In addition, we assessed the effect of several variables that have previously been shown to influence accuracy of genomic evaluation such as relatedness between training and prediction (Clark *et al.*, 2012), and the accuracy of EBV used for validation (Hayes *et al.*, 2009a).

In conclusion, the overarching goal of this dissertation was to implement a scheme for the adoption of genomic selection techniques into swine breeding, that would be optimally fit to the structure and requirements of this population. We aimed to present a solution that would yield highly accurate predictions while maintaining cost efficiency with respect to initial and long-term investment. In addition, by releasing all data and code used to obtain these results we hope to facilitate further research addressing issues within this specific population, but also the translation of our approach to designing a genomic selection scheme for other populations and species not currently employing these techniques.

Specifically the objectives of this dissertation were:

1. Estimate LD in the Duroc, Hampshire, Landrace and Yorkshire pig breeds using SNP genotypes obtained using the Illumina PorcineSNP60 Genotyping BeadChip. Determine persistence of phase between breeds to assess the potential of mixed breed reference panels for both imputation and genomic selection in the future.

2. Assess tagSNP selection strategies to obtain a maximally informative subsets of tagSNP that effectively span the genome for each breed. In addition, report on imputation accuracy of the recently released GeneSeek Genomic Profiler for Porcine LD (GGP-Porcine, GeneSeek a Neogen Company, Lincoln, NE), a commercially available 10K tagSNP panel.

3. Perform GEBV prediction for economically important production traits using high density SNP genotypes for the Yorkshire breed, as well as genotypes imputed from the GGP-Porcine. Assess the loss in accuracy when genotypes in selection candidates were imputed from low to high density.

Chapter 2

Estimation of linkage disequilibrium in four US pig breeds

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., & Steibel, J. P. (2012). Estimation of linkage disequilibrium in four US pig breeds. BMC genomics, 13(1), 24.

2.1 Background

The extent of non-random association of gametes at different loci, or linkage disequilibrium (LD), has become the focus of many recent studies in both humans and animals (Amaral *et al.*, 2008; Conrad *et al.*, 2006; Reich *et al.*, 2001; Corbin *et al.*, 2010). Gaining knowledge of the distribution of LD in livestock populations is important for genetic mapping of economically important traits such as disease resistance (Pritchard and Donnelly, 2001), and it can reveal population history and breed development (Nordborg and Tavare, 2002; Tenesa *et al.*, 2007). Moreover, genome wide association (GWAs) studies as well as genomic selection in livestock rely on the existence of LD between causative variants and genetic markers (Hayes *et al.*, 2009a; Goddard and Hayes, 2009). Recent advances in genotyping technology allow high density genotyping of single nucleotide polymorphisms (SNP) for several livestock species such as cattle (Matukumalli *et al.*, 2009), chicken (Groenen *et al.*, 2009), and pigs (Ramos *et al.*, 2009) Obtaining high density genotypes from a sample of individuals allows for the estimation of genome-wide LD and persistence of phase among breeds (Goddard *et al.*, 2006). Previous studies have shown that the extent and persistence of LD in livestock (de Roos *et al.*, 2008; Sargolzaei *et al.*, 2008; Uimari and Tapio, 2011) is much larger than that found in human populations (Reich *et al.*, 2001), due to selection and smaller effective population size in livestock species (Amaral *et al.*, 2008; Harmegnies *et al.*, 2006). Using dense markers to cover the genome increases the likelihood of SNP markers to be in high LD with causative genes altering important production phenotypes (Goddard, 2008). Meuwissen and Goddard (2001) proposed that the merit of these markers in livestock would be in the parallel use of all markers to derive genomic breeding values (GEBV) as a composite score of all individual SNP effects rather than improving mapping of quantitative trait loci

(QTL). The implementation of genomic selection using GEBV has been successful in dairy cattle (Hayes *et al.*, 2009a; VanRaden *et al.*, 2009; Goddard and Hayes, 2007), and is currently being tested in laying chickens (Wolc *et al.*, 2011) and pigs (Cleveland *et al.*, 2010). The reliability of GEBV prediction relies on the level of LD between markers and QTL, the origin of such LD (either within family or population-wise), the number of animals used in the training population as well as heritability of the trait (Hayes *et al.*, 2009a). In this study it is our objective to estimate and describe genome wide levels of LD in four pig breeds using high density genotypes. We also estimate population-wise LD for a variety of panels with lower marker density in order to estimate the number of markers needed to reach a given level of LD. We estimate persistence of phase between the four breeds in this study as a measure of relationship between these populations.

2.2 Results

2.2.1 Estimation of Linkage Disequilibrium

To estimate LD, we genotyped 351 animals in 117 sire/dam/offspring trios across four breeds of pigs (Duroc, Hampshire, Landrace and Yorkshire) using the Illumina Porcine SNP60 BeadChip (Ramos *et al.*, 2009). We used BEAGLE (Browning and Browning, 2009) to build haplotypes and estimated pairwise r^2 for all SNP on the same chromosome using equation (2.1). Average r^2 between adjacent markers within breed was estimated using equation (2.2). Average r^2 at various distances was computed by grouping all SNP combinations by their pairwise distance in classes of 100 kb of length starting at 0 to 10 Mb. Figure 2.1 displays an overview of the decline of r^2 over distance in each breed. In addition, Table 2.1 displays average r^2 for adjacent markers and at 0.5, 1 and 5 Mb. The average r^2 between adjacent

SNP was largest in the Duroc animals (r^2=0.46), followed by Hampshire (r^2=0.44), whereas Yorkshire and Landrace exhibited the smallest average r^2 (0.39 and 0.36 respectively; Table 2.1). Marker pairs with an average distance of 1 Mb had an average r^2 of 0.20 for Hampshire, 0.19 for Duroc, 0.16 for Yorkshire and 0.15 for Landrace. For all breeds, at least 54% of the adjacent SNP had an $r^2 \geq 0.2$ and 44% $r^2 \geq 0.3$. For most chromosomes, average r^2 between adjacent SNP in Duroc and Hampshire was larger than average r^2 in Landrace or Yorkshire. In addition to estimating average r^2 within distance classes, we also computed average r^2 between adjacent markers for different marker densities. To obtain marker sets with various SNP densities we sequentially removed markers from the current map using every second, fourth, 10th, 50th, 100th and 200th marker (Table 2.2). Average r^2 decreased between 6% for Yorkshire to 15% for Hampshire when only 50% of the markers were used, with highest average r^2 for Duroc (r^2=0.40) followed by Hampshire (r^2=0.37), Yorkshire (r^2=0.34) and the lowest for Landrace (r^2=0.30). Using only every 10th marker, average r^2 decreased to around 50% of the original r^2 (r^2=0.20-0.25), and using every 100th marker average r^2 ranged from 0.05-0.07 at an average marker distance of 6.5 Mb, which was comparable to the results found for average r^2 at 5 Mb.

Table 2.1: Average r^2 at various distances in four breeds

Breed	Adjacent	0.5Mb[2]	1Mb[2]	5Mb[2]
Duroc	0.46	0.26	0.19	0.06
Hampshire	0.44	0.25	0.20	0.08
Landrace	0.36	0.19	0.15	0.06
Yorkshire	0.39	0.21	0.16	0.05

[1] Average r^2 for SNP with adjacent map positions (exact spacing: 70 kb for Duroc, 74 kb for Hampshire, 60 kb for Landrace, and 61kb Yorkshire).
[2] Average r^2 for SNP spaced 0.5 Mb, 1 Mb and 5 Mb apart

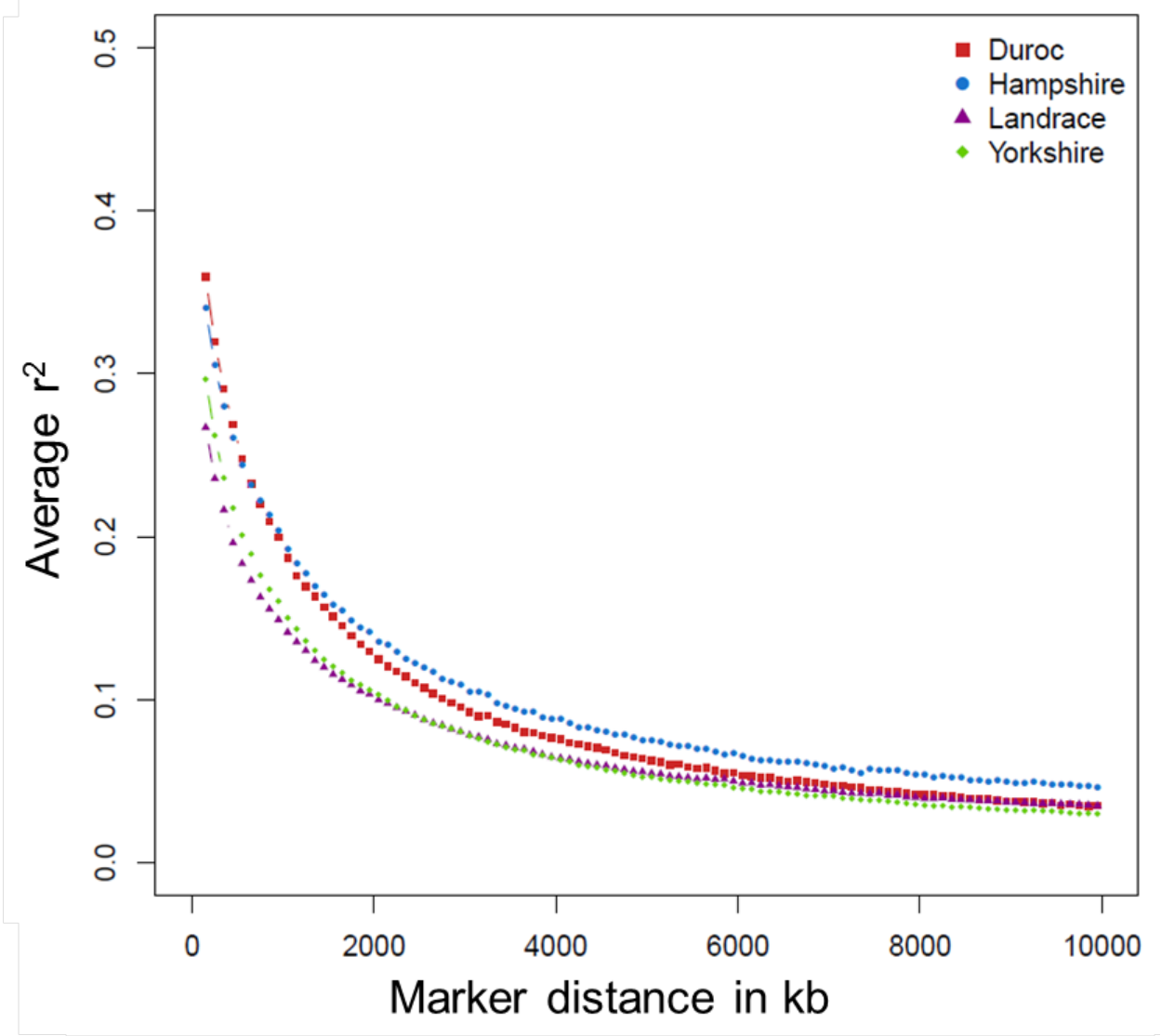

Figure 2.1: Decay of average r^2 over distance
Average r^2 between markers in Duroc, Hampshire, Landrace and Yorkshire at various distances in base pairs ranging from 0 to 10 Mb.
For interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation.

Table 2.2: Average r^2 between adjacent SNP for sparse marker panels

% of SNP kept [1]	Duroc		Hampshire		Landrace		Yorkshire	
	average r^2 [2]	average distance in kb [3]	average r^2 [2]	average distance in kb [3]	average r^2 [2]	average distance in kb [3]	average r^2 [2]	average distance in kb [3]
50%	0.4	141	0.37	148	0.3	120	0.34	123
25%	0.34	281	0.31	296	0.25	239	0.28	246
10%	0.25	703	0.23	740	0.2	597	0.21	613
2%	0.1	3,507	0.11	3,693	0.09	2,978	0.09	3,056
1%	0.05	7,026	0.06	7,399	0.05	5,963	0.05	6,127
0.50%	0.02	14,120	0.04	14,872	0.03	11,977	0.02	12,313

[1] Percentage of SNP included in the current set of markers
[2] Average r^2 for SNP with adjacent map positions for the current set of markers
[3] Average distance in *kb* for SNP with adjacent map positions in the current set of markers

2.2.2 Persistence of Phase

Persistence of phase is a measure of the degree of agreement of LD phase for pairs of SNP between two populations. To estimate persistence of phase, we calculated r_{ij} as the square root of r^2_{ij} in equation (2.1) between all possible combinations of SNP i and j respectively, using the sign of the non-squared numerator. If r^2 between two markers is equal in two populations, but their corresponding r has opposite sign, the gametic phase is reversed (Uimari and Tapio, 2011). Persistence of phase over a certain genomic distance interval can be estimated as the pairwise Pearson correlation coefficient ($R_{k,k'}$) of inter-marker r_{ij} between two populations k and k (Equation 2.3). For all pairwise comparisons of breeds we estimated $R_{k,k'}$ and the percentage of SNP with reversed sign of r. Similar to our computation of average r^2, we grouped SNP pairs in classes of inter-marker distances 100 kb long and computed persistence of phase within each class starting at 0 up to 10 Mb (Figure 2.2). In theory, the Pearson correlation coefficient ranges between -1 and 1. Large negative values are a result of high LD (r^2) in both breeds but phase is reversed between them. High positive values are a result of high r^2 and equal phase in both breeds (Uimari and Tapio, 2011). Correlation of phase between SNP less than 100 kb apart ranged from 0.73 for Duroc with Hampshire and Yorkshire to 0.82 for Landrace with Yorkshire. Considering SNP pairs with an average distance of 0.9 to 1 Mb, correlation of phase decreased to 0.41 for Duroc with Hampshire and to 0.57 for Yorkshire with Landrace (Table 2.3). Persistence of phase decreased with increasing marker distance at a rate comparable to that observed for the decrease in average r^2 with increasing marker spacing. The slope of the decline was lower for the correlation between Landrace and Yorkshire when compared to other breed comparisons. Applying a z-test with Fishers transformation (Cohen *et al.*, 2003) to the correlation of phase

at <10 kb, the correlation of phase between Landrace and Yorkshire was significantly larger ($p < 0.001, n = 1520$) than all other breed combinations. Results for the correlation of phase were not significantly different ($p > 0.05, n = 1520$) in the Duroc-Hampshire, Duroc-Landrace, Duroc-Yorkshire, Hampshire-Yorkshire, and Hampshire-Landrace pairings (Table 2.3). For these five population comparisons, the average proportion of SNP with r having opposite sign ranged between 9-11% for SNP spaced within 10 kb and up to 45-49% for SNP spaced between 4.9 and 5 Mb (Table 2.3). In general, the estimates of r with reversed sign for the Landrace-Yorkshire were lower ranging from 9% to 45%. These results suggested a closer population relationship between the Landrace and Yorkshire populations than among all other populations.

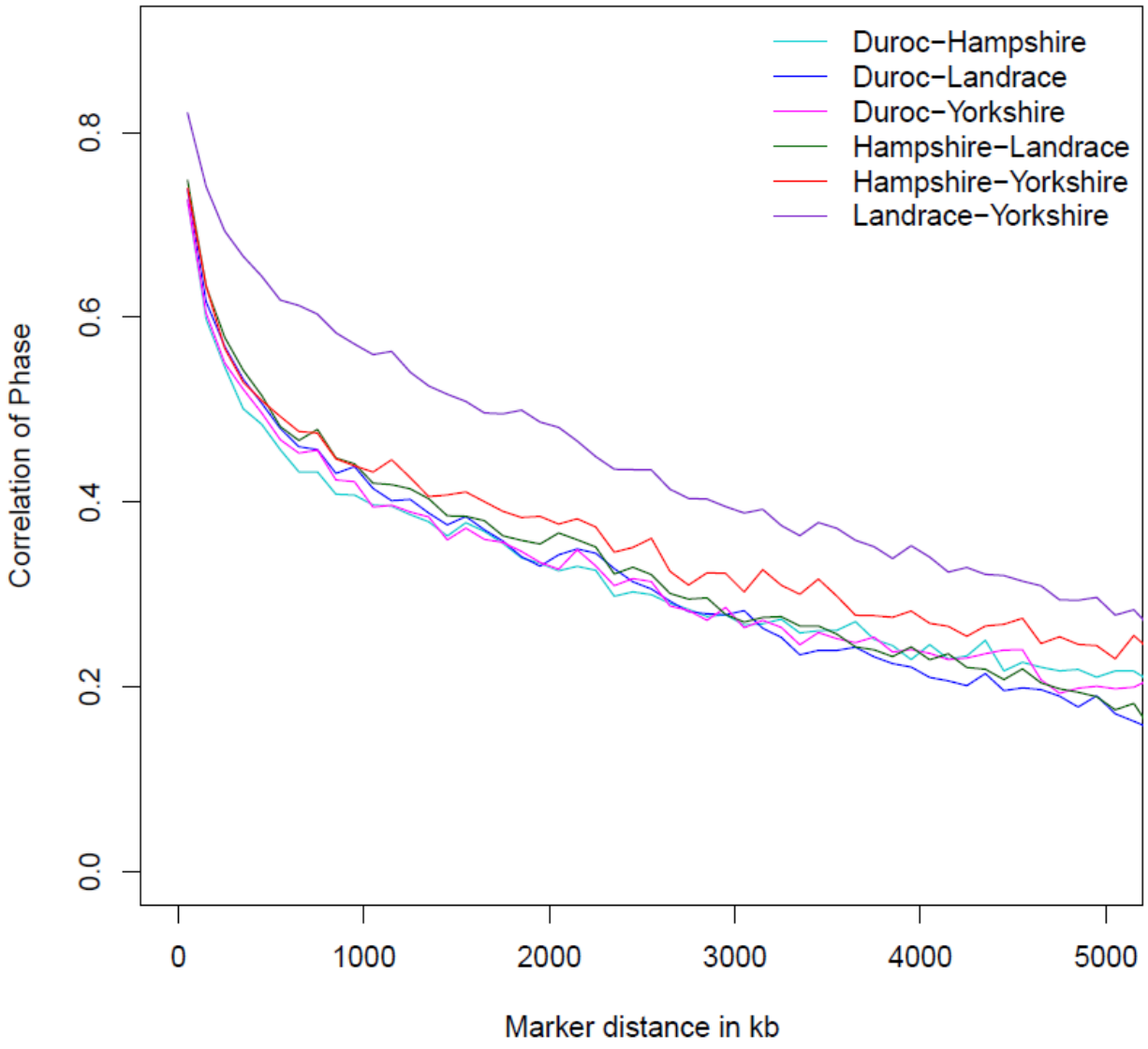

Figure 2.2: Correlation of gametic phase compared across breeds over distance Correlation of Phase between breeds for SNP pairs grouped by distance in intervals 100 kb long covering 0 to 5 Mb across the genome.

Table 2.3: Pairwise breed comparison of correlation of phase and proportion of phase agreement at various distances

Breeds Compared	Distance[1]	Proportion of r with opposite sign[2]	Correlation of $r_{ij(k)}$ and $r_{ij(k)}$ [3]
Duroc Hampshire	0-10kb	0.107	0.875
	10-50kb	0.184	0.762
	50-100kb	0.246	0.668
	0.9-1Mb	0.391	0.408
	4.9-5Mb	0.469	0.21
Duroc Landrace	0-10kb	0.108	0.872
	10-50kb	0.186	0.773
	50-100kb	0.251	0.681
	0.9-1Mb	0.395	0.438
	4.9-5Mb	0.485	0.19
Duroc - Yorkshire	0-10kb	0.104	0.87
	10-50kb	0.195	0.761
	50-100kb	0.252	0.67
	0.9-1Mb	0.396	0.422
	4.9-5Mb	0.468	0.201
Hampshire - Landrace	0-10kb	0.099	0.882
	10-50kb	0.184	0.776
	50-100kb	0.242	0.697
	0.9-1Mb	0.392	0.441
	4.9-5Mb	0.475	0.189

Table 2.3: (cont'd)

Breeds Compared	Distance[1]	Proportion of r with opposite sign[2]	Correlation of $r_{ij(k)}$ and $r_{ij(k)}$ [3]
Hampshire - Yorkshire	0-10kb	0.113	0.871
	10-50kb	0.189	0.771
	50-100kb	0.249	0.686
	0.9-1Mb	0.389	0.439
	4.9-5Mb	0.459	0.245
Landrace - Yorkshire	0-10kb	0.087	0.921
	10-50kb	0.16	0.842
	50-100kb	0.204	0.783
	0.9-1Mb	0.353	0.571
	4.9-5Mb	0.448	0.297

[1]Interval length in kb
[2]Proportion of SNP pairs having r with reversed sign within the interval
[3]Correlation of phase between two breeds (k and k) within the given interval

2.3 Discussion

2.3.1 Extent of Linkage Disequilibrium

Current effective population size of the breeds used in this study was previously estimated, using pedigree information, to be between 74 (Landrace) and 113 (Duroc, Yorkshire, (Welsh *et al.*, 2010)). Consistent with having the largest current effective population size, we find that long range r^2 (10 Mb, Fig. 2.1) estimated from our data was smallest for Duroc and Yorkshire (0.035, 0.03). In Hampshire, a smaller effective population of 109 corresponded to

higher r^2 at 10 Mb (0.046). Due to the similar long range r^2 (0.035) at 10 Mb we would have expected the Landrace population to have an effective population size comparable to that of Duroc and Yorkshire. However, using pedigree data Welsh *et al.* (2010) estimated the current effective population size of Landrace to be 74. The reason for this discrepancy remains unknown. Several previous studies investigated LD in pigs using reduced numbers of microsatellite markers and fewer animals from commercial populations (Harmegnies *et al.*, 2006; Nsengimana *et al.*, 2004). Nsengimana *et al.* (2004) found relatively large estimates of LD (D) from 0.29 to 0.41 using 15 microsatellite markers. In contrast, using r^2 instead of D and thereby correcting for minor allele frequency, Harmegnies *et al.* (2006) found r^2 ranging from 0.15 to 0.19 for marker distance < 1 cM and 0.10-0.12 for markers spaced between 1 cM and 5 cM, using 29 microsatellite markers on SSC15, comparable to our results of r^2 between 0.16-0.22 for marker spaced between 1 and 5 Mb. Du *et al.* (2007) estimated r^2 from 4,500 SNP markers in six commercial lines of pigs and found estimates of average $r^2 = 0.51$ for markers less than 0.1 cM apart, and estimates of 0.21 and 0.07 at marker distances of 1 cM and 5 cM respectively. Similarly, our populations had average r^2 of 0.15 to 0.20 and 0.05 to 0.08 at marker distances of 1 Mb and 5 Mb, respectively. A recent study conducted by Uimari and Tapio (2011) used the same genotyping platform as our study to estimate r^2 and effective population size in Finnish Landrace and Yorkshire populations. Uimari and Tapio (2011) found average r^2 of 0.43 and 0.46 for adjacent markers in the Finnish Landrace and Yorkshire populations, respectively, which was higher than our results of 0.36 for Landrace and 0.39 for Yorkshire. In addition, Uimari and Tapio (2011) reported that the r^2 for markers spaced at 5 Mb decreased to 0.09 and 0.12 in the Finnish Landrace and Yorkshire breeds, respectively. In the present study, r^2 declined further to 0.05-0.06 at 5 Mb marker spacing for Landrace and Yorkshire (Table 2.1). The higher average r^2 for distant ($r^2 > 0.2$ for 1

Mb) markers in the Finnish populations could be explained by smaller effective population size of the Finnish populations, causing higher r^2 on average. This is partially confirmed by comparing the estimated effective size of the Finnish populations ($N_e = 91$, 61 for Landrace and Yorkshire, respectively) (Uimari and Tapio, 2011), to estimated effective population sizes of the populations used in the current study reported by Welsh *et al.* (2010) ($Ne = 74$, 113 for Landrace and Yorkshire, respectively), where the current effective population size for Finnish Yorkshire is approximately half that of our Yorkshire population. Compared to recent estimates from Canadian populations we found estimates of average r^2 for markers with pairwise distance below 100kb to be consistent in Landrace (US: 0.34, Canadian: 0.31) and Yorkshire (US: 0.37, Canadian: 0.32; Jafarikia *et al.*, 2010). However, in Duroc estimates of average r^2 for markers with pairwise distance below 100kb were considerably higher in the US population (0.42) compared to the Canadian population (0.31, Jafarikia *et al.*, 2010).

2.3.2 Persistence of Phase

Persistence of phase can be used to infer upon the history of a species and relatedness of breeds within that species as well as on reliability of across population GWA and GEVB prediction (de Roos *et al.*, 2008). Persistence of phase was previously reported for three Canadian swine breeds (Duroc, Landrace, Yorkshire, Jafarikia *et al.*, 2010). For SNP with pairwise distance below 50kb we estimated persistence phase to be 0.88 between Landrace and Yorkshire and 0.82 for both Landrace and Yorkshire with Duroc. In the Canadian breeds persistence of phase also indicates a closer relationship between Landrace and Yorkshire (0.82) and a more distant relationship between Landrace/Yorkshire and Duroc (Jafarikia *et al.*, 2010). We found correlation of phase of 0.82 for Landrace/Yorkshire with Duroc, while the Canadian breeds had 0.66/0.65, indicating less agreement of phase even at short

pairwise distance (Jafarikia *et al.*, 2010). Our results showed that correlation of phase for the pig breeds in this study ranged between 0.87 for Duroc-Yorkshire and 0.92 for Landrace-Yorkshire for markers with pairwise distance <10 kb. Previous research in Australian cattle breeds (de Roos *et al.*, 2008) showed correlation of phase between 0.68 for Australian Angus-New Zealand Jersey to 0.97 for Dutch Holstein-Black and White. At increasing marker distance, correlation of phase for the pig breeds in this study decreased (range in r: 0.41 to 0.57) at an average pairwise marker distance of 1 Mb. This decrease however was less than the decrease de Roos *et al.* (2008) observed in all but two of the cattle breeds they considered (< 0.4 for markers spaced 1 Mb). While correlation of phase was similar between these pig breeds and dairy cattle at short range (<10kb), the pig breeds showed generally larger correlation of phase than the dairy cattle de Roos *et al.* (2008) at increasing marker distances. If two populations diverged from a common ancestral population, their correlation of phase can be expressed as $r_0^2(1-c)^{2T}$, where r_0^2 is a measure of LD in the common ancestral population, c is the recombination distance between markers, and T is time since breed divergence in generations (Sved *et al.*, 2008). For markers as close as 10 kb the recombination distance c will be almost 0, so that correlation of phase at those short distances can serve as an estimation of r_0^2 in the common ancestral population. Since correlation of phase was comparable in the pig populations (0.87-0.92) for markers with pairwise distance below 10 kb to that reported in Australian cattle (0.80-0.97, de Roos *et al.*, 2008), LD in the common ancestral pig population is likely to be similar to that in the common ancestral population of Australian cattle breeds. Larger correlation of phase at increasing marker distance (1 Mb) in the pig populations used in this study (0.41-0.57) compared to Australian cattle breeds (< 0.40) suggests that T is smaller in our pig breeds than it is in the cattle breeds. The expected correlation of r between two breeds can be expressed as e^{-2cT} (de Roos

et al., 2008). To estimate the time since breed divergence for the pig breeds in this study we used SNP with pairwise distance between 10kb and 300kb, and estimated correlation of phase for each 2.5kb interval. We calculated the linear regression of the natural logarithm of the estimated correlation of phase onto the average pairwise distance c. The slope of this regression is an estimate of $-2T$. Consequently, the slope divided by -2 is the number of generations (T) since these two breeds have diverged (de Roos *et al.*, 2008). Results suggest that the pig breeds in this study diverged approximately 40-66 generations ago. The expected correlation of phase would decrease to 0.41 and 0.02 at 1 cM and 5 cM distance respectively in the Yorkshire-Landrace comparison, assuming T of 40 and r_0^2 of 0.92. We observed a correlation of phase of 0.57 and 0.30 at 1 Mb and 5 Mb, respectively, between these two breeds, indicating that a T of 40 may overestimate the actual time since breed divergence. One possible cause of this observation is admixture between these two breeds, causing more common LD between them than what would be expected from fully diverged breeds (de Roos *et al.*, 2008). We obtained the date of herd book closure for each of the breeds in this study, and assuming a generation interval of approximately 2 years (Welsh *et al.*, 2010), Duroc, Hampshire, Landrace, and Yorkshire have existed as distinct breeds for at least 38.5, 44.5, 31.5, and 30.5 generations, respectively. The time of herd book closure does not directly indicate the time since breed divergence, since distinguishable breeds must have existed before herd book closure. Nevertheless, the time of herd book closure further supports our observation that Landrace and Yorkshire have developed as separate breeds later than Duroc and Hampshire.

2.3.3 Implications of estimated levels of LD for GEBV implementation

Our results have several important implications for future implementation of genomic selection in swine. Accuracy of prediction of genome wide marker assisted selection can be directly affected by the chosen marker density (resulting in average r^2 between markers and QTL), and the size of the training population (Hayes *et al.*, 2009a). The currently used marker panel, containing approximately 40,000 usable markers, had average r^2 of approximately 0.4 between adjacent markers for all four breeds. That exceeds the level of $r^2 = 0.2$ simulated by Meuwissen *et al.* (2001) to reach prediction GEBV accuracy around 0.85. Furthermore, our results indicated that reducing the original marker panel to 10% of the markers (3,000-4,000 SNP) still resulted in average r^2 for adjacent markers exceeding 0.2 in all four breeds. On the other hand, recent research in Australian Holstein Friesian cattle has shown (Moser *et al.*, 2010) that using subsets of 3000-5000 SNP to estimate direct genomic breeding values (DGV) could only reach 80% of the prediction accuracy previously estimated using approximately 42,000 SNP. Such a reduction in prediction accuracy will be unacceptable for most practical implementations. However, the accuracy of GEBV predicted by low density panels can be increased through the use of genotype imputation (Weigel *et al.*, 2010a), where high density genotypes are imputed using low density SNP genotypes and a high density reference panel of haplotypes (Browning and Browning, 2009). Weigel *et al.* (2010b) used approximately 10% of 2,693 SNP from Bos Taurus chromosome 1 to impute the full SNP set in a Jersey population. They found that using a high density reference genotype panel ($n = 2,542$ animals), the imputation accuracy of the non-typed markers was between 0.86 and 0.94. Average r^2 in our populations ranged from 0.36 to 0.48 for markers

less than 100 kb apart, comparable to average $r^2 = 0.38$ for markers spaced at <100 kb in the Jersey population (Villa-Angulo *et al.*, 2009). Assuming a comparable decline of LD for increasing marker distance between the Jersey population and our pig populations, we would expect to accurately impute approximately 90% of the high density genotypes, using a low density panel containing 10% of the markers. More recent results reported even higher average accuracy of imputation (approximately 95%) when imputing 42,000 SNP in the Bovine 50K using the 3K subset in Holstein cattle (Johnston *et al.*, 2011). To assess the accuracy of GEBV estimated from imputed genotypes Weigel *et al.* (2010a), used the same Jersey population from their previous study (Weigel *et al.*, 2010b), and they found that the accuracy of GEBV based on imputed markers was 95% of the accuracy of the GEBV estimated using the observed genotypes (Weigel *et al.*, 2010a). As noted above average r^2 is similar between the American Jersey population and our pig populations, suggesting that future research in genomic selection in swine should explore the use of imputed low density genotypes to increase cost efficiency. Previous research in humans (Huang *et al.*, 2009), and European Holstein cattle (Dassonneville *et al.*, 2011) indicated that combining haplotypes from closely related populations can increase the accuracy of genotype imputation, while research in sheep suggests that breed specific reference haplotypes would yield better accuracy (Hayes *et al.*, 2012). The success of combined haplotypes for genotype imputation depends on the relatedness between the populations. Further research is necessary to determine if persistence of phase is large enough in our pig populations to increase imputation accuracy when combining reference haplotypes across breeds. As noted by Goddard (2008), the accuracy of GEBV prediction can be expressed as a function of the LD of between marker and QTL and the accuracy of estimated SNP effects. The loss in accuracy of GEBV prediction caused by imputing instead of observing genotypes could be compensated by increasing the

number of animals used to estimate SNP effects. If not enough animals are available for the estimation of SNP effects, animals from different, but closely related, populations could be combined to estimate SNP effects for GEBV prediction in both populations (Goddard *et al.*, 2006; Ibánez-Escriche *et al.*, 2009). The squared short-range (<10 kb) correlation of phase can also serve as the accuracy with which we can predict a marker-QTL association in one population using known marker-QTL associations from another population. For the pig breeds reported in this study the squared correlation of phase for close markers (0-100kb) ranged from 0.53 to 0.67. To evaluate whether these accuracies would warrant the use of a combined training population to estimate SNP effects accurately for both populations we refer to a simulation study conducted by de Roos *et al.* (2009) estimating the accuracy of GEBV prediction for combined training populations of highly, moderately and lowly related populations. Correlation of phase for populations diverged approximately $T = 30$ generations ago was reported to be below 0.80 for markers with pairwise distance below 0.055 cM (de Roos *et al.*, 2009). We found correlation of phase between Landrace-Yorkshire of around 0.80 at a corresponding marker distance. de Roos *et al.* (2009) concluded that reliability of GEBV prediction could be increased between 0.05-0.10 points in two populations, when approximately 40,000 marker genotypes are available, heritability is $h^2 = 0.3$ or higher, 1000 animals from each population were used to estimate SNP effects, and under the assumption that QTL effects are the same for both populations (de Roos *et al.*, 2009). In addition, they found that for genetically distant populations, at least 1,000 animals with genotypes and phenotypes available in each population were needed to avoid a decrease in the reliability of prediction (de Roos *et al.*, 2009). When SNP effects estimated in one population are used to calculate GEBV for another population which diverged approximately $T = 30$ generations ago, the reliability of the predicted GEBV was 0.65 assuming both high marker

density ($M = 40,000$) and heritability $h^2 = 1$ (de Roos *et al.*, 2009). Consequently, combining animals into a multi-breed panel to estimate SNP effects is likely to be only marginally beneficial for the pig breeds in this study, given the estimated correlation of phase and the large number of animals and markers required (de Roos *et al.*, 2009).

2.4 Conclusions

We used the PorcineSNP60 chip (Ramos *et al.*, 2009) to obtain high density genotypes (34,000-40,000 SNP) from pig trios in four breeds. From this data we estimated r^2 as a measure of LD across the genome as well as correlation of r, which measures phase agreement between breeds. We found r^2 of approximately 0.4 for markers less than 100 kb apart, which is higher than comparable estimates reported for North American Holstein cattle (Sargolzaei *et al.*, 2008) as well as various Australian cattle breeds (de Roos *et al.*, 2008). The same was true for average r^2 between markers with pairwise distance larger than 1 Mb, indicating a smaller past effective population size of these pig breeds. We also report a relatively slow rate of decay of LD over distance, observing r^2 around 0.2 at 1 Mb. The comparably high long range LD is an indicator that good accuracy can be expected for future implementations of GEBV in pigs using 10% (3,000-4,000) of SNP used in the current assay or less, along with genotype imputation. We would encourage future research in genomic selection in swine to especially focus on the possible benefits of the combined use of reduced marker panels and genotype imputation. To successfully promote the use of genomic selection in swine it will be necessary to increase cost efficiency while maintaining high accuracy of prediction. Currently no low density panels for SNP genotyping are publicly available for swine, but the presented results will be available to aid in the development of efficient SNP platforms.

Relatively low persistence of phase reported here implies that the use of multi-breed panels estimating SNP effects for genomic selection will likely be limited, especially when using low density genotypes, but the merit of combining reference haplotypes for genotype imputation should be further investigated.

2.5 Methods

2.5.1 Sample Design

For this study sire/dam/offspring trios of the Duroc, Hampshire, Landrace and Yorkshire breeds were selected from the National Swine Registry (NSR) pedigree. Selected parents were unrelated for at least two generations. All animals were genotyped using the Illumina PorcineSNP60 (Number of markers $M = 62,163$) Genotyping BeadChip (Illumina Inc.) (Ramos *et al.*, 2009) at a commercial laboratory (GeneSeek, a Neogen Company, Lincoln, NE). All SNP showing Mendelian inconsistencies for a trio were set missing in that particular trio. If one or more animals within a trio had missing genotypes in more than 10% of the SNP that trio was eliminated from further analysis. Similarly, SNP were removed if they did not have genotypes available for at least 90% of the samples across all breeds ($M_{CallRate} < 0.9 = 5080$). Only autosomal SNP were considered in this study, leading to the exclusion of all SNP with an uncertain map position on build 10 of the pig genome sequence, as well as SNP on the sex chromosomes ($M_{non-autosomal} = 9308$). To exclude non-segregating SNP from the analysis we removed markers with minor allele frequency (MAF) below 5% within each breed separately. The number of fixed SNP varied substantially between breeds: we excluded $M_{MAF} < 5\% = 13,646$ SNP in Duroc, $M_{MAF} < 5\% = 15,405$ SNP in Hampshire, $M_{MAF} < 5\% = 7,631$ SNP in Landrace, and $M_{MAF} < 5\% = 8,665$ SNP in

Yorkshire. Additionally, SNP were excluded for failure to meet Hardy Weinberg Equilibrium ($p < 0.001$) within breeds causing $M_{HWE} < 0.001 = 117$, 85, 146, and 176 SNP to be discarded in Duroc, Hampshire, Landrace, and Yorkshire respectively. After applying the described filtering criteria, a total of 30, 26, 29, and 32 trios were included for the Duroc, Hampshire, Landrace and Yorkshire breeds, respectively. And a total of 34,129, 32,370, 40,144 and 39,110 SNP spaced at an average distance of 70, 74, 60 and 61 kb satisfied the SNP selection criteria for Duroc, Hampshire, Landrace and Yorkshire, respectively.

2.5.2 Estimation of average LD and persistence of phase

Haplotypes were obtained for the founder animals using the trio option of BEAGLE (Browning and Browning, 2009), phasing the genotypes by chromosome. Sampling animals in trios was shown to yield improved accuracy of estimated haplotypes (Marchini *et al.*, 2006). To further increase haplotype accuracy, BEAGLE was set to run 100 iterations of the phasing algorithm and sample 100 haplotype pairs for each individual per iteration. Additionally, a short simulation experiment was conducted showing that for MAF above 5% average r^2 can be reliably estimated from the current sample size (results not shown). Alleles for each SNP were re-coded as 0/1, keeping the reference allele constant across all four populations, allowing for later determination of phase agreement. Haplotypes and code needed to reproduce these results are publicly available at `https://www.msu.edu/~steibelj/JP_files/LD_estimate.html`. For all pairs of SNP r^2 was estimated, using allelic frequencies of the founding animals, according to the following equation:

$$r^2_{ij} = \frac{(p_{ij} - p_i p_j)^2}{p_i(1-p_i)p_j(1-p_j)} \tag{2.1}$$

where p_i, p_j are the marginal allelic frequencies at the i^{th} and j^{th} SNP respectively and p_{ij} is the frequency of the two-marker haplotype (Devlin and Risch, 1995), using the freely available software R (Team, 2011). Marker pairs were grouped by their pairwise physical distance into intervals of 100 kb starting from 0 up to 10 Mb. Average r^2 for SNP pairs in each interval was estimated as the arithmetic mean of all (Equation 1), with the pairwise distance between the i^{th} and j^{th} element of the currently considered interval:

$$\bar{r}^2 = \frac{1}{\sum_{i=1}^{18}(M_l - 1)} \sum_{i=2}^{M_l - 1} r^2_{i,i+1} \tag{2.2}$$

where $\bar{r}^2$ is the average of all adjacent SNP across 18 autosomes (l), with M_l SNP per chromosome. To estimate average r^2 between adjacent markers for different marker densities a certain percentage of markers (50%, 75%, 90%, 95%, 99%, and 99.5%) were removed before average r^2 was estimated using equation 2. To select markers, an increasing proportion of SNP were sequentially removed solely considering their map position, so that for instance: to reduce a panel to 50%, every second marker was kept for analysis, for 25% every fourth was kept and so on. To estimate persistence of phase only markers with minimum MAF of 5% in all breeds were included in the analysis, resulting in 22,340 common SNP across all breeds. Correlation of phase was estimated for intervals of 100 kb (from 0 to 10 Mb). We excluded markers with pairwise distance above 10 Mb to decrease the computational load. Estimates of average r^2 at larger distances are close to zero, which would cause correlation of phase to be close to zero as well. Persistence of phase was then estimated as:

$$R_{k,k\prime} = \frac{\sum_{(i,j)\in p}(r_{ij(k)} - \bar{r}_{(k)})((r_{ij(k\prime)} - \bar{r}_{(k\prime)})}{s_{(k)}s_{(k\prime)}} \tag{2.3}$$

where $R_{k,k'}$ is the correlation of phase between $r_{ij(k)}$ in population k and $r_{ij(k)}$ in population k, $s_{(k)}$ and $s_{(k)}$ are the standard deviation of $r_{ij(k)}$ and $r_{ij(k)}$ respectively, and $\bar{r}_{(k)}/\bar{r}_{(k')}$ are the average $r_{i}j$ across all SNP i and j within interval p for population k and k respectively.

Chapter 3

Methods of tagSNP selection and other variables affecting imputation accuracy in swine

Badke, Y. M., Bates, R. O., Ernst, C. W., Schwab, C., Fix, J., Van Tassell, C. P., & Steibel, J. P. (2013). Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC genetics, 14(1), 8.

3.1 Background

Recent advances in genotyping technology have facilitated the availability of high density genotyping platforms in many livestock species. High density platforms including several thousand single nucleotide polymorphisms (SNP) are available for cattle (Boichard *et al.*, 2012; Matukumalli *et al.*, 2009; VanRaden *et al.*, 2011), chicken (Groenen *et al.*, 2009), sheep (Archibald *et al.*, 2010), and pig (Ramos *et al.*, 2009).

These platforms can be used to increase the efficiency and accuracy of breeding programs by implementing genomic selection (Hayes *et al.*, 2009a; Meuwissen *et al.*, 2001). Using SNP data to inform breeding decisions allows animal breeders to select breeding stock prior to the animals having progeny of their own, thereby accelerating genetic progress through shortened generation intervals (Hayes *et al.*, 2009a; Meuwissen *et al.*, 2001).

Currently, genomic selection has been successfully implemented in dairy cattle based on genotypes from the Illumina BovineSNP50 chip (Hayes *et al.*, 2009a). In an effort to increase cost efficiency, the use of low density (tagSNP) genotyping platforms was exploited for dairy cattle (Dassonneville *et al.*, 2011; Weigel *et al.*, 2010a). If high density genotypes are imputed from tagSNP with high accuracy, the loss of reliability of predicted genomic breeding values is minimal (Berry and Kearney, 2011; Dassonneville *et al.*, 2011; Weigel *et al.*, 2010a). High accuracy of imputed genotypes depends on the selection of tagSNP, as well as the composition and size of the reference panel of haplotypes used for imputation.

If close relatives of all imputation candidates are genotyped at high density, untyped markers can be recovered through linkage and segregation analysis (Habier *et al.*, 2009), where haplotypes can be traced through generations of directly related individuals using the rules of Mendelian inheritance. However, in some species it may not be feasible to genotype

a large proportion of the pedigree at high density. In that case a small panel of reference haplotypes can be used to impute all untyped markers by exploiting population-wide linkage disequilibrium (LD) (Browning and Browning, 2009; Scheet and Stephens, 2006). This approach was initially proposed in human genome-wide association studies (GWAS) and has recently found application in plant (Hickey *et al.*, 2012) and animal breeding (Berry and Kearney, 2011; Hayes *et al.*, 2012; Weigel *et al.*, 2010a). A combination of imputation based on segregation analysis and population-wide LD is currently being used in dairy breeding (Dassonneville *et al.*, 2011). While combining both approaches will increase accuracy of imputation, eventually becoming the default method, cost-effective implementation of genomic selection in novel populations is likely to initially rely more on LD based imputation. Consequently, in this paper we will concentrate on LD based imputation by investigating tagSNP selection and haplotype reference panel construction.

Human geneticists have proposed a variety of approaches to select an optimal low density set of tagSNP to achieve cost efficient imputation in GWAS (He and Zelikovsky, 2007). These approaches include statistical criteria based on a pairwise threshold of LD between SNP (e.g. Qin *et al.*, 2006) and predictive ability, selecting tagSNP that provide the most accurate prediction of all non-typed markers (He and Zelikovsky, 2006). On the other hand, tagSNP sets used in livestock are mainly selected for equidistant spacing based on physical position along the genome, and high minor allele frequency (MAF) to ensure segregation (e.g. Boichard *et al.*, 2012).

Crucial to successful implementation of genotype imputation using population wide LD is the availability of a representative panel of reference haplotypes (Howie *et al.*, 2011; Huang *et al.*, 2009). These panels are commonly built by genotyping a small number of trios or a larger number of relatively unrelated individuals. The overall goal in either case is to

collect genotypes that can be accurately phased (Marchini *et al.*, 2006) into haplotypes representative of population frequencies. As a result, we began our study by genotyping and phasing a small number of trios in four US pig breeds (Badke *et al.*, 2012, $N_{Trios} \sim 30$) and further enriching this panel for the Yorkshire breed with a set of high density genotypes from largely unrelated individuals ($N_{samples} = 889$).

The objective of this study was to develop guidelines for the implementation of genotype imputation in livestock populations having little or no prior use of genome-wide marker-assisted-selection. First, we compared imputation accuracy resulting from three methods of tagSNP selection using Yorkshire pigs genotyped with a high density SNP set (Illumina PorcineSNP60). This includes a report on imputation accuracy of the recently developed commercially available 9K tagSNP set referred to as the GeneSeek Genomic Profiler for Porcine LD (GGP-Porcine, GeneSeek a Neogen Company, Lincoln, NE). Second, we assess accuracy of imputation based on an increasing number of reference haplotypes to inform the selection of an optimal reference panel of haplotypes. Finally, we discuss imputation accuracy as a function of chromosomal location and MAF of non-observed SNP.

3.2 Methods

3.2.1 Genotypes

High density genotypes for approximately 30 sire/dam/offspring trios were obtained and phased for each of four breeds of pigs (Duroc, Hampshire, Landrace, Yorkshire) in a previous study (Badke *et al.*, 2012). To ensure accurate phasing, the reference panel for imputation used in this study was the 128 haplotypes from the Yorkshire sire/dam pairs previously genotyped as parents in those trios. Animal protocols were approved by the Michigan State

University All University Committee on Animal Use and Care (AUF# 03/09-046-00). The haplotypes of these animals are freely available at `https://www.msu.edu/~steibelj/JP_files/LD_estimate.html`.

Detailed information about data cleaning procedures, descriptive statistics of LD, and correlation of phase between Yorkshire and other US pig breeds can be found in Badke *et al.* (2012). In addition, DNA samples were collected from 920 Yorkshire pigs and sent to a commercial laboratory (GeneSeek, a Neogen Company, Lincoln, NE) to be genotyped on the Illumina PorcineSNP60 (Number of markers M=62,163) Genotyping BeadChip (Illumina Inc.) (Ramos *et al.*, 2009). Only animals with more than 90% genotype call rate were considered for analysis, resulting in 889 animals used as the testing panel for this study. All SNP included in the 128 haplotype Yorkshire reference panel were used for analysis. All data from this study is available at `https://www.msu.edu/~steibelj/JP_files/imputation.html`.

In our previous study (Badke *et al.*, 2012) we reported breed specific LD and persistence of phase among breeds for Duroc, Hampshire, Landrace, and Yorkshire pigs. We found that persistence of phase between Yorkshire and the other breeds ranged between 0.42 and 0.57 for SNP spaced approximately 1MB apart (Badke *et al.*, 2012). As a result the amount of LD within the Yorkshire breed that could be recovered through haplotypes from another breed ranges between 0.18 and 0.33, such that adding haplotypes of a second breed to impute Yorkshire genotypes did not appear to be beneficial. For genomic selection, a simulation study conducted by de Roos *et al.* (2009) found that persistence of phase between breeds needs to be much larger than the reported value between Yorkshire and any of the other three breeds to implicate any advantage for the use of mixed breed training panels. For this reason we decided to use only Yorkshire haplotypes in the reference panel for imputation in

this paper.

3.2.2 Genotype imputation and estimation of imputation accuracy

All imputations in this study were done using BEAGLE version 3.3.1 (Browning and Browning, 2009), a genotype imputation software that uses a reference panel of haplotypes to estimate phase and impute missing genotypes in a set of unrelated individuals. Beagle was run separately for each chromosome using 128 reference haplotypes from the trio design (Badke *et al.*, 2012, phased file) to phase and impute genotypes of the 889 un-phased testing animals. All SNP, except tagSNP, were masked as missing in the testing set. Beagle was run for ten iterations of the phasing algorithm, drawing four samples per iteration. Previous results from another study (Hayes *et al.*, 2012), as well as a short experiment conducted in this study (data not shown) found no increase in imputation accuracy when the number of iterations or samples per iteration were increased. The output files from BEAGLE contained the most likely imputed genotypes (AA, AB, BB), posterior genotype probabilities ($P(AA)$, $P(AB)$, $P(BB)$), and posterior expected allelic dosage of the B allele derived from the posterior genotype probabilities (i.e. $0 * P(AA) + 1 * P(AB) + 2 * P(BB)$) (Browning, 2011).

Imputation accuracy was estimated using three different measures that reflect different influences of MAF and error counting. The proportion of correctly imputed alleles was computed as

$$IA = 1 - \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} |g_{ij} - \hat{g}_{ij}|}{2 * \sum_{i=1}^{M} N_i} \tag{3.1}$$

where g_{ij} is the observed allelic dosage of SNP i in individual j, $\hat{g}_{ij}$ is the corresponding posterior expected allelic dosage obtained from BEAGLE output, M is the total number of imputed SNP, and N_i is the number of individuals with called genotypes for SNP i. This overall measure of imputation accuracy can be further decomposed into SNP-specific accuracy ($IA_{i.} = 1 - \frac{\sum_{j=1}^{N_i} |g_{ij} - \hat{g}_{ij}|}{2 * N_i}$) and animal specific accuracy ($IA_{.j} = 1 - \frac{\sum_{i=1}^{M_j} |g_{ij} - \hat{g}_{ij}|}{2 * M_j}$). This measure of imputation accuracy will be biased upwards, especially for SNP with low MAF, because even if imputation ignores LD information and is based solely on allele frequency, the major allele would be correctly imputed for a large proportion of genotypes (Hayes *et al.*, 2012; Hickey *et al.*, 2012). As tagSNP density decreases, imputation accuracy of rare alleles further decreases as rare haplotypes become harder to identify due to longer sequences of SNP missing (Hickey *et al.*, 2012). Estimating the total percentage of correctly imputed alleles for SNP with low MAF will be biased due to the large number of correctly imputed major alleles masking the small number of misspecified minor alleles, which can be overcome through the use of a more sensitive measure of accuracy for these SNP (Hickey *et al.*, 2012). In addition, if individuals carrying the minor allele are not correctly identified and their phenotype cannot be matched for GWAS this relatively small proportion of incorrectly imputed alleles will further decrease power. A variety of measures have been introduced to obtain estimates of imputation accuracy unbiased by MAF (Hayes *et al.*, 2012; Hickey *et al.*, 2012; Zheng *et al.*, 2011). We estimated the proportion of correctly imputed alleles adjusted for MAF using the formula presented by Hayes *et al.* (2012):

$$IA_{MAF} = \frac{IA - IA_{Freq}}{1 - IA_{Freq}} \tag{3.2}$$

where IA is computed as described in equation (3.1) and IA_{Freq} is the accuracy of imputation based on genotypic frequencies estimated as:

$$IA_{Freq} = p(AA)_{ref} * p(AA)_{val} + p(AB)_{ref} * p(AB)_{val} + p(BB)_{ref} * p(BB)_{val} \quad (3.3)$$

where $p(AA)_{ref_i}$, $p(AB)_{ref_i}$, and $p(BB)_{ref_i}$ are the observed frequencies for genotypes AA, AB, and BB for SNP i in the reference haplotypes and $p(AA)_{val_i}$, $p(AB)_{val_i}$, and $p(BB)_{val_i}$ are the predicted genotypic frequencies in the testing population for SNP i. IA_{Freq} can be interpreted as the expected probability of correctly imputing a genotype in the testing population by assigning a randomly sampled genotype from the haplotypes in the reference panel. This measure was computed on a SNP-wise basis and averaged across all SNP. To account for a slightly different number of genotypes observed within each SNP (due to missing at random) the average was obtained by weighting the accuracy of each SNP by the number of individuals with observed genotypes within each SNP.

Alternatively, another measure of imputation accuracy robust to MAF is the squared correlation between the observed and imputed allelic dosage (Hickey *et al.*, 2012). The correlation was obtained on a SNP by SNP basis using the correlation function in R (Becker *et al.*, 1988). SNP wise correlation measures were weighted by the number of available observations within the SNP to obtain an overall average imputation accuracy.

3.2.3 Methods of tagSNP selection

TagSNP were selected using three approaches: 1) evenly spaced based on physical position, 2) based on minimum pairwise LD with non-tagSNP (statistical selection), and 3) based on marker predictive ability to accurately impute non-observed SNP genotypes (predictive

selection).

To select evenly spaced SNP the total length of each chromosome was partitioned into segments corresponding to the total number of tagSNP to be selected. Then, within each segment the SNP closest to the segment center was identified and added as a tagSNP. If a given segment was empty, no tagSNP was selected in that segment.

To implement a statistical search for tagSNP (He and Zelikovsky, 2006) we used the freely available software package FESTA (Qin *et al.*, 2006). FESTA performed a greedy search, where each SNP i was either an element of the tagSNP set or in LD higher than a threshold (r_t^2) with an existing element of the tagSNP set. FESTA was run repeatedly for increasing r_t^2 ranging from 0.1 to 0.9 in 0.1 increments using estimates of LD based on 128 reference haplotypes.

To implement predictive tagSNP selection, we applied the following forward search algorithm: First, we split the 64 Yorkshire reference animals into a randomly sampled set of 10 individuals (training set) and 54 individuals (reference haplotypes). Second, all SNP except one tagSNP in the training set were masked and imputed using the reference haplotypes. Third, accuracy of all imputed SNP was estimated and saved. Steps two and three were then repeated until all estimates of imputation accuracy were available for all potential tagSNP (at first, the potential tagSNP are all SNP on the chromosome). Fourth, the SNP that yielded the highest average accuracy of imputation among those not already chosen as tagSNP was selected as a new tagSNP. Steps two through four were repeated until the maximum number of tagSNP or a target imputation accuracy were reached. Because of the high computational demand of this methodology, this approach was only applied to the smallest available chromosome (SSC18), selecting tagSNP from 786 candidate SNP.

Concurrent to this research, a set of 9390 tagSNP (Release Date: April 2012) was as-

sembled by GeneSeek (Lincoln, NE) for the development of a commercial platform for low density genotyping in swine. This assay has been marketed as the GeneSeek Genomic Profiler for Porcine LD (GGP-Porcine; GeneSeek, Lincoln, NE). After production, the GGP-Porcine contains approximately 8500 tagSNP (Jeremy Walker, personal communication). TagSNP covering the entire genome were selected based on MAF in 13 commercial lines of pigs represented by four breeding companies and four purebred populations. The MAF were provided by the breeding companies (identified simply as company A, B, C, and D). The number of lines provided by these companies were 1, 1, 4, and 7. Additional estimates of MAF used to identify tagSNP were obtained from our previous study (Badke *et al.*, 2012) of four pure breeds: Duroc, Hampshire, Landrace, and Yorkshire. The freely available SNPspace software (C.P. Van Tassell, unpublished data) was used to select tagSNP. SNPspace was initially developed to select SNP for the Illumina BovineSNP50 beadchip (Matukumalli *et al.*, 2009). The conceptual framework of SNPspace is briefly described in that study (Matukumalli *et al.*, 2009), but additional features have been added since that time. Relative weights on lines or breeds of pigs ranged from 0.00625 to 0.25. SNPspace is based on a greedy algorithm, where SNP scores account for breed or line specific MAF, region of the genome, and position of SNP relative to previously selected tagSNP. Density of tagSNP was doubled within 5 Mbp of the chromosomal extremes, which has been shown to improve average accuracy of imputation compared to tagSNP evenly spaced across the entire chromosome (Boichard *et al.*, 2012; Dassonneville *et al.*, 2012).

3.2.4 Increasing reference panel size

To assess the effect of the number of reference haplotypes on imputation accuracy, we split the available sample of 889 Yorkshire pigs into two groups: 1) a 200 animal testing panel,

and 2) a 689 animal set of supplemental reference sires. Assignment to the two panels was random.

To obtain imputation accuracy for a decreased set of reference haplotypes, we split the original 128 reference haplotypes obtained from 64 Yorkshire animals into two groups of 64 reference haplotypes (corresponding to 32 animals) and estimated average imputation accuracy in the 200 animal testing set. Then, we split the two groups of 64 reference haplotypes further into two groups of 32 reference haplotypes and obtained four estimates of imputation accuracy that were averaged into a single measure.

Subsequently, we compared imputation accuracy using trio based reference panels to imputation accuracy based on randomly sampled reference panels. To this end, we randomly sampled 16 animals from the 689 animal supplemental reference set and continued to add individuals at random to obtain reference sets of 24, 32, 48, 64, 96, 128, 256, and 512 animals. Each of these sets was phased individually using BEAGLE (Browning and Browning, 2009) and then those haplotypes were used as reference panel to impute the 200 testing animals.

Finally, we assembled reference panels of haplotypes combining the original 128 haplotypes from trios, with an increasing number of supplemental reference sires. To form these reference panels 64, 128, 192, and 448 supplemental reference sires were randomly selected and phased using the trio haplotypes as a reference panel. Both, the trio reference haplotypes and an increasing number of supplementary reference haplotypes were then used to impute the 200 animal testing set.

Because imputation accuracy was constant across chromosomes (see Results, section 3.1) we conducted this experiment on chromosome SSC14, a medium sized chromosome that has uniform coverage of SNP across its length. We expect results to extrapolate to all other chromosomes.

3.3 Results

3.3.1 Comparison of methods for tagSNP selection

Due to the high computational demand, we initially performed a comparison of methods for tagSNP selection only on the smallest chromosome (SSC18). Statistical tagSNP selection requires fixing an r^2 threshold (r_t^2). Setting $r_t^2 = 0.2$ resulted in the selection of 165 tagSNP, which produced imputation accuracy of 0.936. Increasing r_t^2 to 0.3, led to a panel of 235 tagSNP and an increased imputation accuracy of 0.956. In comparison, imputation accuracy based on 165 and 235 tagSNP selected for predictive ability was 0.93 and 0.945, respectively. Direct comparison to tagSNP sets selected for even spacing is more difficult because of empty intervals, for which no tagSNP were selected, resulting in smaller than targeted tagSNP sets. The evenly spaced tagSNP sets closest in size to 165 and 235 tagSNP were as expected slightly smaller (161 and 224 tagSNP), and the resulting imputation accuracies were slightly lower than those obtained using the other sets (0.92 and 0.941, respectively). As expected, imputation accuracy increased with increasing densities of tagSNP regardless of the selection method (Figure 3.1). Statistically selected tagSNP performed slightly better than both, predictive and evenly spaced tagSNP (Figure 3.1), but all three methods resulted in similar imputation accuracy. Selection of tagSNP using predictive ability required an at least 500-fold increase in computation time for SSC18 compared to statistical and evenly spaced selection. However, results of imputation accuracy indicate that predictive tagSNP did not yield significantly higher imputation accuracy compared to tagSNP selected by other methods. Therefore, only statistical and evenly spaced tagSNP were selected in an exhaustive evaluation of imputation accuracy across all autosomes (Figure 3.1).

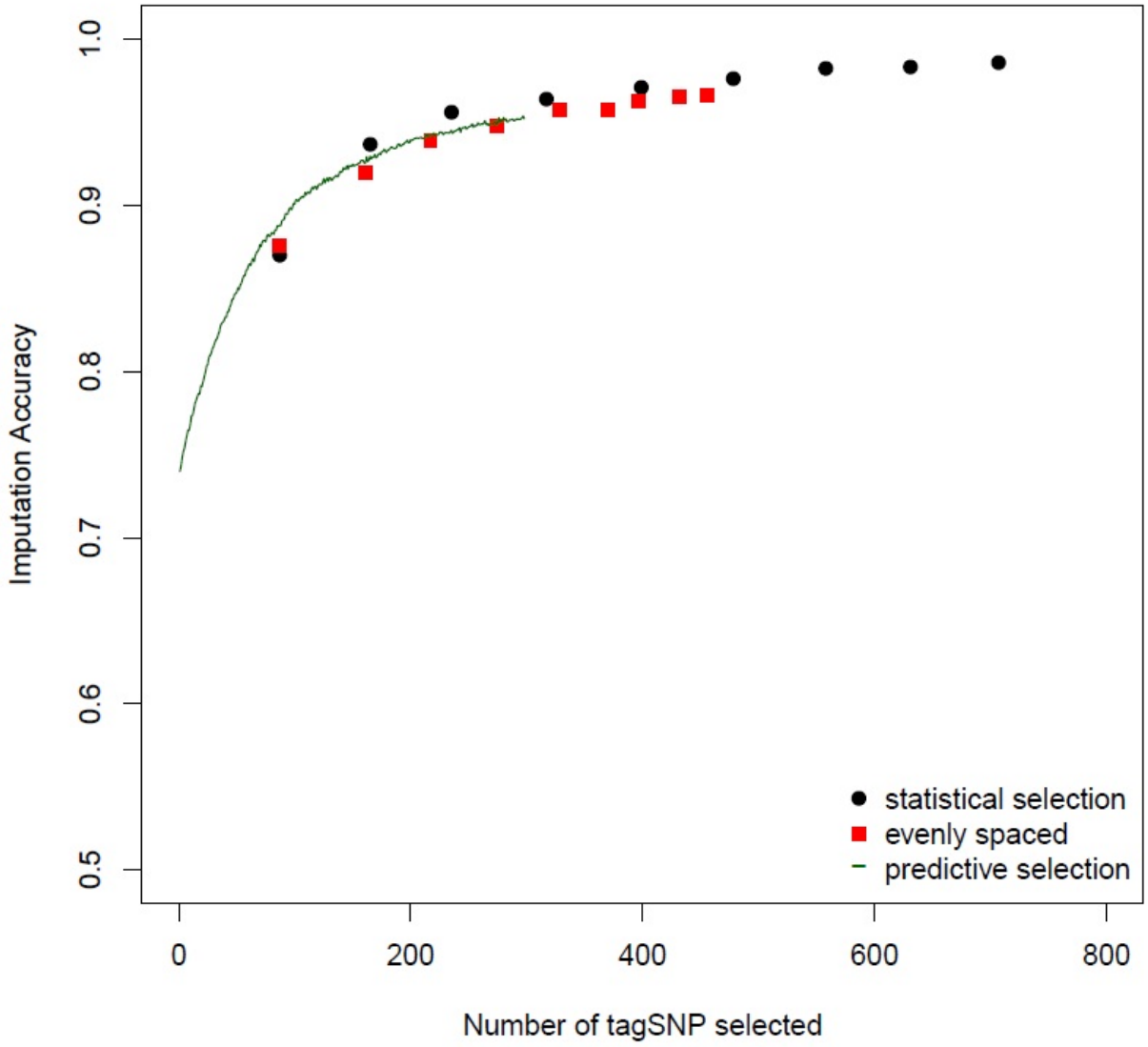

Figure 3.1: Imputation accuracy based on tagSNP selected using 3 different methods Average imputation accuracy (IA) as a function of the number of tagSNP selected using three methods of tagSNP selection for SSC18: 1) evenly spaced (red square), 2) statistical selection (black circle), or 3) predictive selection (green line).

When imputing across all autosomes, as observed on SSC18, imputation accuracy using statistically selected tagSNP was slightly higher than that using evenly spaced tagSNP (Figure 3.2). In particular, to attain imputation accuracy of 0.95, 7036 statistically selected tagSNP were necessary ($r_t^2 = 0.3$). In comparison, 10540 evenly spaced tagSNP were necessary to reach similar imputation accuracy. Imputation accuracy was virtually uniform across

chromosomes ranging from 0.92 to 0.94 for $r_t^2 = 0.2$ and from 0.94 to 0.96 for $r_t^2 = 0.3$.

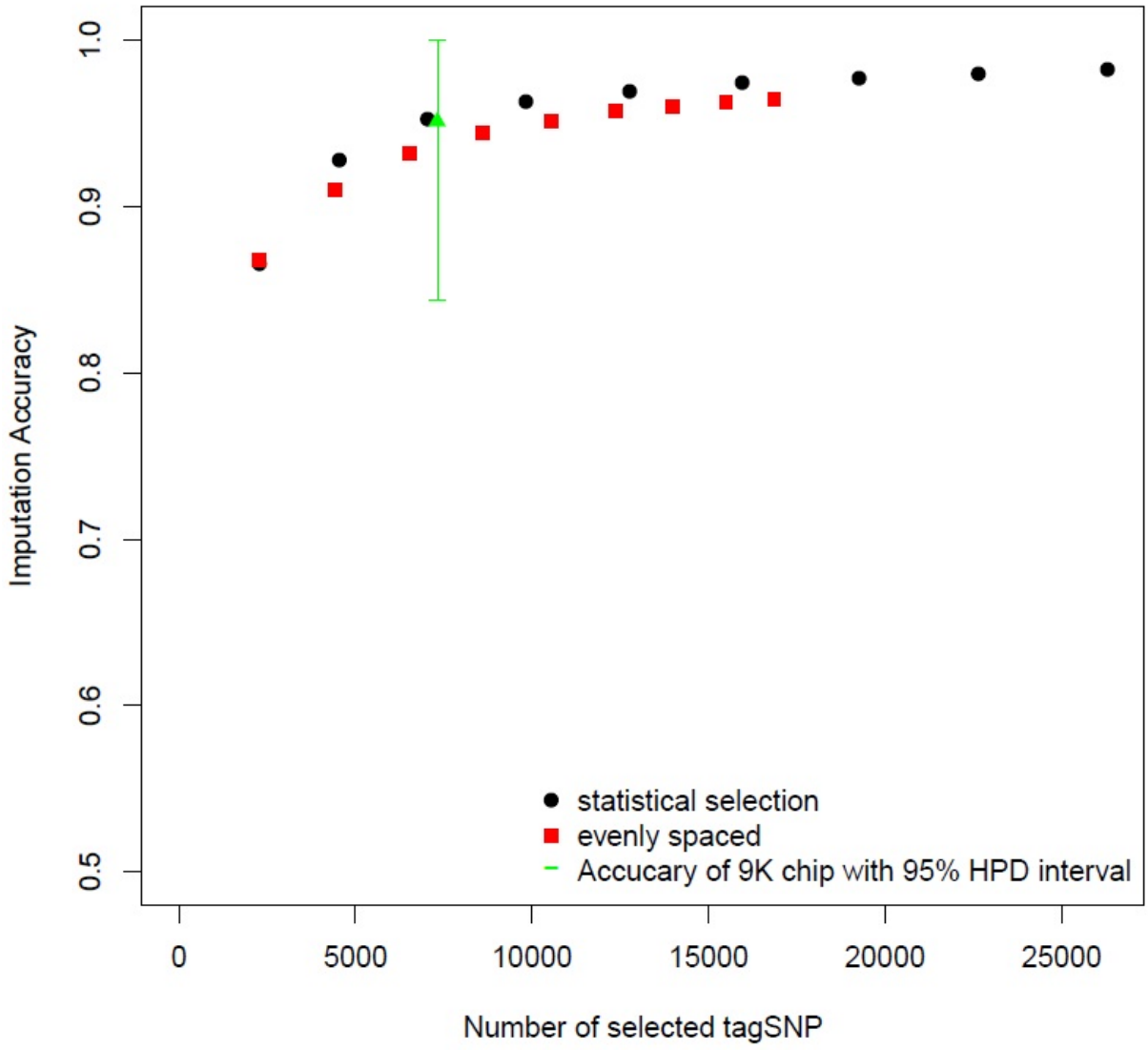

Figure 3.2: Imputation accuracy using evenly spaced or statistically selected tagSNP Average imputation accuracy (IA) as a function of the number of tagSNP selected for: 1) even spacing (red square), or 2) statistical selection (black circle) across all autosomes. Imputation accuracy for 7323 tagSNP from the commercial 9K tagSNP set (green triangle) with 95% highest posterior density interval.

We computed imputation accuracy based on 7323 tagSNP from the original list of 9K tagSNP provided by GeneSeek that passed quality control in this study ($MAF > 0.05$, $CallRate > 0.9$, assembled to an autosome under map build10) resulting in imputation

accuracy of 0.951 with a SNP-wise 95% highest posterior density interval equal to $[0.84, 1]$ (Figure 3.2). Accuracy of imputation using the commercial tagSNP was similar to that obtained using statistically selected SNP, at comparable density ($r_t^2 = 0.3$, $M_{tagSNP} = 7036$). The advantage of the proposed commercial platform is that it is not based on population specific LD, thereby making it applicable across swine populations. For this reason all subsequent results of imputation accuracy will be based on the tagSNP element of the Genomic Profiler for Porcine LD.

3.3.2 Imputation accuracy using the commercial 9K tagSNP set

To assess accuracy of imputation as a function of chromosomal location we plotted imputation accuracy of each individual SNP versus chromosomal position (Figure 3.3). SNP within 5% of the chromosomal extremes had on average slightly lower imputation accuracy (0.949) than the 10% in the center of the chromosome (0.972). As mentioned before, this property of imputation accuracy has previously been observed for other low density sets (Boichard *et al.*, 2012; Dassonneville *et al.*, 2012) and was anticipated during the tagSNP set design. Based on these reports (Boichard *et al.*, 2012; Dassonneville *et al.*, 2012), the density of tagSNP was approximately doubled within 5 Mbp of the chromosomal ends in the commercial 9K tagSNP set.

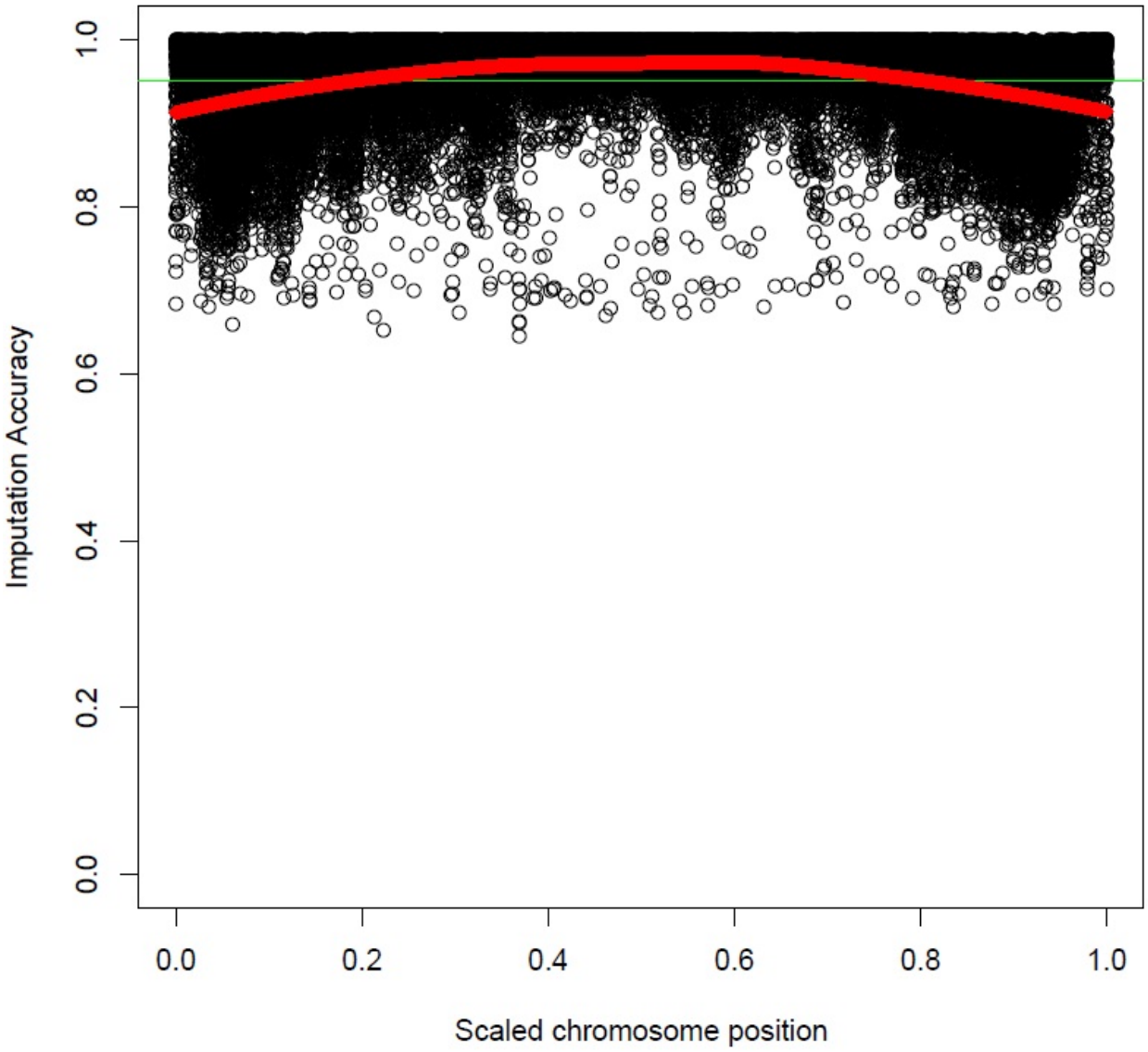

Figure 3.3: SNP-wise imputation accuracy by chromosomal location
SNP-wise imputation accuracy ($IA_{i.}$) vs. the scaled chromosomal location of the SNP. The red line is the weighted mean average estimated using a loess smoother (Cleveland *et al.*, 1992), and the green line represents average imputation accuracy ($IA = 0.951$).

Animal-wise imputation accuracy ($IA_{i.}$) averaged 0.951 but the corresponding highest posterior density interval ($[0.917, 0.978]$) was shorter than that observed for SNP-wise accuracy. Overall, all but 12 animals had imputation accuracy > 0.90 and 551 animals (62%) had imputation accuracy above 0.95. Also, seven of the animals had a dam, sire, or grand-sire in the reference panel (Badke *et al.*, 2012), which resulted in on average higher accuracy of

imputation in these animals (0.959). A group of 15 animals was identified with consistently low imputation accuracy (i.e. < 0.91 for 9K tagSNP) across all sets of tagSNP selected. An ongoing research project in our laboratory investigating breed composition, identified all of the 15 low accuracy individuals as potentially having mixed breed ancestry (YiJian Huang, unpublished data). Further assessing the pedigree of these 15 animals, we found that nine of them were imported to the US, which could result in a slightly different haplotype composition and the observed low accuracy of imputation, when only American Yorkshire pigs had been used as reference. Another three animals of the remaining six US Yorkshires with low imputation accuracy were identified as a family (sire, two offspring), such that the observed low accuracy in the offspring is likely a result of the mixed breed ancestry of their sire.

As noted before, to assess the effect of MAF of imputed SNP on imputation accuracy required adjusting estimates of imputation accuracy for MAF. Imputation accuracy as a function of MAF is presented in Figure 3.4, where imputation accuracy was estimated as a) proportion of alleles correctly imputed, b) coefficient of determination (R^2) between observed and imputed allelic dosage (Hickey *et al.*, 2012; Zheng *et al.*, 2011), and c) proportion of alleles correctly imputed adjusted for MAF (Hayes *et al.*, 2012). The red line in all plots represents the weighted mean average estimated using a loess smoother (Cleveland *et al.*, 1992). Loess consists of fitting smooth piecewise polynomial regressions to local subsets of data and it is widely used in normalization of micro-array experiments (Steibel *et al.*, 2009). At first inspection, it can be seen that accuracy estimated as the proportion of correctly imputed alleles (Figure 3.4a) is highest for low frequency alleles and exhibits a small decrease as MAF increases. However, the observed high proportion of correctly imputed alleles in SNP with low MAF is based on the fact that high frequency alleles can be im-

puted with high accuracy even if imputation is solely based on allele frequency (Hayes *et al.*, 2012; Hickey *et al.*, 2012). For this reason, we computed R^2 and the proportion of correctly imputed alleles adjusted for MAF that provide estimates of imputation accuracy unbiased by allele frequency. In other words, these measures are indicative of the performance of the imputation algorithm in comparison to a baseline imputation based on genotypic frequencies (Hayes *et al.*, 2012; Hickey *et al.*, 2012). When imputation accuracy is adjusted for MAF, estimated accuracy is generally higher for intermediate allele frequencies ($MAF \sim 0.5$) and declines as MAF decreases (Figure 3.4 b/c). Average imputation accuracy considering only the added benefit of the imputation algorithm was lower ($IA_{MAF} = 0.91$, $R^2 = 0.81$) than the total proportion of correctly imputed alleles ($IA = 0.951$). The difference between the proportion of correctly imputed alleles adjusted for MAF (Figure 3.4c) and estimates of R^2 (Figure 3.4b) can be explained by the difference in error counting between these measures. While the proportion of correctly imputed alleles adjusted for MAF is obtained by counting the total number of wrongly imputed alleles, R^2 is obtained from the squared difference in imputed and observed alleles, thereby more heavily penalizing large differences between observed and imputed allelic dosage.

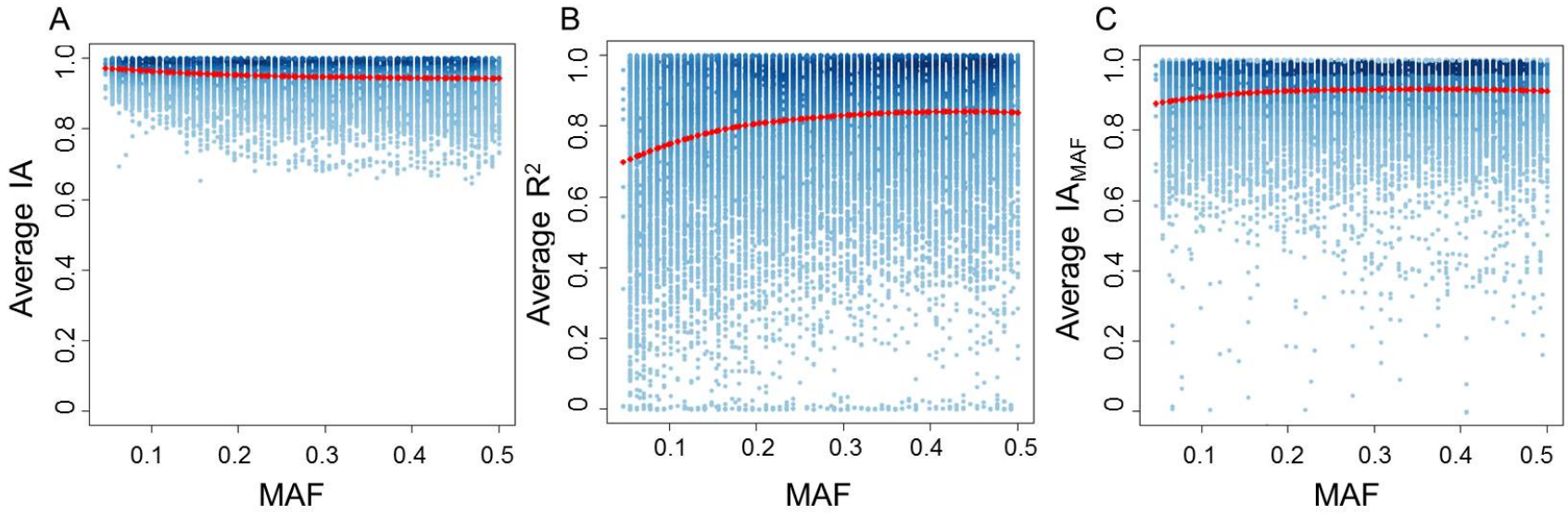

Figure 3.4: Three measures of SNP-wise imputation accuracy by MAF
SNP-wise imputation accuracy computed as **A**) the proportion of correctly imputed alleles ($IA_{i.}$), **B**) the correlation between imputed and observed allelic dosage (R^2), and **C**) the proportion of correctly imputed alleles adjusted for MAF ($IA_{MAF_{i.}}$), as a function of MAF of the SNP. The red line is the weighted mean average estimated using a loess smoother.

3.3.3 Effect of numbers of reference haplotypes on imputation accuracy

For all previous analyses in this paper we imputed genotypes of 889 individuals across all autosomes using a reference panel of 128 Yorkshire haplotypes obtained from a sire/dam/offspring genotyping design (Badke *et al.*, 2012), phased with higher accuracy (Marchini *et al.*, 2006). Reducing the number of imputation animals from 889 to 200 had no impact on the observed imputation accuracy. Imputation accuracy using all 128 haplotypes from the original reference panel was 0.959 on SSC14, which reduced to 0.939 when 64 haplotypes were used, and further to 0.904 when imputation was based on 32 haplotypes (Figure 3.5). Therefore, imputation accuracy larger than 0.90 can be obtained using the commercial 9K tagSNP set with a reference panel of only 32 haplotypes, given that these haplotypes were phased at high accuracy.

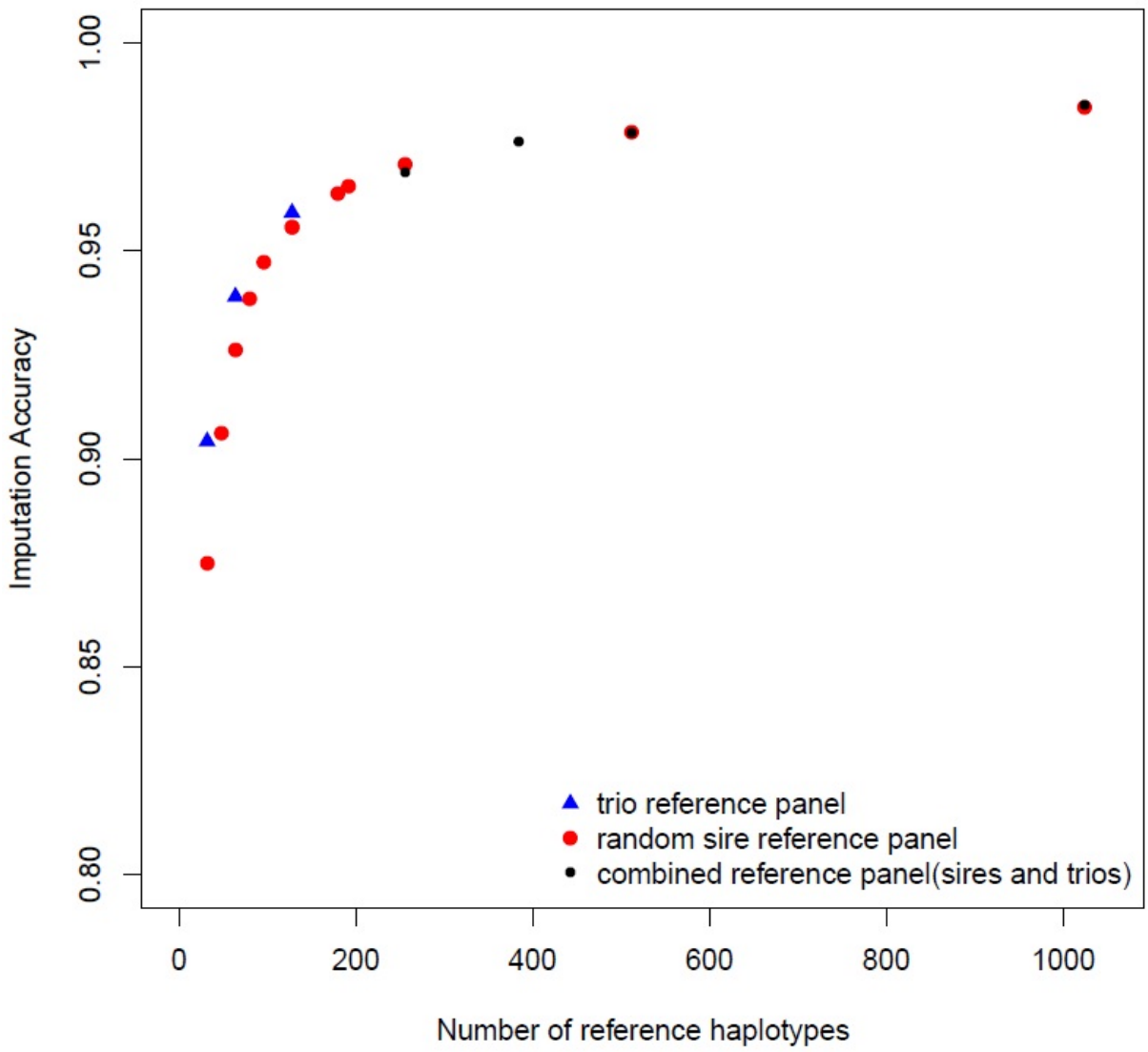

Figure 3.5: Effect of number of reference haplotypes on imputation accuracy
Average imputation accuracy (IA) as a function of the number of haplotypes in the reference panel used for imputation. Imputation accuracy was estimated for reference panels composed of haplotypes from a trio design (blue triangle), reference panels composed of haplotypes from randomly sampled sires (red circle), and reference panels composed of both haplotypes from a trio design and haplotypes from randomly sampled sires (black circle).

We further investigated if it is necessary to obtain reference haplotypes from a trio design, or if accuracy can be replicated using a reference panel of randomly sampled individuals genotyped at high density. In comparison to imputation accuracy obtained using a trio

reference panel, imputation accuracy based on 32 and 64 reference haplotypes derived from selected sires was slightly lower (0.875 and 0.926, respectively). However, accuracy from 128 reference haplotypes obtained from 64 randomly sampled individuals was 0.955, which is practically identical to results obtained using 128 reference haplotypes from trios. Therefore, if the reference panel of haplotypes is composed of more than 128 haplotypes, there is no longer an advantage in using haplotypes obtained from a trio design. Alternatively, the cost of assembling panels of 32, 64, and 128 reference haplotypes obtained from a trio design involves the same genotyping cost as assembling panels of 48, 96, and 192 haplotypes obtained from randomly sampled individuals. This is due to the fact that in a trio design the offspring haplotypes are not used as part of the reference panel, since they are identical to the parents transmitted haplotypes. Imputation accuracies for 48, 96, and 192 reference haplotype panels from randomly sampled individuals were estimated to be 0.906, 0.947, and 0.965, respectively, which is either equivalent or higher than accuracy of imputation obtained using the cost equivalent trio based reference panels (Figure 3.5). In addition, we compared imputation accuracy from reference haplotypes of either 64 randomly selected individuals or the 64 oldest individuals and found no difference in imputation accuracy (0.956, 0.953 respectively). Consequently, if no reference panel of haplotypes is available for a population, according to the results of this study, it would be most cost efficient to assemble high density haplotypes of randomly sampled individuals.

Previous research has indicated an increase in imputation accuracy can be expected as the number of available reference haplotypes increases to a certain point (Howie *et al.*, 2011; Huang *et al.*, 2009). We added randomly selected individuals to the reference panel and obtained 256, 512, and 1024 reference haplotypes. These panels resulted in average accuracy of imputation of 0.971, 0.978, and 0.985 respectively (Figure 3.5). Imputation

accuracy only marginally increased when more than 256 haplotypes were used as reference panel for imputation (up to 1.4% gain). Additionally, we assessed accuracy of imputation from reference panels composed of the original 128 reference haplotypes from a trio design and an increasing number of randomly sampled individuals added to that panel. In this case, imputation accuracy based on reference panels with 256 and 512 haplotypes was 0.969 and 0.978 respectively, which is virtually identical to results obtained using reference panels solely from randomly sampled individuals (Figure 3.5).

In addition to assessing the effect of an increased number of reference haplotypes on average accuracy, we also investigated how it affects individual SNP with different MAF and physical location. We found that as the size of the reference panel increases, imputation accuracy (quantified as R^2) improved more markedly for SNP with MAF below 0.1, such that when the size of the reference panel is increased from 256 haplotypes to 512 haplotypes the increase in accuracy for SNP with MAF below 0.1 was on average 0.06 points, while for all other SNP the increase was only 0.02 points (Additional file 1: Figure 3.6). When imputation was based on 1024 reference haplotypes imputation accuracy appears to be uniform across allele frequencies. Similarly, we observed that imputation accuracy (proportion of correctly imputed alleles) for SNP located in the 10% chromosomal extremes (5% on either side) could be improved through an increase in the number of reference haplotypes (Additional file 2: Figure 3.7). A reference panel containing 512 haplotypes was necessary to obtain maximal imputation accuracy ($IA = 0.99$) for SNP located in the chromosomal center, while SNP in the chromosomal extremes were imputed with accuracy of only 0.97, even when the number of reference haplotypes was doubled (1024 reference haplotypes). Imputation accuracy observed in SNP located in the chromosomal extremes was more than 0.02 accuracy units lower than the average imputation accuracy of all remaining SNP irrespective of the reference panel

size.

3.4 Discussion

3.4.1 Methods for tagSNP selection

Current algorithms for genotype imputation exploit population-wise LD (Browning and Browning, 2009; Scheet and Stephens, 2006), familial LD from identity by descent (Abecasis *et al.*, 2002), or a combination of both (Hickey *et al.*, 2011) to infer unobserved genotypes conditional on tagSNP information. Virtually all methods for tagSNP selection aim at identifying tagSNP that carry the maximum amount of information to impute unobserved markers. This is attained by either directly quantifying the tagSNP ability to predict non-typed SNP (predictive tagSNP selection) or indirectly by selecting tagSNP in high pairwise LD with non-tagSNP (statistical tagSNP selection) (He and Zelikovsky, 2007).

A goal of this study was to select a minimal set of tagSNP that would yield acceptable accuracy of imputation of non-tagSNP (Dassonneville *et al.*, 2012; Weigel *et al.*, 2010a). Since genotype imputation utilizes information about the structure of LD to infer non-observed SNP, we expected that tagSNP sets selected based on LD information, such as statistical and predictive tagSNP selection, would yield higher accuracy of imputation than tagSNP selected based solely on their physical location. In addition, we expected that directly assessing the ability of each tagSNP to predict non-observed SNP (predictive selection) would yield an improvement in imputation accuracy compared to tagSNP selected purely based on pairwise thresholds of LD (statistical selection).

We found that at the lowest examined tagSNP density (1 tagSNP per Mb) accuracy of imputation was below 0.87 irrespective of the method of tagSNP selection and that at least 2

tagSNP per Mb were necessary to increase accuracy to at least 0.91. Accuracy of imputation increased as tagSNP density increased, reaching a plateau accuracy of approximately 0.98 when tagSNP were spaced at an average distance of less than 125kb with negligible increases beyond such density. Our results compare well to those of Weigel *et al.* (2010b), where randomly selected tagSNP at an approximate density of 300kb were necessary to obtain accuracy larger than 0.90 in the US Jersey cattle population using a similar type of imputation. In our study, imputation accuracy of approximately 0.95 was obtained using between 7000 (average tagSNP spacing of 340kb) and 10000 tagSNP (average tagSNP spacing of 230kb), depending on the method of tagSNP selection.

As expected, predictively and statistically selected tagSNP did yield higher accuracy of imputation than evenly spaced tagSNP, but we found no difference in imputation accuracy between tagSNP sets selected statistically or based on predictive ability. Comparing 300 tagSNP selected using predictive ability to 317 tagSNP obtained using statistical selection ($r_t^2 = 0.4$) on SSC18, we observed the same imputation accuracy ($IA = 0.95$). However, the two sets are qualitatively different. For instance, the 300 predictive tagSNP only provide statistical coverage ($r^2 \leq 0.4$) to 37% of non-tagSNP. The tagSNP sets also have on average different MAF ($MAF_{predictive} = 0.30$, $MAF_{statistical} = 0.27$). We attribute the equivalence in imputation accuracy of two different tagSNP sets to the extent of LD observed across the genome in Yorkshire pigs ($r^2 = 0.16$ at 1 Mb, Badke *et al.*, 2012). Under these conditions, precision of estimates of individual tagSNP imputation accuracy is likely compromised by collinearity, making selection of a single best predictive tagSNP at each step of the forward search complicated (Vittinghoff *et al.*, 2005). For example, the initial step of the forward search for predictive tagSNP resulted in six SNP with predictive ability within 0.002 accuracy units of each other. Each of these SNP could have been selected as a starting

point of the greedy search, resulting in different sets of tagSNP selected. Furthermore, the implemented predictive forward search requires $\frac{M_{ti}(2M_i - M_{ti}+1)}{2}$ imputation operations per iteration step compared to only two with statistical selection, where M_i is the number of SNP per chromosome and M_{ti} is the number of selected tagSNP on that chromosome. Consequently, even though both methods result in different tagSNP sets, statistical selection is a computationally efficient proxy for predictive tagSNP selection when moderate LD between consecutive markers is present.

We show that tagSNP sets strictly selected for even spacing are slightly outperformed by statistical or predictive tagSNP selection. However, it is possible to enhance the performance of evenly spaced tagSNP through a few simple measures. TagSNP with high MAF seem to be advantageous for genotype imputation (predictive tagSNP selection seemed to favor tagSNP with high MAF) and their likelihood to segregate across populations will ensure that they carry information for imputation in various populations. This has been exploited previously in cattle for the assembly of the 3K platform (Dassonneville *et al.*, 2012), as well as in newer tagSNP sets aimed to further increase imputation accuracy (Boichard *et al.*, 2012). In addition to selecting evenly spaced tagSNP with high MAF, an increase in accuracy can be obtained by increasing tagSNP density in the chromosomal extremes (Boichard *et al.*, 2012). The success of these enhancements of evenly spaced tagSNP is evident in the imputation accuracy we report using the commercial 9K set ($M_{tagSNP} = 7323$, $IA = 0.951$), which is similar to results we found for statistical tagSNP sets for thresholds $r_t^2 = 0.3$ ($M_{tagSNP} = 7036$, $IA = 0.952$). In addition, although the recently released commercially available chip has approximately 10% fewer tagSNP than the original 9K tagSNP list that was used for this analysis, our conclusions regarding imputation accuracy are likely to uphold, due to the fact that we based our analysis on a set of only 7323 tagSNP, which should be representative

of the number of commercial tagSNP that will pass quality control in future study samples.

In summary, efficient tagSNP selection based on MAF and physical location is feasible and more flexible than statistical tagSNP selection. Selecting evenly spaced tagSNP with high MAF requires knowledge of the physical location of the SNP and the MAF across populations of interest, while statistical tagSNP selection requires knowledge of the LD structure, and would be population specific. As a result, selecting a tagSNP set with high MAF and an increased density in the chromosomal extremes is more versatile than tagSNP sets selected for predictive ability or based on statistical criteria while yielding the same accuracy of imputation. In addition, the tagSNP set selected based on physical location and MAF is expected to be useful for imputation as long as the 60K chip is being used for genomic selection, because we do not expect selection to alter LD or MAF of selected SNP in any particular way. If such tagSNP sets will be used across multiple closely related populations it will be necessary to include a number of SNP that will be specific to a subset of populations. In the case of the 9K tagSNP set more than 9000 tagSNP were selected based on MAF across several populations and physical location of the SNP, but only 7323 of these SNP passed quality editing for the Yorkshire data in this study.

3.4.2 Factors affecting imputation accuracy

Accuracy of imputation is affected by several factors including the selection and density of tagSNP as detailed above, the MAF and the physical location of the imputed SNP, as well as the size and composition of the reference panel.

When evaluating imputation accuracy as a function of the tagSNP selection method and density we have focused on average accuracy as a measure of overall performance. Assessing the average accuracy of imputation is a good indicator of the performance of imputed

genotypes, when all genotypes are used simultaneously to obtain a global measure. Such a measure could be prediction of GEBV, which would be based on all SNP simultaneously, such that a small number of wrongly imputed SNP is unlikely to greatly affect the accuracy of prediction. Alternatively, some applications of imputed high density genotypes may require high accuracy across all SNP. One example would be GWAS based on imputed genotypes. For GWAS, SNP associations are assessed on a SNP by SNP basis, such that wrongly imputed alleles for low frequency SNP are more likely to cause bias in the estimated association, especially since phenotypes of interest are suspected to be associated with low frequency alleles (Howie *et al.*, 2011).

One of the factors directly related to the individual SNP imputation accuracy, is the allele frequency of that particular SNP. To investigate imputation accuracy as a function of MAF we used two measures of imputation accuracy that were unbiased by MAF (i.e. IA_{MAF}, R^2). The adjusted proportion of correctly imputed alleles (IA_{MAF}) and the correlation between observed and imputed allelic dosage (R^2) are scaled differently, such that the observed accuracy differs as a function of scale, but the comparative difference in imputation accuracy as a function of MAF can be observed using either of the two accuracy measures. We found that estimates of imputation accuracy adjusted for MAF (R^2, IA_{MAF}) are lower ($R^2 = 0.73$, $IA_{MAF} = 0.89$) for SNP with MAF below 0.1, compared to SNP with MAF above 0.1 ($R^2 = 0.82$, $IA_{MAF} = 0.91$), which has been previously noted by Hayes *et al.* (2012) reporting results of genotype imputation in sheep and Hickey *et al.* (2012) in lines of maize.

Another factor relating to individual SNP imputation accuracy is the physical location of the SNP. Previous studies designing low density genotyping platforms have pointed out the need to increase coverage of tagSNP in the chromosomal extremes due to difficulties

in correctly imputing SNP located in those regions (Boichard *et al.*, 2012; Dassonneville *et al.*, 2012). In the commercial 9K tagSNP set the density of tagSNP within 5Mbp of the chromosomal extremes was approximately doubled to aid imputation accuracy. We found that imputation accuracy using the 9K commercial tagSNP was still slightly lower in the extreme regions (0.949) when compared to the chromosome center (0.972). The effect however was alleviated in comparison to an equally spaced tagSNP set of comparable density, where the average imputation accuracy in the chromosomal extremes was only 0.89.

We found a group of 15 animals that produced consistently low imputation accuracy ($IA \leq 0.90$), compared to all remaining animals ($IA = 0.951$). Nine of these animals were identified as imports, such that the observed low accuracy of imputation is likely a result of differences in haplotype frequencies between the US Yorkshire population that was used as reference for imputation, and the population(s) from which these animals originated. The remaining six animals were all identified as having potentially mixed breed ancestry based on results of a concurrent research project in our laboratory (YiJian Huang, unpublished data). In addition, we can infer from these results that if a population contains heterogeneous sub-populations, such as a large number of imported animals or animals with cross-bred ancestry, imputation accuracy will be decreased if this sub-structure is not accounted for when sampling reference haplotypes.

We found that increasing the number of reference haplotypes led to an increase in average imputation accuracy. In addition to the number of reference haplotypes, their average relatedness to the imputation candidates (Gualdron Duarte *et al.*, 2012; Hayes *et al.*, 2012; Hickey *et al.*, 2012; Huang *et al.*, 2012), as well as accurate phasing of these haplotypes (Browning and Browning, 2011) directly affect the resulting accuracy of imputation. In this paper, we assessed the effect of phasing accuracy and the number of reference haplotypes.

Previous research comparing phasing accuracy of unrelated or randomly sampled individuals and trio designs (sire/dam/offspring), found that genotypes from trios can be phased with higher accuracy (Marchini *et al.*, 2006). The initial reference panel available in this study was composed of the haplotypes of sire/dam pairs from a previous sample of trios that were unrelated for at least two generations and therefore sampled to efficiently cover the Yorkshire population (Badke *et al.*, 2012). We found that for haplotype panels composed of 64 or less haplotypes, imputation accuracy was higher when these haplotypes were obtained from the trio design rather than a random sample of individuals (Figure 3.5). This advantage of the trio design is likely due to the superior phasing accuracy as well as the sampling strategy used to obtain these samples. However, adjusting sample size for the increased genotyping cost in a trio design, we observed that imputation accuracy was equal or higher when imputation was based on haplotypes obtained from randomly sampled individuals instead of trio reference haplotypes (Figure 3.5). Therefore, we conclude that if no reference panel is available in a population the most cost efficient method for reference panel construction is genotyping a random sample of individuals across the population.

Next we assessed imputation accuracy as a function of increasing reference panel size. We found that a reference panel of 256 to 512 reference haplotypes is sufficient to obtain imputation accuracy of $IA = 0.97$. If the size of the reference panel is increased beyond 1024 haplotypes ($IA = 0.985$) any further gain in imputation accuracy appears to be very small. A similar type of response has been observed in human genotype imputation (Huang *et al.*, 2009). This relatively small number of reference haplotypes necessary to obtain high imputation accuracy ($IA = 0.97$) is likely due to the relatively small effective population size of the Yorkshire population ($N_e = 113$, (Welsh *et al.*, 2010)), and consequently high average LD even at decreased tagSNP density.

After determining that increasing the size of the reference panel would increase accuracy of imputation, we assessed whether accuracy would differ as a function of reference panel composition. In general, when imputation experiments are conducted the older animals are used as reference panel, while the younger animals serve as imputation candidates (Weigel *et al.*, 2010b; Zhang and Druet, 2010). We assessed whether imputation accuracy would differ depending on the reference panel being composed of randomly selected individuals or older individuals and found no advantage in imputation accuracy when selecting a reference panel composed of older animals.

In addition to the observed increase in overall imputation accuracy, we found that increasing the size of the reference panel is especially efficient at increasing the individual imputation accuracy of SNP that exhibited below average imputation accuracy (Howie *et al.*, 2011). SNP with MAF below 0.1 were imputed poorly in comparison to SNP with MAF above 0.1 (accuracy measure R^2, IA_{MAF}), but as reference panel size increased imputation accuracy of these SNP improved, and for imputation based on 1024 haplotypes we observed a uniform distribution of imputation accuracy (quantified as R^2) across levels of MAF (Additional file 1: Figure 3.6). An increase in the size of the reference panel increases the precision of estimated frequencies of haplotypes containing rare alleles, which appears to more efficiently boost imputation accuracy for the corresponding SNP (Howie *et al.*, 2011). SNP located in the 10% chromosomal extremes (5% on either side) also had on average lower imputation accuracy than the remaining SNP. As reference panel size was increased very little improvement could be observed in imputation accuracy of SNP located in the center of the chromosome, due to these SNP already being imputed with accuracy close to 1. However, imputation accuracy of SNP in the chromosome ends improved as reference haplotypes were added to the panel, until reaching accuracy within 0.02 points of the average

imputation accuracy of SNP in the remainder of the chromosome for imputation based on a reference panel containing 1024 haplotypes (Additional file 2: Figure 3.7).

Although, we did not assess the accuracy of GEBV prediction based on imputed genotypes in this paper, we can use results from dairy cattle breeding that show the promise of imputed genotypes to predict GEBV. Based on the average imputation accuracy we observed for Yorkshire pigs and previous results for GEBV prediction based on imputed genotypes in dairy cattle we could expect that losses in accuracy of GEBV prediction as a result of genotype imputation will be negligible. Wiggans *et al.* (2012) and Dassonneville *et al.* (2011) reported correlation of GEBV from imputed genotypes ($IA \geq 0.96$) with GEBV estimated from high density genotypes larger than 0.93. Moreover, Weigel *et al.* (2010a), reported a loss in accuracy of GEBV, estimated as the correlation between GEBV and direct genomic value, between 0 and 5% when using genotypes imputed with low accuracy ($IA = 0.91$). Since our estimates of imputation accuracy in the Yorkshire population are within the range of those reported in dairy cattle (Dassonneville *et al.*, 2011; Weigel *et al.*, 2010a,b; Wiggans *et al.*, 2012), we expect GEBV estimated from imputed genotypes in Yorkshire pigs to be as accurate as those currently used in the dairy breeding industry. Furthermore, these results are expected to hold in other swine breeds with similar levels of LD (Badke *et al.*, 2012).

3.5 Conclusion

In conclusion, high ($IA \geq 0.95$) genotype imputation accuracy can be achieved in pigs combining the newly available commercial 9K tagSNP set and a relatively small reference haplotype panel (128 haplotypes), even when imputation is based only on population-wide LD. Further improvements in imputation accuracy could be achieved through the inclusion

of additional reference animals ($IA = 0.97$ with 512 reference haplotypes) and the use of pedigree relations between reference and imputation animals in the imputation algorithm (Gualdron Duarte *et al.*, 2012; Huang *et al.*, 2012; Wiggans *et al.*, 2012). An important result from this study is that an efficient design for reference panel construction is randomly sampling individuals instead of specifically sampling older animals or trios. In addition, a relatively small panel of reference haplotypes ($\geq$ 128) can efficiently serve as a reference panel for genotype imputation, such that any available high density genotypes in a livestock population could potentially serve this purpose. For the pig species such panels are already available for several populations (Badke *et al.*, 2012). Finally, prospects for the use of imputed genotypes in GEBV prediction are very positive based on the results from dairy breeding that routinely use similarly accurately imputed genotypes for genomic evaluation (Dassonneville *et al.*, 2011; Weigel *et al.*, 2010a; Wiggans *et al.*, 2012). The methodology used in this paper for construction of tagSNP sets and reference haplotype panels can be easily applied in any future study population. Code and data to obtain and reproduce the results presented is publicly available at `https://www.msu.edu/~steibelj/JP_files/imputation.html`.

3.6 Supplementary Materials

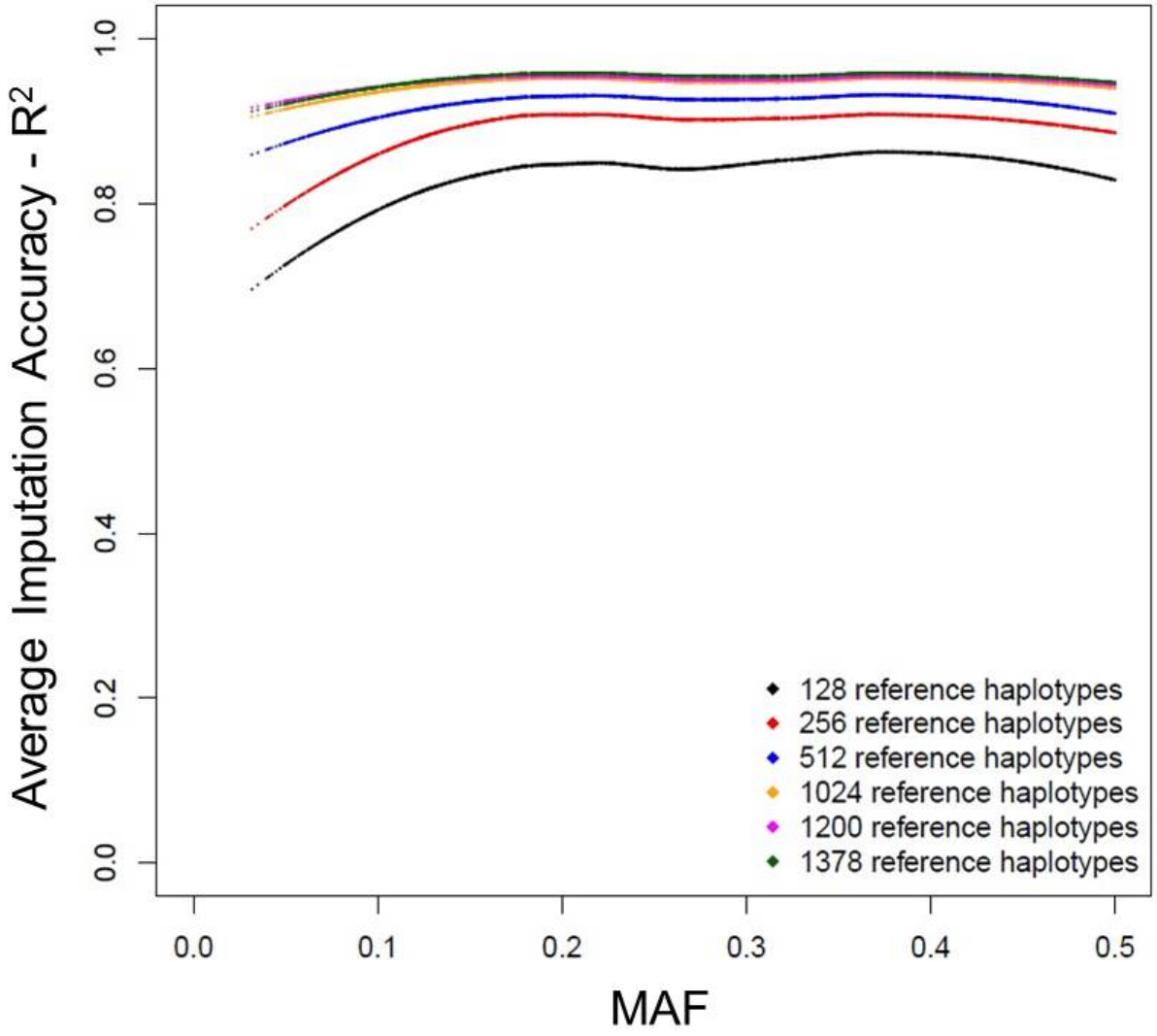

Figure 3.6: Effect of reference panel size on imputation accuracy of SNP as a function of their MAF

Weighted mean average imputation accuracy (quantified as R^2) as a function of MAF depicted for imputation based on haplotype reference panels of increasing size.

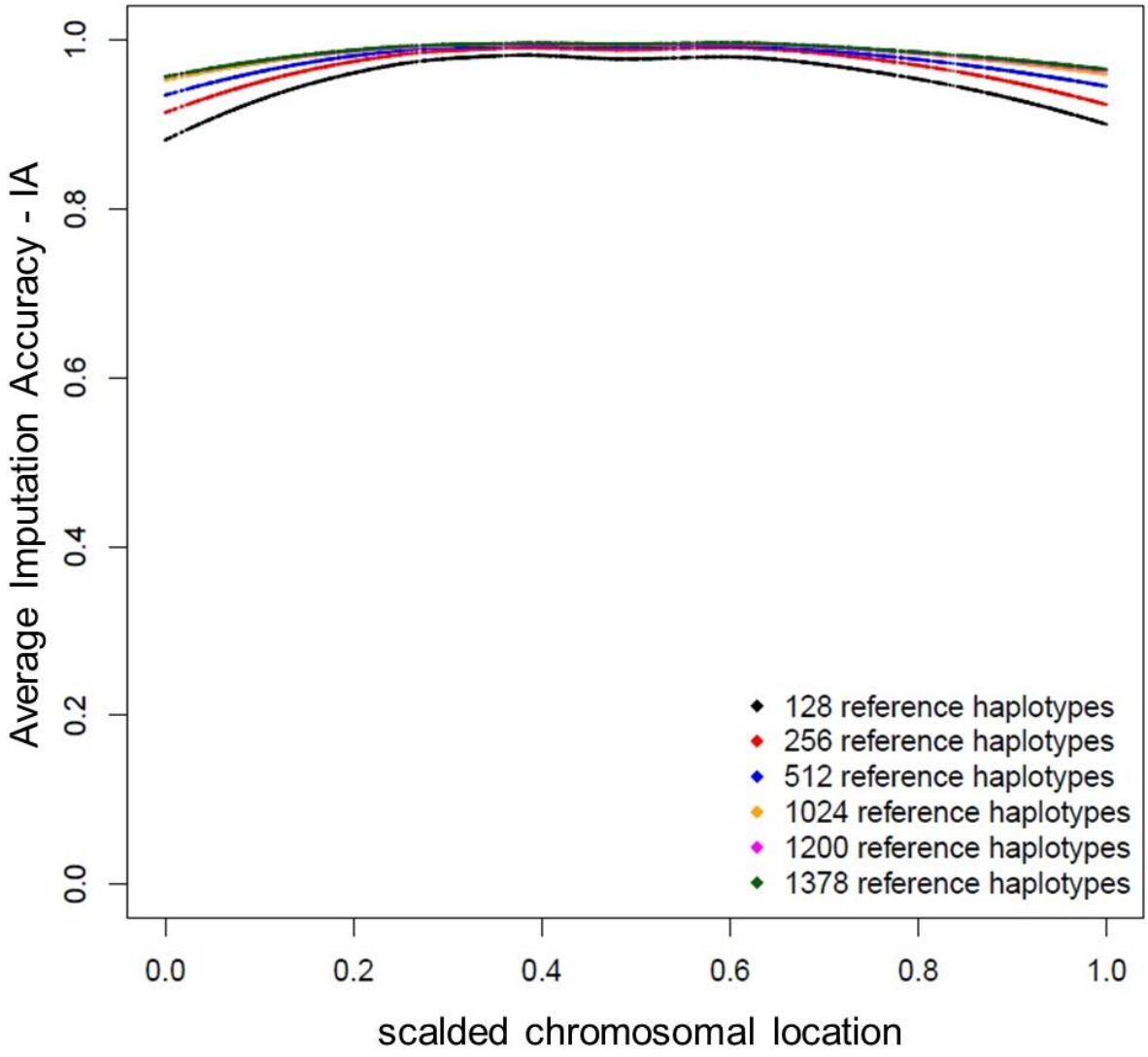

Figure 3.7: Effect of reference panel size on imputation accuracy of SNP as a function of scaled physical location

Weighted mean average imputation accuracy (quantified as IA) as a function of the scaled chromosomal location for imputation based on haplotype reference panels of increasing size.

Chapter 4

Accuracy of estimation of genomic breeding values in pigs using low density genotypes and imputation

Badke, Y. M., Bates, R. O., Ernst, C. W., Fix, J., & Steibel, J. P. (2013). Accuracy of estimation of genomic breeding values in pigs using low density genotypes and imputation. submitted to G3

4.1 Introduction

Genetic improvement through breeding for lean growth, reproductive performance, meat quality, and health traits is an important tool in the pig breeding industry to assure its continued competitiveness and success. The use of traditional estimated breeding values (EBV) derived from pedigree information has lead to important advances in genetic improvement but has several limitations (Dekkers *et al.*, 2010). A number of important phenotypes, such as disease resistance or lifetime productivity traits are difficult and expensive to observe or can only be measured later in life, which impairs the estimation of highly accurate EBV. Furthermore, EBV estimated from performance records of close relatives are based on the expected instead of observed proportions of identity by descent between the animals.

The use of genomic breeding values (GEBV), estimated using a large number of genetic markers across the genome is expected to overcome a number of these limitations (Meuwissen *et al.*, 2001; Dekkers *et al.*, 2010). Genomic prediction is expected to improve the accuracy of breeding values, especially for lowly heritable and complex traits and allow for the selection of animals at a young age thereby shortening generation intervals (Hayes *et al.*, 2009a; VanRaden *et al.*, 2009; Wiggans *et al.*, 2011). A number of review papers have reported the progress and success of genomic selection in dairy cattle (Hayes *et al.*, 2009a; VanRaden *et al.*, 2009; Wiggans *et al.*, 2011), and it is expected to be equally useful in pigs (Tribout *et al.*, 2012). High density genotypes in pigs can be obtained from the PorcineSNP60 BeadChip (Illumina, San Diego, CA) containing roughly 62K SNP (Ramos *et al.*, 2009). Previous research based on this chip has reported on extent of linkage disequilibrium (LD) (Uimari and Tapio, 2011; Badke *et al.*, 2012), effective population size (Uimari and Tapio, 2011), the correlation of phase between populations (Badke *et al.*, 2012), and genome wide association

studies (GWAS) performed for various traits e.g. boar taint (Duijvesteijn *et al.*, 2010).

First implementations of genomic prediction in pigs included evaluations for total number of pigs born in a litter and percent stillborn (Cleveland *et al.*, 2010). Results of this study indicated that GEBV in pigs can reach accuracies comparable to those observed in dairy cattle if the training population is large enough (Cleveland *et al.*, 2010). In addition, several strategies to increase cost efficiency through the use of low density genotypes have been explored but accuracy of GEBV was reasonable only for certain traits, likely due to differences in the genetic architecture of the traits (Cleveland *et al.*, 2010). However, when genotypes were imputed with high accuracy results for genomic evaluation were promising for several traits in a commercial pig population (Cleveland and Hickey, 2013).

The relatively high genotyping cost per animal currently limits the widespread commercial use of high density genotypes for genomic selection purposes in pigs. One strategy to improve the cost efficiency of genotyping schemes is the use of genotype imputation for a portion of the population. In the interest of cost efficiency it is likely that selection candidates will not be genotyped using a high density array such as the PorcineSNP60, but rather will be genotyped on a low density array like the recently released GeneSeek Genomic Profiler for Porcine LD (GGP-Porcine: GeneSeek Inc., a Neogen Co., Lincoln, NE), a subset of the PorcineSNP60 containing roughly 10K SNP. We showed (Badke *et al.*, 2013) that genotypes in pigs can be imputed from the GGP-Porcine to the PorcineSNP60 with accuracy of $R^2 = 0.88$ using LD based imputation algorithms with a reference panel of haplotypes as small as 128 haplotypes. Imputation accuracy can be further improved by adding animals to the reference panel (Badke *et al.*, 2013), or in case of a pedigreed population exploiting Mendelian segregation and population wide LD (Huang *et al.*, 2012; Gualdrón Duarte *et al.*, 2013). We use genotypes imputed based on population wide LD, offering a strategy that

can be applied universally in any population, for which a suitable reference panel can be assembled.

Our objective was to assess how using imputed instead of observed genotypes would affect the accuracy of genomic evaluations using an efficient G-BLUP fitting the prediction equation to the realized genomic relationship matrix (Hayes *et al.*, 2009b). We used two sets of reference haplotype panels, small ($N = 128$) or a large ($N \sim 1800$), to evaluate how an increase in imputation accuracy affects the accuracy of genomic predictions.

4.2 Materials & Methods

4.2.1 Materials

4.2.1.1 Animals and Genotypes

Data used in this study was collected from 983 Yorkshire sires. High density genotypes for these animals were obtained from samples provided by the National Swine Registry (NSR). Genotyping was performed at a commercial laboratory (GeneSeek, a Neogen Company, Lincoln, NE) using the Illumina PorcineSNP60 BeadChip. The same dataset was previously used to assess the effect of genotype imputation (Badke *et al.*, 2013) and is publicly available at:

`https://www.msu.edu/~steibelj/JP_files/imputation.html`. Animal protocols were approved by the Michigan State University All University Committee on Animal Use and Care (AUF# 03/09-046-00). Genotyping rate of at least 90% of both animals and SNP and a minor allele frequency of at least 5% were required for genotypes to be included in the analysis, leaving a total of 41248 markers in 983 animals. SNP that were not assigned to an

autosomal position in map build 10.2 were excluded from the analysis. It was our goal to estimate the genomic breeding value (GEBV) of male offspring of a sire and since sires will not pass an X chromosome to their male offspring, these SNP do not contribute to the sons GEBV (VanRaden *et al.*, 2009). In addition to genotypes for 983 Yorkshire sires, a set of 128 Yorkshire haplotypes was available as a reference panel for genotype imputation from a previous study (Badke *et al.*, 2012). These haplotypes are also freely available at `https://www.msu.edu/~steibelj/JP_files/LD_estimate.html` and details on the design and phasing can be found in Badke *et al.* (2012).

4.2.1.2 Phenotypes

For every animal and their parents, estimated breeding values (EBV) and accuracies were obtained for three traits from NSR through their traditional genetic evaluation. These traits were: backfat thickness (BF), number of days to 250lb (D250), and loin muscle area (LEA). Descriptive statistics of EBV and accuracies are presented in Table 4.1. All code and data used in this paper has been assembled into an R package, accessible at: `http://tinyurl.com/MSURGEBV`.

4.2.2 Methods

4.2.2.1 De-regression of breeding values

De-regressed breeding values (dEBV) were used as response variables throughout the analysis. We computed individual animal dEBV and their weights (w_i) with the parent average removed following the procedure outlined by Garrick *et al.* (2009). After de-regression and filtering a total of 965, 936, and 938 animals remained for the traits BF, D250, and LEA

respectively (Table 4.1).

Table 4.1: Descriptive statistics of EBV

	BF	D250	LEA
$E\bar{B}V$	-0.03	4.57	0.61
$\bar{r}^2_{EBV}$ $^{\times}$	0.74	0.67	0.75
N *	965	936	938
h^2	0.45	0.26	0.47

$^{\times}$ average reliability of EBV
* number of animals with usable EBV

4.2.2.2 Estimation of genomic relationship matrix

The genomic relationship matrix was estimated from observed or imputed high density ($\sim 41K$) SNP genotypes. Genotypes were expressed as allelic dosage, which is the number of copies of the minor allele, such that genotypes were entered into a marker matrix **M** as a decimal number in the interval $[0, 2]$. We obtained matrix **Z** by subtracting twice the allelic frequency of the minor allele (p_i), from columns of **M** (VanRaden, 2008). The genomic relationship matrix was then calculated as:

$$\mathbf{G} = \frac{\mathbf{ZZ'}}{2\sum_{i=1}^{M} p_i(1-p_i)} \tag{4.1}$$

where $2\sum_{i=1}^{M} p_i(1-p_i)$ is a normalizing constant (Wang *et al.*, 2012) summing expected variances across markers scaling **G** towards the numerator relationship matrix (VanRaden, 2008). The allele frequency p_i was obtained using all available animals (N=983). Average relatedness between animals was obtained from the row/column vectors of **G**. We quantified relatedness in this study as the average of the top 10 relationships observed within the **G**

matrix ($rel10$).

4.2.2.3 Implementation of prediction model

Using the genomic relationship matrix from equation (4.1) an animal-centric model for genomic evaluations can be written as:

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{a} + \mathbf{e} \tag{4.2}$$

where $\mathbf{y}$ is the vector of dEBV, μ is the overall mean, $\mathbf{a}$ is the vector of n animal effects ($\mathbf{a} \sim N(\mathbf{0}, \mathbf{G}\sigma^2_a)$), and $\mathbf{e}$ is a vector of random residuals ($\mathbf{e} \sim N(\mathbf{0}, \mathbf{R}\sigma^2_e)$). The variance of the dEBV is $var(\mathbf{y}) = \mathbf{G}\sigma^2_a + \mathbf{R}\sigma^2_e$, where $\mathbf{R}$ is a diagonal matrix with diagonal elements $R_{ii} = \frac{1}{w_i}$, the inverse of the weights of the dEBV (VanRaden *et al.*, 2011). Equivalently, the information in $\mathbf{G}$ can also be included in the incidence matrix of the animal effects $\mathbf{a}$ as follows (Vazquez *et al.*, 2010):

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{C}\mathbf{a}^* + \mathbf{e} \tag{4.3}$$

where $\mathbf{C}$ is the Cholesky decomposition of $\mathbf{G}$, such that $\mathbf{G} = \mathbf{C}\mathbf{C}'$, μ is the overall mean, $\mathbf{a}^*$ is the vector of animal effects with $\mathbf{a}^* \sim N(\mathbf{0}, \mathbf{I}\sigma^2_{a^*})$ noticing that $\mathbf{a} = \mathbf{C}\mathbf{a}^*$, and $\mathbf{e}$ is a vector of residual effects $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R}\sigma^2_e)$ such that $var(\mathbf{y}) = \mathbf{C}\mathbf{C}'\sigma^2_{a^*} + \mathbf{R}\sigma^2_e = \mathbf{G}\sigma^2_{a^*} + \mathbf{R}\sigma^2_e$. The variance terms for models (4.2) and (4.3) are equal, such that the two models are in fact equivalent if variance components are assumed known. Likewise, when estimating the parameters under these two models we found virtually identical results, but model (4.3) was computationally more efficient resulting in a two fold reduction in compute time (results not shown). The `BLR` package (Pérez *et al.*, 2010) in R (Team, 2011) was used to fit the mixed model equations. Model parameters σ^2_e and σ^2_{a*} were sampled from their corresponding full

conditional distribution using a Gibbs sampler. Prior distributions were elicited based on equations presented by Pérez *et al.* (2010). The prior distribution of σ^2_e and σ^2_{a*} were an inverse χ^2 distribution with degrees of freedom *df* and scale S. To ensure proper priors with finite expectations we set $df = 3$. The scale parameters were obtained as a function of the *df* and assuming values of the genetic variance (V_a) and error variance (V_e) (Pérez *et al.*, 2010):

$$\sigma^2_e \sim \chi^{-2}(df_e = 3, S_e = V_e(df_e + 2))$$

$$\sigma^2_{a*} \sim \chi^{-2}(df_a = 3, S_a = \frac{V_a(df_a + 2)}{\bar{A_{ii}}})$$

where $\bar{A_{ii}}$, is the average inbreeding coefficient, set equal to 1 in this case, assuming no inbreeding. Heritability was assumed to be $h^2 = 0.5$, such that after the value for V_e was arbitrarily set to 0.4, V_a was estimated $V_a = \frac{V_e h^2}{1-h^2}$. The Gibbs sampler implemented in `BLR` (Pérez *et al.*, 2010) was used to obtain a total of 100,000 samples, 10,000 of which were discarded as burn-in. The reported estimates of σ^2_e, σ^2_{a*}, animal effects ($\mathbf{a}^*$), and GEBV ($\hat{\mathbf{y}}$) were based on the posterior means of the remaining 90,000 iterations. We assessed convergence of the MCMC chain as well as sensitivity to priors to ensure robustness of estimates to priors (results not shown).

4.2.3 Genomic prediction under cross-validation

Accuracy of genomic evaluation was estimated in a 10 fold cross-validation design. Approximately 10 % of the animals were randomly assigned to a validation panel (V) in which predictions would be made, while the remaining 90% were used as the training panel (T) to estimate the parameters necessary for prediction. A total of 10 separate datasets were created such that each animal would be used for validation once. Across cross-validation

datasets we fit model (4.3) to the training animals, we refer to that subset by adding a subindex T:

$$\mathbf{y}_T = \mathbf{1}n_T\mu + \mathbf{C}_T\mathbf{a}_T^* + \mathbf{e}_T \tag{4.4}$$

to estimate the BLUP of $\hat{\mathbf{a}}_T^*$ (VanRaden *et al.*, 2011):

$$\hat{\mathbf{a}}_T^* = \mathbf{C}_T'(\mathbf{G}_T + \mathbf{R}_T\frac{\sigma_e^2}{\sigma_a^2})^{-1}(\mathbf{y}_T - \mathbf{1}n_T\hat{\mu}) \tag{4.5}$$

where the matrices $\mathbf{G}$ and $\mathbf{C}$ are partitioned into block structure such that

$$\begin{bmatrix} \mathbf{G}_T & \mathbf{G}_{TV}' \\ \mathbf{G}_{TV} & \mathbf{G}_V \end{bmatrix} = \begin{bmatrix} \mathbf{C}_T & 0 \\ \mathbf{C}_{TV} & \mathbf{C}_V \end{bmatrix} \begin{bmatrix} \mathbf{C}_T' & \mathbf{C}_{TV}' \\ 0 & \mathbf{C}_V' \end{bmatrix} = \begin{bmatrix} \mathbf{C}_T\mathbf{C}_T' & \mathbf{C}_T\mathbf{C}_{TV}' \\ \mathbf{C}_{TV}\mathbf{C}_T' & \mathbf{C}_{TV}\mathbf{C}_{TV}' + \mathbf{C}_V\mathbf{C}_V' \end{bmatrix} \tag{4.6}$$

The relation between the BLUP for $\mathbf{a}$ based on model (4.2) and $\hat{\mathbf{a}}^*$ based on model (4.3) can be expressed as:

$$\begin{bmatrix} \mathbf{a}_T \\ \mathbf{a}_V \end{bmatrix} = \begin{bmatrix} \mathbf{C}_T & 0 \\ \mathbf{C}_{TV} & \mathbf{C}_V \end{bmatrix} \begin{bmatrix} \mathbf{a}_T^* \\ \mathbf{a}_V^* \end{bmatrix} \tag{4.7}$$

The genomic breeding value of training animals in model (4.2) were computed as:

$$\hat{\mathbf{a}}_T = \mathbf{C}_T\hat{\mathbf{a}}_T^* = \mathbf{C}_T\mathbf{C}_T'(\mathbf{G}_T+\mathbf{R}_T\frac{\sigma_e^2}{\sigma_a^2})^{-1}(\mathbf{y}_T-\mathbf{1}n_T\hat{\mu}) = \mathbf{G}_T(\mathbf{G}_T+\mathbf{R}_T\frac{\sigma_e^2}{\sigma_a^2})^{-1}(\mathbf{y}_T-\mathbf{1}n_T\hat{\mu})$$

Subsequently, the genomic breeding values of the validation animals $\hat{\mathbf{a}}_V$ were estimated from $\hat{\mathbf{a}}_T$ using the following equation:

$$\hat{\mathbf{a}}_V = \mathbf{G}_{TV}\mathbf{G}_T^{-1}\hat{\mathbf{a}}_T = \mathbf{C}_{TV}\mathbf{C}_T'(\mathbf{G}_T + \mathbf{R}_T\frac{\sigma_e^2}{\sigma_a^2})^{-1}(\mathbf{y}_T - \mathbf{1}n_T\hat{\mu}) \tag{4.8}$$

where σ^2_e, σ^2_a, and $\hat{\mu}$ are estimated using model (4.4) which is equivalent to applying model (4.3) to the training animals.

4.2.3.1 Estimation of accuracy

Accuracy of genomic evaluation is the correlation between the estimated GEBV and the unknown true breeding values (TBV) (Hayes *et al.*, 2009a). However, the TBV are unknown. Consequently, the accuracy of genomic evaluation has to be approximated using the available information. Hayes *et al.* (2009a) proposed to express the correlation between GEBV and TBV as a function of the correlation between GEBV and EBV:

$$r_{(GEBV,TBV)} = \frac{cor(GEBV, EBV)}{cor(EBV, TBV)} = \frac{cor(GEBV, EBV)}{\sqrt{r^2_{EBV}}} \tag{4.9}$$

where r^2_{EBV} is the estimated reliability of the EBV. VanRaden *et al.* (2009) replaced r^2_{EBV} with the arithmetic mean of the reliability of the EBV. Daetwyler *et al.* (2013) proposed to report a simple Pearson correlation coefficient between GEBV and EBV to allow for comparability of results across studies. We estimate accuracy of genomic evaluation as the Pearson correlation coefficient between GEBV and EBV ($r_{(GEBV,EBV)}$) and the Pearson correlation coefficient adjusted for the average accuracy of the EBV to facilitate such comparison ($\frac{r_{(GEBV,EBV)}}{\bar{r}_{EBV}}$).

Accuracies of individual GEBV were obtained analogous to the accuracy of EBV in an animal model (Goddard *et al.*, 2011) through inversion of the mixed model equations (Mrode, 2005; VanRaden, 2008; VanRaden *et al.*, 2009; Strandén and Garrick, 2009; Clark *et al.*, 2012). The accuracy of $\hat{\mathbf{a}}$ of the model (4.2) can be expressed as (Mrode, 2005;

Strandén and Garrick, 2009; Clark *et al.*, 2012):

$$r_{\hat{a}} = \sqrt{1 - (PEV/\sigma_a^2)} \tag{4.10}$$

where PEV is the prediction error variance of $\hat{\mathbf{a}}$:

$$PEV = var(\mathbf{a} - \hat{\mathbf{a}}) = (\mathbf{R}^{-1}\frac{1}{\sigma_e^2} + \mathbf{G}^{-1}\frac{1}{\sigma_a^2})^{-1} \tag{4.11}$$

and σ_a^2 is the genetic variance such that:

$$var(\mathbf{a}) = \mathbf{G}\sigma_a^2 \tag{4.12}$$

Strandén and Garrick (2009) showed, that $r_{\hat{a}}$ for all animals can be obtained from the diagonals of:

$$r_{\hat{a}} = \sqrt{\frac{\left\{\mathbf{G}(\mathbf{G} + \mathbf{R}\frac{\sigma_e^2}{\sigma_a^2})^{-1}\mathbf{G}\right\}_{ii}}{\{\mathbf{G}\}_{ii}}} \tag{4.13}$$

and VanRaden (2008) showed that the accuracy of GEBV of the validation animals can be obtained from the diagonals of:

$$r_{\hat{\mathbf{a}}_V} = \sqrt{\frac{\left\{\mathbf{G}_{TV}(\mathbf{G}_T + \mathbf{R}_T\frac{\sigma_e^2}{\sigma_a^2})^{-1}\mathbf{G}'_{TV}\right\}_{ii}}{\left\{\mathbf{G}_V\right\}_{ii}}} \tag{4.14}$$

This equation was used to estimate the accuracy of individual GEBV for validation animals.

4.2.3.2 Genotype imputation

Linkage disequilibrium (LD) based genotype imputation was performed with BEAGLE version 3.3.1 (Browning and Browning, 2009). We used the standard settings for BEAGLE: ten iterations of the phasing algorithm, drawing four samples per iteration. Previous results from our group (Badke *et al.*, 2013) and other studies (Hayes *et al.*, 2012) showed negligible improvement in imputation accuracy as a result of an increase in iterations or samples per iteration.

A matrix of 'observed' genotypes was created by imputing randomly missing genotypes in the 983 Yorkshire sires ($\leq$ 0.05%) supplementing the data with a reference panel of 128 Yorkshire haplotypes to improve imputation accuracy. Due to the small ($\leq$ 0.05%) percentage of randomly missing genotypes we expected accuracy of imputation very close to 100%, and consequently treated these genotypes as 'observed' genotypes for all further analysis.

We implemented two separate imputation experiments, which differ in the size of the high density reference panel used for imputation: 1) a reference panel of 128 Yorkshire haplotypes or 2) a reference panel combining the 128 Yorkshire haplotypes with the haplotypes of all animals that are part of the training panel ($\sim$ 1700 additional haplotypes) in the respective cross-validation dataset. To assess the effect of genotype imputation on genomic prediction we considered the following four scenarios: 1) the reference scenario where genomic evaluation was based on observed genotypes in training and validation animals, 2) genomic evaluation based on observed genotypes in the training animals and genotypes imputed from a large reference panel ($\sim$ 1800 haplotypes) in the validation animals, 3) genomic evaluation based on observed genotypes in the training animals and genotypes imputed from a small

reference panel (128 haplotypes) in the validation animals, and 4) genomic evaluation based on imputed genotypes in training and validation animals using a small (128 haplotypes) but representative reference panel for imputation. All genotype imputation and subsequent estimation of imputation accuracy was implemented using the R package `impute.R` (Badke *et al.*, 2013). To compare average accuracy of genomic evaluation across these four scenarios we fitted a linear model with the average accuracy of genomic evaluation as response variable and the genotype imputation scenario as independent variable, adding the effect of the random cross-validation dataset in which accuracy of genomic evaluation was estimated as a random blocking factor.

4.3 Results

4.3.1 Accuracy of genomic evaluation and GEBV using observed genotypes

When genotypes were observed in both training and prediction animals, accuracy of genomic evaluation, measured as the weighted mean of the Pearson correlation coefficient between EBV and predicted GEBV across 10 cross-validation datasets, was 0.68, 0.66, and 0.65 for BF, D250, and LEA respectively (Table 4.2). When the measure of accuracy was adjusted for the average reliability of the EBV of the training animals the observed accuracy of genomic evaluation was 0.80, 0.82, and 0.76 for BF, D250, and LEA respectively (Table 4.2).

Table 4.2: Estimates of accuracy for genomic evaluation and individual GEBV across imputation scenarios

trait	scenario $^{\times}$	imputation accuracy *	$r_{EBV,GEBV}$ $^{\diamond}$	$\bar{r}_{EBV}$ $^{\mp}$	$\frac{r_{EBV,GEBV}}{\bar{r}_{EBV}}$	$\bar{r}_{GEBV}$	HPD $^{\triangleleft}$
BF	1	(1, 1)	0.6810^{a}	0.8510	0.7998	0.6852	$[0.5395, 0.8211]$
	2	(1, 0.95)	0.6795^{a}		0.7981	0.6861	$[0.5467, 0.8164]$
	3	(1, 0.88)	0.6585^{b}		0.7734	0.6684	$[0.5498, 0.7909]$
	4	(0.88, 0.88)	0.6598^{b}		0.7749	0.7014	$[0.5727, 0.8267]$
D250	1	(1, 1)	0.6603^{a}	0.8020	0.8229	0.6575	$[0.5073, 0.7948]$
	2	(1, 0.95)	0.6555^{ab}		0.8170	0.6585	$[0.5187, 0.7962]$
	3	(1, 0.88)	0.6521^{ab}		0.8127	0.6412	$[0.5213, 0.7771]$
	4	(0.88, 0.88)	0.6463^{b}		0.8054	0.6750	$[0.5345, 0.7985]$
LEA	1	(1, 1)	0.6516^{a}	0.8529	0.7639	0.6859	$[0.5386, 0.8325]$
	2	(1, 0.95)	0.6491^{a}		0.7610	0.6868	$[0.5377, 0.8214]$
	3	(1, 0.88)	0.6278^{c}		0.7360	0.6684	$[0.5519, 0.8054]$
	4	(0.88, 0.88)	0.6364^{d}		0.7461	0.7040	$[0.5667, 0.8330]$

$^{\times}$ scenarios 1: no imputation, 2: imputation in prediction panel ($R^2 = 0.95$), 3: imputation in prediction panel ($R^2 = 0.88$), and 4: imputation in all animals ($R^2 = 0.95$)

* accuracy of genotype imputation R^2 for training and validation animals: (R^2_T, R^2_V)

$^{\diamond}$ Tukey HSD post-hoc comparison of accuracy of genomic evaluation across imputation scenarios

$^{\mp}$ average accuracy of EBV in the validation panel

$^{\triangleleft}$ 95% highest posterior density (HPD) interval of GEBV accuracy across validation animals

We observed a significant difference between the estimates of accuracy of genomic evaluation across ten randomly assigned cross-validation datasets for three traits (Table 4.3). That variation across cross-validation datasets was partially explained by a significant effect of the average EBV accuracy of validation animals on accuracy of genomic evaluation (Table 4.3) in three traits and a significant effect of top 10 relatedness on accuracy of genomic evaluation in D250. Another source of difference of accuracy of genomic evaluation across cross-validation datasets could be the population structure. This would be revealed through differences in estimated variance components. We did not expected differences in variance components estimated from randomly assigned validation datasets. We confirmed this assumption by studying the distribution of estimated heritability ($\frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$) and included the obtained results in Supplementary Figure 4.3.

Table 4.3: Significance of variables affecting accuracy of genomic evaluation

	folds $^{\times}$		$rel10$ *		$\bar{r}_{EBV}$ $^{\diamond}$	
trait	F $^{\triangleleft}$	p	F †	p	F †	p
BF	258	< 0.001	2.83	0.1013	11.73	0.0016
D250	229	< 0.001	5.18	0.0291	7.238	0.0109
LEA	311	< 0.001	2.06	0.1605	3.430	0.0725

$^{\times}$ accuracy of genomic evaluation by randomly assigned folds of the cross-validation
* accuracy of genomic evaluation by average of the top 10 genomic relationship estimates of animals in the validation set
$^{\diamond}$ accuracy of genomic evaluation by average accuracy of EBV of validation animals by fold
$^{\triangleleft}$ $df = c(9, 27)$
† $df = c(1, 35)$

The average accuracy of the genomic evaluation and the assessment of the accuracy of individual GEBV using equation 4.14 is equally important in a practical implementation of

genomic selection. Average accuracy of individual GEBV was 0.69, 0.66, and 0.69 for BF, D250, and LEA respectively with a 95% highest posterior density (HPD) interval ranging from roughly 0.51 to 0.80 across all traits (Table 4.2).

As can be seen in Figure 4.1 accuracy of GEBV (r_{GEBV}) and accuracy of EBV (r_{EBV}) are not linearly related. accuracy of GEBV (r_{GEBV}) and accuracy of EBV (r_{EBV}) are not linearly related. Accuracy of EBV was higher than the estimated accuracy of GEBV for most animals in three traits, especially when $r_{EBV} > 0.8$. For a few animals with r_{EBV} between 0.4 and 0.8 accuracy of GEBV was higher than their respective EBV accuracy. We observed there was an almost linear increase (Figure 4.2) in r_{GEBV} as top 10 relatedness increased, which we found was statistically significant ($p < 0.01$).

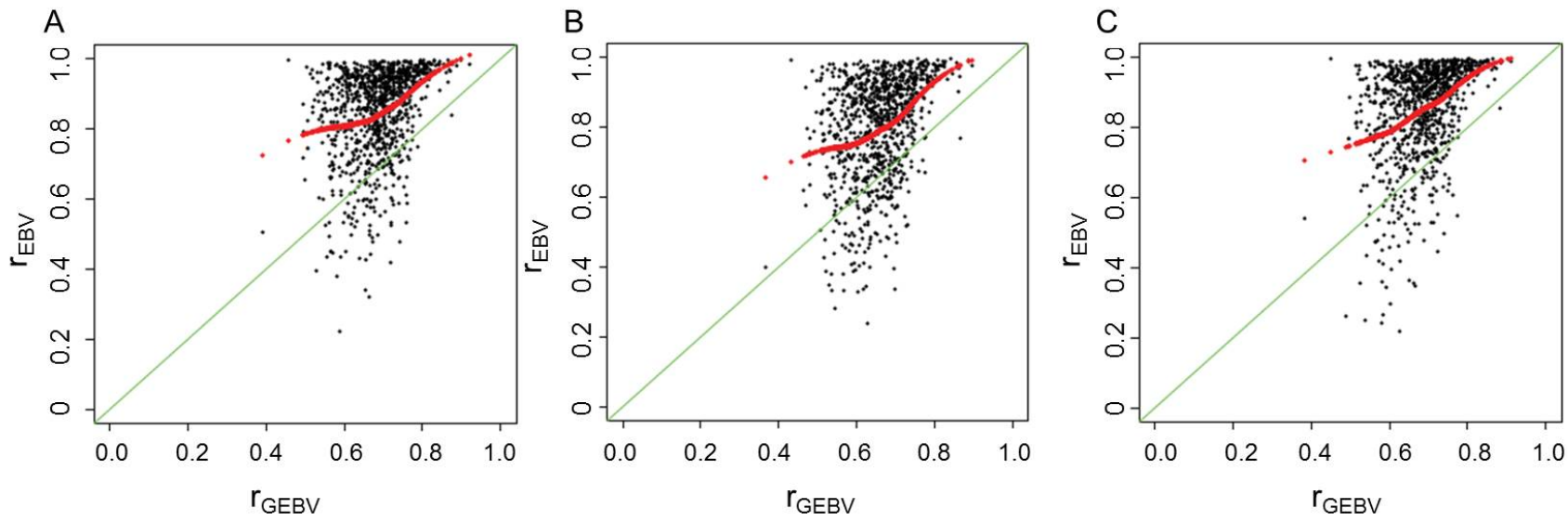

Figure 4.1: Accuracy of GEBV by observed accuracy of EBV for a) BF, b) D250, and c) LEA
r_{GEBV} in relation to the animals r_{EBV}, with the 1-1 line of the regression (green line) and a loess smoother (red line), which is a local weighted mean of the r_{GEBV}

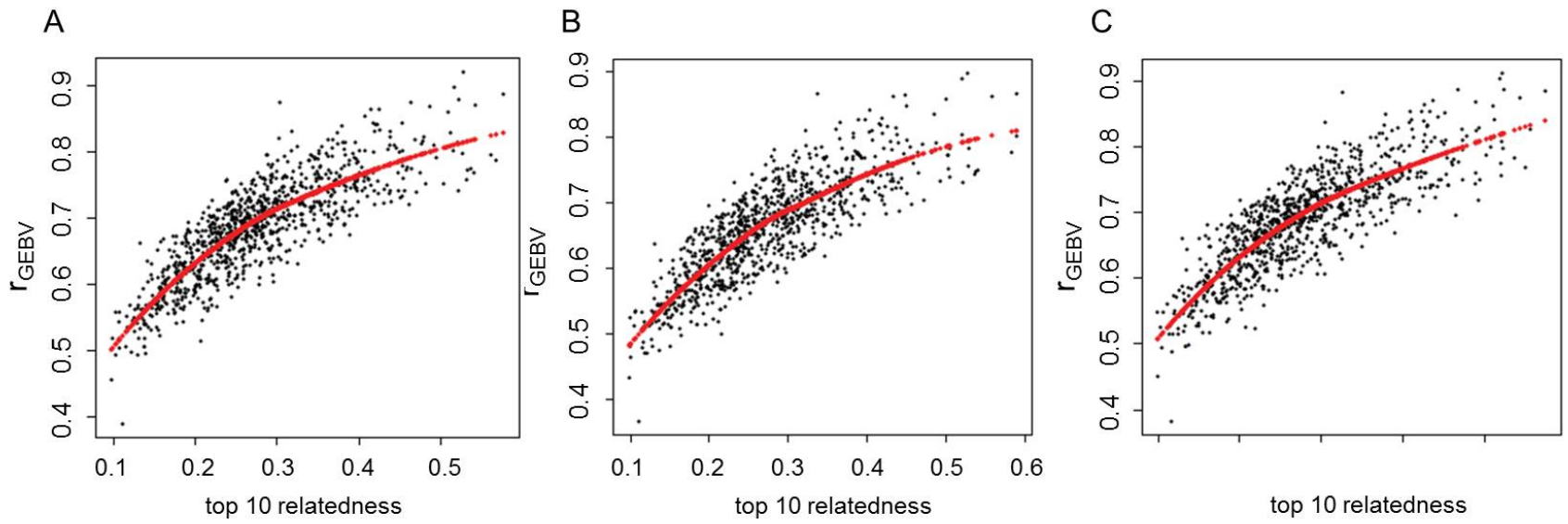

Figure 4.2: Accuracy of GEBV by average top 10 relatedness between the individual and training panel for (A) BF, (B) D250, and (C) LEA
r_{GEBV} in relation to the animals $rel10$, a loess smoother (red line), which is a local weighted mean of the r_{GEBV}

4.3.2 Effect of genotype imputation on accuracy of genomic evaluation and GEBV

Accuracy of imputation (R^2) for each animal was measured as the squared correlation between the observed and imputed allelic dosage across all SNP (Badke *et al.*, 2013). Average accuracy of imputation was $R^2 = 0.88$ for the first scenario using a small (128) haplotype reference panel, and it increased to $R^2 = 0.95$, when a larger reference panel ($\sim$ 1800 haplotypes) was utilized. In our previous study (Badke *et al.*, 2013) we found that increasing the size of the reference panel led to an improved imputation especially of SNP that appear difficult to impute, such as SNP with low (≤ 0.1) MAF and those located in the chromosomal extremes. These results were replicated in this study (Supplementary Figure 4.4).

For BF we found that the average accuracy of genomic evaluation under scenario 2 ($r_{GEBV,EBV} = 0.6795$), where genotypes in the validation animals were imputed with high accuracy ($R^2 = 0.95$), was not significantly different from the accuracy observed in the reference scenario ($r_{GEBV,EBV} = 0.681$), where all genotypes were observed. However average accuracy of genomic evaluation was significantly lower ($r_{GEBV,EBV} = 0.6585$, $r_{GEBV,EBV} = 0.6598$), when genotypes were imputed with lower accuracy ($R^2 = 0.88$) when using a small reference panel of haplotypes (scenarios 3 & 4). For D250 there was no significant difference in accuracy of genomic evaluation between the reference design ($r_{GEBV,EBV} = 0.6603$) and the two scenarios where genotypes were imputed in the validation animals (Table 2). However, when genotypes were imputed in both training and validation (scenario 4) accuracy of genomic selection was significantly lower ($r_{GEBV,EBV} = 0.6463$). For LEA there was also no difference in accuracy of genomic evaluation between the reference scenario ($r_{GEBV,EBV} = 0.6516$) and scenario 2 ($r_{GEBV,EBV} = 0.6491$).

There was a significant decrease in accuracy of genomic evaluation when genotypes were imputed with lower accuracy ($R^2 = 0.88$) in scenarios 3 ($r_{GEBV,EBV} = 0.6278$) and 4 ($r_{GEBV,EBV} = 0.6364$).

To assess the effect of genotype imputation on the results of a genomic evaluation we compared the top 5% sires ($n = 46$), ranked by their estimated GEBV across imputation scenarios. Again, scenario 1 was used as a reference design to compare how many of the top 5% ranked animals were also top ranked under the imputation scenarios (2-4). The proportion of top 5% ranked sires that were conserved when genotypes were imputed in validation animals with high accuracy (scenario 2) was 0.96 for BF and 0.98 for D250 and LEA. When genotypes were imputed in validation animals with lower accuracy (scenario 3) the proportion of top 5% ranked sires decreased to 0.86, 0.92, and 0.87 for BF, D250, and LEA respectively. When genotypes were imputed in training and validation the proportion of top 5% sires conserved in comparison to the reference design showed a small increase compared to the design with only validation animals imputed for BF (0.88), a small decrease for D250 (0.89), and a more substantial decrease for LEA (0.81).

Accuracy of individual GEBV is estimated using the genomic relatedness between training and validation animals. Using genotypes imputed with high accuracy ($R^2 = 0.95$) the estimated r_{GEBV} remained constant in all traits, compared to estimates obtained from observed genotypes. When genotypes were imputed with less accuracy ($R^2 = 0.88$) r_{GEBV} slightly decreased. However, when the genomic relationship matrix was obtained from imputed genotypes in both training and prediction animals ($R^2 = 0.88$) the observed accuracy of GEBV was higher than even the reference scenario across traits. Examining the estimation procedure for r_{GEBV} we found that this difference was due to smaller estimates of the diagonal elements of the genomic relationship matrix between the validation elements

($\mathbf{G}_V$) in the scenario with all imputed genotypes. These diagonal elements were used to scale values of r_{GEBV} (equation 4.14), and smaller values in the denominator resulted in the larger estimates of r_{GEBV} we saw for animals in scenario 4. Comparing unscaled values of r_{GEBV} individual accuracy was higher in the reference scenario for all animals.

4.4 Discussion

4.4.1 Accuracy of genomic evaluation and GEBV using observed genotypes

The size of the training population used to train the prediction equation in this study was small compared to previous genomic evaluations published in swine (Cleveland *et al.*, 2010, 2012), and especially compared to studies applying genomic evaluation in European (Dassonneville *et al.*, 2011) or US dairy cattle (Weigel *et al.*, 2010a; Wiggans *et al.*, 2012). Observed accuracy of genomic evaluation in this study was in good agreement with previously published results for genomic evaluation in pigs, assessing five unspecified commercial traits with comparable heritability (Cleveland *et al.*, 2012) and earlier results for two reproductive traits (Cleveland *et al.*, 2010). Accuracy of genomic evaluation was high across three traits (BF: $r_{GEBV} = 0.6810$ D250: $r_{GEBV} = 0.6603$, LEA: $r_{GEBV} = 0.6516$). In addition, we report accuracy adjusted for the fact that the Pearson correlation between EBV and GEBV will underestimate the true quantity of interest (Luan *et al.*, 2009). Assessing the variation in accuracy of genomic evaluation across datasets of the cross-validation, we found that the $\bar{r}_{EBV}$ of the validation animals and their relatedness to the training animals were significantly associated to the average accuracy of genomic evaluation. Higher accuracy of

genomic evaluation of prediction animals with close relatives in the training population (Habier *et al.*, 2010; Clark *et al.*, 2012) and within closely related populations, with relatively small effective population size, has been previously reported (Daetwyler *et al.*, 2013). Accuracy of genomic evaluation in this study was high in spite of the limited number of animals available for training and the inclusion of animals with relatively low EBV accuracy. Furthermore, we obtained accurate genomic predictions using an equivalent model fitting the genomic relationship matrix instead of a marker based matrix (Hayes *et al.*, 2009b), thereby greatly reducing the computational load. We expect that accuracy of genomic evaluation in this population, and other US swine populations with comparable population structure and LD (Badke *et al.*, 2012), will be feasible for commercial implementation and could be further increased through the inclusion of additional training animals with highly accurate EBV.

Besides assessing the accuracy of genomic evaluation we also reported accuracies for individual GEBV. The accuracy of GEBV will be important to influence selection decisions, but as proposed by Goddard *et al.* (2011), can also be approximated prior to the implementation of genomic evaluation and used to inform the design of genomic selection in a population. As expected, we observed that accuracy of GEBV increased with increased relatedness between the animal and the training panel. Several previous studies in other populations and simulation experiments also showed the importance of relatedness for the prediction of accurate GEBV (Habier *et al.*, 2010; Clark *et al.*, 2012), especially when the training population was small (Wientjes *et al.*, 2013) as was the case in our study. In addition, we observed that accuracy of GEBV was higher than accuracy of EBV for only a few animals that had mostly low accuracy of EBV. This finding is further supported by previous reports that implementation of genomic evaluation would be most beneficial for young animals with little information on their own and subsequently low accuracy of traditional EBV (VanRaden, 2008).

4.4.2 Effect of genotype imputation on accuracy of genomic evaluation and GEBV

Genotype imputation is an efficient tool to decrease the cost of obtaining high density genotypes for selection candidates. It was the goal of this study to quantify the loss on accuracy of genomic evaluation if GEBV were estimated from imputed rather than observed genotypes in selection candidates. Comparing accuracy of genomic evaluation across four scenarios of genotype imputation we found that for three traits there was no significant difference between genomic evaluation as a function of genotypes in validation animals being observed or imputed with high accuracy ($R^2 = 0.95$). Accuracy of genomic evaluation decreased in comparison to the reference scenario when genotypes in selection candidates were imputed with lower overall accuracy ($R^2 = 0.88$). Especially when imputation was applied in training and prediction animals we observed a decrease in accuracy of genomic evaluation, such that while this would be the most cost efficient scenario, it would not be feasible as practical implementation to obtain maximally accurate results of genomic evaluation. Accuracy of genotype imputation is a function of the SNP density, the size of the reference panel, the level of LD between adjacent SNP, and the SNP chromosomal location and MAF (Badke *et al.*, 2013). Combined with the results from this study that a minimum accuracy of imputed genotypes was necessary to conserve accuracy of genomic evaluation, these variables can be used to design an optimally cost efficient scheme of genomic selection with genotype imputation in a population without any current use of molecular markers to estimate genetic merit. Once genomic selection has been successfully implemented, the constant influx of additional animals with high density genotypes and accurate EBV will only serve to increase the accuracy of both imputation and subsequent genomic evaluation. Previously published

results support that while it is not feasibly to implement genomic prediction based on low density genotypes (Habier *et al.*, 2009; Cleveland *et al.*, 2010), even if SNP were preselected for association with the phenotype, accuracy of genomic evaluation is still feasible for practical implementation when genotypes in selection candidates are accurately imputed to high density (Weigel *et al.*, 2010a; Cleveland and Hickey, 2013). In addition, several studies also support that an increase in imputation accuracy will facilitate results of genomic evaluation nearly indistinguishable from those obtained from observed genotypes (Dassonneville *et al.*, 2011; Wiggans *et al.*, 2012; Cleveland and Hickey, 2013) while the cost efficiency of low density genotypes allows a much larger proportion of the population to be included in the genomic evaluation procedure (Wiggans *et al.*, 2012). In conclusion, an implementation of genomic selection based on observed genotypes for training of the prediction equation and GEBV predictions obtained from genotypes imputed with high accuracy appears to be a promising approach to provide the swine breeding industry with a cost efficient procedure to obtain GEBV for animals at a young age. A recent study assessing the accuracy of genomic evaluation using high density genotypes and various imputation schemes in a commercial pig population further supports these findings (Cleveland and Hickey, 2013).

We found that accuracy of individual GEBV was a linear function of the relatedness between a validation animal and the respective training set. As has been previously shown in the literature, animals that are highly related to the training population will have higher r_{GEBV} (Habier *et al.*, 2010; Clark *et al.*, 2012). For scenarios with observed genotypes used in training animals the average r_{GEBV} was in good agreement with the overall accuracy of genomic evaluation. However, when genotypes were imputed in training and prediction animals, the average r_{GEBV} was notably larger than the accuracy of genomic evaluation, which we found was an artifact of lower estimates of the diagonal elements of the **G** matrix.

This was caused by a decrease in the variance of the allelic dosage of imputed genotypes due to the relatively small number of reference haplotypes available. When the variance of imputed allelic dosages was decreased, the deviation from the expected value estimated from MAF ($2\mathbf{p}$) also decreased, causing overall smaller estimates of $\mathbf{Z}$ and the resulting diagonal elements of the $\mathbf{G}$ matrix. This increase in the homogeneity of allelic dosages in the imputed genotypes causes the observed inflation in accuracy of estimated GEBV, such that in any case when GEBV are obtained from imputed genotypes the estimated accuracy of GEBV should be used with caution. The average GEBV accuracy notably exceeded the expected accuracy of genomic evaluation in that scenario.

In conclusion, we found that results for accuracy of GEBV further support the notion that genomic evaluation using high density genotypes imputed with high accuracy for selection candidates is a feasible method to implement a cost efficient design for genomic selection in swine. When genotypes were imputed with lower accuracy in training and prediction animals accuracy of genomic evaluation was significantly decreased and estimates of accuracy of GEBV were inflated. As mentioned before, all code and data used in this paper has been made available through an R package, accessible at: `http://tinyurl.com/MSURGEBV`

4.5 Supplementary Materials

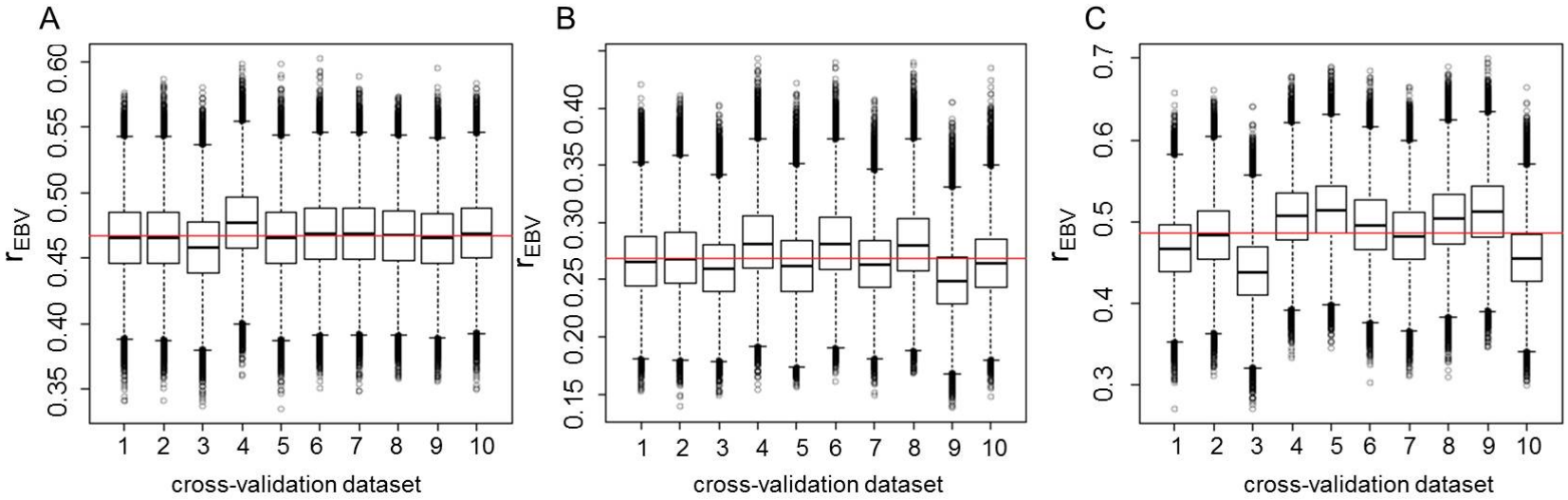

Figure 4.3: Distribution of genomic heritability across 10 folds for a) BF, b) D250, and c) LEA

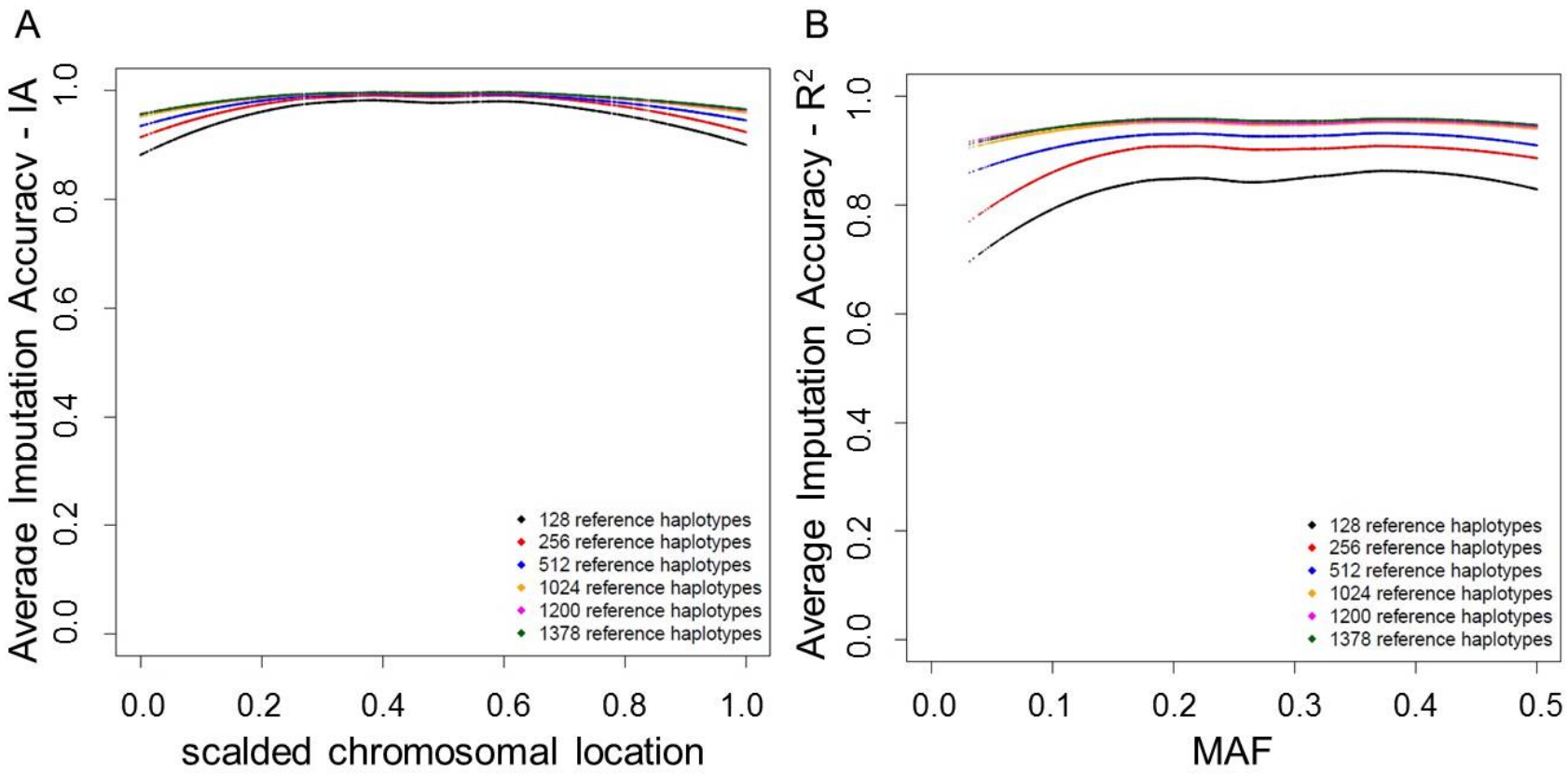

Figure 4.4: Average accuracy of genotype imputation for imputation from a small (blue) or large (red) reference panel as a function of (A) chromosomal location of SNP and (B) MAF

Chapter 5

General Discussion

Genomic selection is a valuable tool to genetically improve livestock and plant species (Daetwyler *et al.*, 2013). Direct use of genome wide marker information allows for accurate selection of superior animals for breeding at a very young age, even if the phenotype under selection is difficult to measure (Daetwyler *et al.*, 2013). Furthermore, results from dairy cattle breeding show that implementation of genomic selection can shorten generation intervals while increasing the rate of genetic gain (VanRaden *et al.*, 2009). Through the availability of the PorcineSNP60 BeadChip (Ramos *et al.*, 2009) implementation of genomic selection for swine breeding has become a probable development (Cleveland and Hickey, 2013). It was the objective of this dissertation to assess variables affecting accuracy and cost efficiency of genomic selection programs in four US pure-breed pig populations. We intended to show how genomic selection could be optimally implemented meeting the swine breeding industry's requirements of high prediction accuracy and cost efficiency. In addition, we released computer algorithms and data to facilitate further research in these specific populations and allow researchers working with other species to implement a similar set of steps to investigate the usability and optimal design of genomic prediction.

5.1 Objectives revisited and their impact on genomic selection in swine breeding

1. Estimate LD and in the Duroc, Hampshire, Landrace and Yorkshire pig breeds using SNP genotypes obtained using the Illumina PorcineSNP60 Genotyping BeadChip. Determine persistence of phase between breeds to assess the potential of mixed breed reference panels for both imputation and genomic selection in the future.

LD describes the non-random association between gametes at different loci across the genome. High genome wide pair-wise estimates of LD are an important precursor for genome wide association studies (GWAS) and accurate genomic prediction. The level of LD in pigs was previously measured using low density markers (Du *et al.*, 2007; Harmegnies *et al.*, 2006; Nsengimana *et al.*, 2004), but our study of LD in four US pig breeds (Badke *et al.*, 2012) was among the first to report genome wide levels of LD using a high density SNP panel (Jafarikia *et al.*, 2010; Uimari and Tapio, 2011). Indicative of a relatively small effective population size we found average LD of approximately 0.4 for markers within 100kb of each other, and 0.2 when average distance between SNP was 1 Mb. Due to these relatively high estimates of pairwise LD, even at increasing distance we projected that implementation of genomic selection has the potential to achieve accuracy of prediction comparable to that observed in dairy cattle populations. We based this projection on the fact that pairwise LD between SNP is an important prerequisite for accurate prediction (i. e. Hayes *et al.*, 2009a), and that LD in these pig breeds was higher than reported estimates for dairy cattle populations (de Roos *et al.*, 2008). Since persistence of LD at increasing distances was high, genotype imputation from a more cost efficient low density SNP panel should be considered in the design of a genomic selection scheme.

Implementation of genomic selection requires a large sample of training individuals with highly accurate EBV to estimate SNP effects. Especially in small populations obtaining a large number of adequate training animals is often cost-prohibitive. Combining training animals across populations to reduce cost has previously been proposed and is based on the assumption that the extent and phase of LD, and the effect of the QTL are conserved across populations. We obtained estimates of correlation of phase, as a measure of persistence of phase, for all pairwise comparisons among four US pig breeds. Our results suggest, that persistence of phase between Duroc, Landrace, Hampshire, and Yorkshire is not sufficient to support the hypothesis that combining any of these populations to train a prediction equation for genomic selection would be beneficial. Persistence of phase will generally increase with increasing marker density, such that this recommendation should be reevaluated as the density of the available SNP chip increases (Goddard *et al.*, 2006). In conclusion, we found promising levels of genome wide LD in these pig populations, but implementation of genomic selection should remain breed specific as marker phase is not sufficiently conserved at the current marker density.

2. Assess tagSNP selection strategies to obtain maximally informative subsets of tagSNP that effectively span the genome for each breed. In addition, report on imputation accuracy of the recently released GeneSeek Genomic Profiler for Porcine LD (GGP-Porcine, GeneSeek a Neogen Company, Lincoln, NE), a commercially available 10K tagSNP panel.

Although results presented in the previous chapter indicate that genome wide LD is sufficient to support the expectation of accurate genomic prediction, implementation based on high density SNP remained unlikely due to cost concerns. A common approach to decrease

genotyping costs used in human (i. e. Howie *et al.*, 2011), plant (Hickey *et al.*, 2012), and animal research (i.e. Wiggans *et al.*, 2012) is genotype imputation. Accuracy of genotype imputation depends on the effective size of the population, the size and composition of the reference panel (Howie *et al.*, 2011; Hayes *et al.*, 2012), the proportion of tagSNP, the location and MAF of untyped SNP (Badke *et al.*, 2013; Hickey *et al.*, 2012), and the type of information used for imputation (Marchini and Howie, 2008). To establish a baseline for imputation accuracy, we applied a population LD based imputation algorithm, mimicking the case of a population with no pedigree available. In case linkage information is available and utilized imputation accuracy is expected to further increase (Huang *et al.*, 2012). Low density SNP panels (tagSNP) selected exploiting population wide LD generally outperformed tagSNP set selected for even spacing along the genome. However, if pairwise LD is exploited the resulting panels are population specific. Therefore, when the GGP-Porcine was released parallel to our efforts of tagSNP selection, we decided to refocus our attention and assess genotype imputation accuracy based on this commercially available 10K platform.

Imputation accuracy (IA), measured as the proportion of correctly imputed alleles, was high (0.95) when imputation was based on the GGP Porcine and a small reference panel of 128 haplotypes. Relatedness to the reference panel had a positive impact on accuracy of imputation even if it was not directly exploited through the imputation algorithm. SNP within the chromosomal extremes or low MAF had on average lower accuracy of imputation (Badke *et al.*, 2013; Hickey *et al.*, 2012). Increased density of tagSNP in the chromosomal extremes was implemented to improve imputation accuracy of SNP located in these areas. In the GGP Porcine tagSNP density is approximately doubled in the 10% extreme locations. Secondly, increasing the size of the reference population did increase overall accuracy of imputation, but especially imputation accuracy of SNP with low MAF or those located in

the chromosomal extremes benefited from adding haplotypes. In our population an ideal number of reference haplotypes to ranged between 1000-1200, beyond which any increase in accuracy was marginal. This information can be utilized to optimize the allocation of funds available to implement genomic selection between genotyping reference animals, genotyping of selection candidates at low density, and phenotype collection. In conclusion, genotype imputation from the commercially available GGP-Porcine is sufficiently high ($IA = 0.95$) to expect that genomic prediction obtained from imputed genotypes will suffer only marginal losses in accuracy compared to those obtained from observed genotypes. The design of the GGP Porcine (Curtis P. Van Tassell, SNPSpace), selecting tagSNP evenly covering the genome with maximal MAF to ensure segregation, can serve as an example for other populations. Our results on individual and SNP specific accuracy can be used to design an optimal reference panel of haplotypes and select individuals for imputation who are expected to yield imputed genotypes with high accuracy.

3. Perform GEBV prediction for economically important production traits using high density SNP genotypes for the Yorkshire breed, as well as genotypes imputed from the GGP-Porcine. We assessed the loss in accuracy when genotypes in selection candidates were obtained through a more cost efficient low density panel (GPP-Porcine) instead of observed high density genotypes (PorcineSNP60).

Results from chapters 2 and 3 indicated that genomic selection in the Yorkshire pig population has the potential to be highly accurate, even if genotypes in selection candidates are imputed. We implemented a ten-fold cross validation study to assess the accuracy of genomic evaluation from observed genotypes, and the loss in accuracy of genomic evaluation if genotypes were imputed. Genomic selection was implemented using a computationally

efficient animal model, and accuracy of individual GEBV was calculated through inversion of the mixed model equations. As expected, we found that accuracy of genomic evaluation for three traits (back-fat thickness, loin muscle area, and # days till 250lb) were in good agreement with the few available results from pig populations (Cleveland *et al.*, 2010, 2012) and mimicked previously observed accuracies for various dairy cattle traits (Dassonneville *et al.*, 2011). Higher relatedness between the training animals and the validation animals had a positive impact on the observed accuracy, as did higher average accuracy of EBV in the validation animals. We implemented three separate imputation scenarios: 1. genotypes were imputed in the validation animals using $\sim$ 1800 reference haplotypes, 2. genotypes were imputed in the validation animals using 128 reference haplotypes, and 3. genotypes were imputed in all animals using 128 reference haplotypes. Accuracy of imputation was assessed as the squared correlation (R^2) between observed and imputed genotypes. The first imputation scenario had the highest observed accuracy of imputation (0.95), which was expected due to the larger number of reference haplotypes. Genomic evaluation based on these genotypes did not show any significant difference from the reference design without imputation in any of the traits. The second and third scenario had slightly lower accuracy of imputation (0.88) and as a result a decrease in accuracy of genomic evaluation compared to the reference design was observed in all three traits. However, while the average accuracy of genomic selection was decreased when genotypes were imputed in all animals, estimates for individual GEBV accuracy were inflated in this scenario. As a result we would not recommend using genotypes imputed in all animals for genomic selection in pig breeding. In conclusion, we were able to replicate previous results from pig (Cleveland and Hickey, 2013) and dairy cattle breeding (Dassonneville *et al.*, 2011; Weigel *et al.*, 2010a; Wiggans *et al.*, 2012), reporting that genomic evaluation based on highly accurate imputed genotypes in

selection candidates is a cost efficient and equally accurate alternative to genomic evaluation based on observed high density genotypes.

5.2 Future Directions

In chapter 2 we exploited haplotypes to estimate LD and persistence of LD across four breeds of US pigs. We focused on levels of LD to assess the future potential of low density genotypes in these populations for genotype imputation and genomic selection. Alternatively, available haplotypes could be exploited for breed identification and quantification of breed purity. Using high density genotypes to estimate breed composition could replace cost and time consuming mating experiments to determine an animals purebred status. Recent results from our group (Huang *et al.*, 2013) were successful in predicting an animals breed composition using regression models, and identifying potentially non-purebred or crossbred animals that should be further evaluated. The animals identified as non-purebred were partially identical with a group of animals that had consistently low imputation accuracy in chapter 3, implicating that this method was in fact able to identify individuals with a distinct haplotype structure. However, for roughly half these animals differences in their haplotype structure was not the result of cross-breeding, instead they were imported from an unrelated population of the same breed. Further research is necessary to separate this effect of breed composition vs. population of origin, and further validate the ability of high density genotypes to reliable predict recent cross-breeding events in an animals pedigree. Especially, it would be interesting to assess the effect of combining different populations of the same breed to be used as a reference for breed composition predictions (e. g. Finnish populations used in Uimari and Tapio, 2011). Another issues with using high density geno-

types to predict breed purity is admixture between breeds. In chapter 2, using short-range persistence of phase we approximated time since breed divergence between four US pig populations to be between 40 and 60 generations, with white breeds having diverged from each other more recently (40). However, long-range persistence of phase between all four breeds is larger than the expected value based on the approximated time of breed divergence, possibly due to more recent admixture between the populations. Using high density genotypes to obtain regression estimates of breed composition will likely reflect these past cross-over events between populations. This creates the need for thresholds of significance to be established indicating whether an individual's divergence from its assigned breed is large enough to necessitate further testing to assess the animals pure-bred status. To assess the ability of haplotypes obtained from high density genotypes to predict breed purity within and across populations of the same breed a panel of reference haplotypes could be assembled. Many populations, especially in commercial settings, will have haplotypes available from previous research. Our results for reference panel design in chapter 3 suggest that in these pig populations haplotypes from randomly sampled animals will not be at a disadvantage compared to more complex sampling designs involving sire/dam/offspring trios or maximally unrelated individuals provided a large ($N \geq 100$) sample is assembled. As a result, using available resources with minimal collection of additional genotypes from populations with no previous genotyping should result in a usable reference panel for breed composition prediction. Validation animals to assess the accuracy of breed composition predictions with known cross-bred ancestry will be necessary. Especially commercial settings, where cross-breeding is commonly implemented to obtain market pigs could provide validation animals with genotypes already available.

In chapter 3 we estimated imputation accuracy from low (GGP Porcine) to high (Porci-

neSNP60) density SNP arrays. Advances in genotyping technology caused a significant decrease of the cost associated with sequencing, such that whole genome sequence is a likely future source of data. Hayes *et al.* (2009a) noted that accuracy of genomic selection is a function of the LD between the observed SNP and the QTL, such that accuracy can be increased through an increase in marker density. If genomic selection is based on sequence, prediction is no longer limited by the extent of LD between marker and causative polymorphism, as the later would be directly observed. As a result, accuracy of genomic selection based on whole genome sequence is expected to be superior compared to results obtained from high density SNP (Druet *et al.*, 2013; Meuwissen and Goddard, 2010). In addition, the currently observed decay in prediction accuracy observed over generations is expected to significantly decrease when causative polymorphism are directly observed (Meuwissen *et al.*, 2013). Results from simulation experiments indicate that a 5-10% increase in accuracy of genomic evaluation can be expected through the use of sequence data (Druet *et al.*, 2013; Meuwissen and Goddard, 2010), which needs to be further validated in real data. This advantage appeared especially pronounced for QTL with low allele frequency (Druet *et al.*, 2013). Inclusion of whole genome sequence will result in a data structure including animals with low, high, and sequence density SNP, such that genotype imputation will continue to be an important tool of genomic analysis. Current implementation of genotype imputation in livestock populations is mainly based on reference panels of haplotypes sampled from within the population (Badke *et al.*, 2013; Hayes *et al.*, 2012; Wiggans *et al.*, 2011). Accuracy of imputation from population specific reference panels appears to be superior to mixed reference panels at the current marker density and persistence of phase between populations (de Roos *et al.*, 2009). However, due to the cost of whole genome sequence, whole-sequence reference haplotypes will likely be sampled across populations, such that the resulting panels will be

highly admixed (e .g `http://www.1000bullgenomes.com/`). Genotype imputation in human populations has shown that imputation in admixed populations from diverse reference haplotypes can be highly accurate (Howie *et al.*, 2009, 2011). In particular, a subroutine implemented in Impute v2 sampling reference haplotypes based on a euclidean distance for each imputation sample, thereby assembling a reference panel of 'related' haplotypes was highly successful (Howie *et al.*, 2011). Differently from human populations effective population size in most livestock species is small and detailed pedigrees are available. As a result across population reference panels can be assembled combining representative samples of highly influential individuals (Druet *et al.*, 2013) from each sub-population, and combining pedigree information, such that whole sequence imputation can be based on both linkage and LD information (e. g. AlphaImpute Hickey *et al.*, 2011).

An initial study investigating the use of whole genome sequence for genomic research in pigs should aim to address the following issues:

1. Optimal strategy for assembling a reference panel of haplotypes at sequence density.

Sequencing coverage is an important variable affecting the certainty with which heterozygotes can be called (Druet *et al.*, 2013). However, higher coverage will also lead to a substantial increase in cost, such that the relation between the number of animals and minimal coverage should be optimized to allow for a maximal sample size of accurately called haplotypes. Chen *et al.* (2013) introduce a design that allows for increased variant detection and genotype calling even at low coverage through the use of trios instead of unrelated individuals. Using simulated data, emulating human populations, they found that 30 trios would outperform 90 unrelated individuals in numbers of variants detected for low coverage (1x-2x), but the effect was reversed as fold coverage increased to 8x (Chen *et al.*, 2013). They attributed this

difference to the additional familial information available to resolve variants and genotypes at low coverage in trios (Chen *et al.*, 2013). In chapter 3 we compared reference haplotype panels consistent of haplotypes derived from a trio design to cost equivalent haplotypes obtained from unrelated individuals ($N_{trio} = 64$, $N_{unrelated} = 96$) and found an advantage in imputation accuracy using trio based panels as reference only for extremely small reference panels ($N \leq 32/48$). Similarly, this advantage of the trio design when the number of reference haplotypes was very small is likely a result of the increased phasing accuracy that can be obtained from a trio design (Marchini *et al.*, 2006). Implementing the concepts of Chen *et al.* (2013) and expanding them include other close relatives to increase the number of variants detected and genotypes called could aid the assembly of whole sequence reference haplotypes. Exploiting the availability of detailed pedigrees and overall high relatedness between animals in livestock populations has the potential to facilitate accurate reference panels derived from cost efficient low coverage whole genome sequence.

2. Implementation of genotype imputation and assessment of strategies to optimally exploit admixed reference haplotype panels.

Accuracy of imputation within a population can be maximized through the combined use of pedigree (linkage) and LD information (Hickey *et al.*, 2011). However, as we expect reference panels for sequence imputation to be highly admixed, a shared pedigree may not be available or very limited, such that direct exploitation of linkage may prove difficult. Implementation of algorithms such as Impute that internally assemble a reference panel of 'related' individuals (Howie *et al.*, 2011) may be better suited to exploit admixed reference panels of haplotypes. One approach to optimize overall accuracy could be an initial imputation based on the few within population reference haplotypes using a combined linkage and LD algorithm.

The results of that imputation are then followed by a secondary imputation focusing on haplotypes that could not be resolved with appropriate certainty using the admixed reference panel and an algorithm establishing indirect relationships between individuals. In addition, variables affecting accuracy of imputation such as sequence coverage, MAF, marker location, and size of the reference panel of haplotypes for high density to sequence imputation should be assessed.

3. Assess the gain in accuracy of genomic evaluation if prediction is based on whole genome sequence (imputed) compared to high density SNP.

Simulation studies have projected between 5-10% gain in accuracy (Meuwissen and Goddard, 2010) of genomic evaluation. However, a study by VanRaden *et al.* (2011) using simulated ultra high density SNP for genomic prediction in comparison to the widely implemented 50K SNP panel found only a small gain in accuracy as a function of increased marker density. They concluded that the size of the training population had a greater impact on improving prediction accuracy than the tested increase in marker density (VanRaden *et al.*, 2011). In chapter 4 we evaluated the accuracy of genomic evaluation in a US Yorkshire population using a training population of approximately 850 animals per fold. Observed significant differences in accuracy between folds of the cross-validation could indicate that this sample-size was not large enough to obtain effects invariant of the current sample. Therefore, a simulation study using the available pedigree and high density genotypes as a basis to obtain realistic ultra high density genotypes (Hickey and Gorjanc, 2012) could be used to compare the effect of increased training panel size vs. an increase in marker density in this population. The possible benefits of genomic evaluation based on whole genome sequence should be weighted against the additional costs collection of sequence data would incur.

4. Assess the potential gain of using whole genome sequence for GWAS

GWAS based on sequence data has the potential to directly identify causative polymorphisms. This would enable research to focus on understanding the molecular processes involved, instead of focusing on potential candidate genes within a QTL region. However, the number of parallel tests would further increase, such that an even larger reference population is likely necessary to provide enough power to allow detection of causative polymorphisms.

Many of these issues could be initially addressed in a simulation study using the genotype data and pedigree available from this research to inform the simulation design, using e. g. the program presented by Hickey and Gorjanc (2012). In addition, if whole genome sequence is already available for some animals the approach presented in Macleod *et al.* (2013) that was utilized by Druet *et al.* (2013) to implement a simulation with similar objectives in cattle could further refine the design. Previous results in our own research (not published) clearly indicate that simple simulation based on the estimated past and current effective population size using a gene-drop model (Sargolzaei and Schenkel, 2009) will likely lead to a pattern of LD that poorly represents real data, such that results cannot be extrapolated to real data examples. High density genotypes are currently publicly available for four US pig populations (Badke *et al.*, 2012) and a US commercial population (Cleveland *et al.*, 2012), while many more research projects have also collected high density genotypes on various pig-breeds worldwide (e. g. Uimari and Tapio, 2011). In addition, a thorough pedigree is available for the Yorkshire population utilized in this research. Initial efforts are underway to collect whole genome sequence on a few select animals in Europe (Martien Groenen, presentation). Collecting and combining this data should provide a solid base to

obtain a simulated dataset that is an adequate representation of whole genome sequence data in swine breeds. The diversity and size of the available data would allow the formation of a 'simulation training panel', used to estimate within and across population structure of LD, time since divergence, as well as size of the ancestral population, and the formation of a 'simulation validation panel', to which the obtained simulated data could be compared. To assess the practicality of this design a smaller initial simulation based on solely the Yorkshire data could use the estimates of r^2 from chapter 2 to determine an approximate size of the ancestral population and use the available pedigree to obtain simulated high density genotypes (Hickey and Gorjanc, 2012). Subsequently, the LD structure within the simulated genotypes, as well as persistence of phase with the observed data of Hampshire, Landrace, and Duroc can be estimated to assess similarity between the simulated and observed genotypes. Phenotypes can be simulated based on a varying number of QTL that will be randomly assigned to markers simulated above. The corresponding substitution effects of QTL could be sampled from a t-distribution and phenotypes obtained by summing the respective QTL effect and adding random residuals sampled from a normal distribution. The data released as part of this study contains EBV, dEBV, and heritabilities for three economically important traits in swine. The heritabilities as well as distribution of EBV can be used to inform the choice of parameters for the simulation of phenotypes with the goal of mimicking the structure of actual traits. Furthermore, an initial GWAS performed on these traits could provide an estimate of the number of QTL with large and medium effects likely affecting each trait, providing a basis for these parameters as well. Obtaining a simulated data set reflective of the diversity and real structure of future data is important to ensure that results can be extrapolated to the real data case and used to inform investment decisions for future research.

In conclusion, the results presented in this dissertation can directly inform design decisions for the implementation of genomic evaluation in the US Yorkshire pure-bred population utilized. Furthermore, due to the similarity of LD structure between the four breeds assessed in chapter 2 we expect estimates of imputation accuracy of and accuracy of genomic evaluation to be similar, once large enough samples of high density genotypes can be assembled, for Landrace, Hampshire, and Duroc. Public availability of all data collected and computer code written for this research facilitates the translation of the methods and design to other populations. Combining the haplotypes of these four US pure-bred swine populations with genotypes of other swine populations can facilitate a more detailed assessment of admixture between populations and breed composition for individual animals. Also, the data can be used as outlined above to inform simulation studies that will be able to allow a first assessment of the usability of whole genome sequence data for genomic evaluation, especially with genotype imputation, and genome wide association studies.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abecasis, G. R., S. S. Cherny, W. O. Cookson, and L. R. Cardon, 2002 Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics **30**: 97–101.

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.

Archibald, A. L., N. E. Cockett, B. P. Dalrymple, T. Faraut, J. W. Kijas, *et al.*, 2010 The sheep genome reference sequence: a work in progress. Animal genetics **41**: 449–53.

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, J. Fix, *et al.*, 2013 Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC genetics **14**: 8.

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.

Becker, R. A., J. M. Chambers, and A. R. Wilks, 1988 *The New S Language*, volume 1 of *Computer Science Series*. Chapman & Hall.

Berry, D. P., and J. F. Kearney, 2011 Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. animal **5**: 1162–1169.

Boichard, D., H. Chung, R. Dassonneville, X. David, A. Eggen, *et al.*, 2012 Design of a Bovine Low-Density SNP Array Optimized for Imputation. PLoS ONE **7**: e34130.

Browning, B. L., 2011 Documentation of BEAGLE 3.3.1.

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**: 210–223.

Browning, S. R., and B. L. Browning, 2011 Haplotype phasing: existing methods and new developments. Nature Reviews Genetics **12**: 703–714.

Chen, W., B. Li, Z. Zeng, S. Sanna, C. Sidore, *et al.*, 2013 Genotype calling and haplotyping in parent-offspring trios. Genome research **23**: 142–51.

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. Genetics, selection, evolution : GSE **44**: 4.

Cleveland, M. A., S. Forni, D. J. Garrick, and N. Deeb, 2010 Prediction of Genomic Breeding Values in a Commercial Pig Population. In *9th WCGALP, Leipzig, Germany*. Leipzig, 0266.

Cleveland, M. A., and J. M. Hickey, 2013 Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. Journal of animal science .

Cleveland, M. A., J. M. Hickey, and S. Forni, 2012 A common dataset for genomic analysis of livestock populations. G3 (Bethesda, Md.) **2**: 429–35.

Cleveland, W. S., E. Grosse, and W. M. Shyu, 1992 *Statistical Models in S*. Chapman & Hall/CRC.

Cohen, J., P. Cohen, S. G. West, and L. S. Aiken, 2003 *Applied Multiple Regression - Correlation Analysis for the Behavioral Sciences*. LAWRENCE ERLBAUM ASSOC Incorporated, Mahwah, NJ, third edition.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall, *et al.*, 2006 A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature genetics **38**: 1251–60.

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

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de Los Campos, and J. M. Hickey, 2013 Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics **193**: 347–65.

Dassonneville, R., R. F. Brondum, T. Druet, S. Fritz, F. Guillaume, *et al.*, 2011 Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. J Dairy Sci **94**: 3679–3686.

Dassonneville, R., S. Fritz, V. Ducrocq, and D. Boichard, 2012 Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. Journal of dairy science **95**: 4136–40.

de Roos, a. P. W., B. J. Hayes, and M. E. Goddard, 2009 Reliability of genomic predictions across multiple populations. Genetics **183**: 1545–1553.

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**: 1503–1512.

Dekkers, J. C. M., P. K. Mathur, and E. F. Knol, 2010 Genetic Improvement of the Pig. In M. F. Rothschild and A. Ruvinsky, editors, *The Genetics of the Pig*, chapter 16. CABI, Cambridge, MA, USA, 2nd edition, 390–425.

Devlin, B., and N. Risch, 1995 A COMPARISON OF LINKAGE DISEQUILIBRIUM MEASURES FOR FINE-SCALE MAPPING. Genomics **29**: 311–322.

Druet, T., I. M. Macleod, and B. J. Hayes, 2013 Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity .

Du, F., A. Clutter, and 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, *et al.*, 2010 A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC genetics **11**: 42.

Garrick, D. J., J. F. Taylor, and R. L. Fernando, 2009 Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol **41**: 55.

Goddard, M., 2008 Genomic selection: prediction of accuracy and maximisation of long term response. Genetica **136**: 245–257.

Goddard, M., B. Hayes, H. McPARTLAN, and A. J. Chamberlain, 2006 Can the same genetic markers be used in multiple breeds. In *Proceedings*. 16–22.

Goddard, M. E., and B. J. Hayes, 2007 Genomic selection. Journal of animal breeding and genetics = Zeitschrift f{ü}r Tierz{ü}chtung und Z{ü}chtungsbiologie **124**: 323–330.

Goddard, M. E., and B. J. Hayes, 2009 Genomic selection based on dense genotypes inferred from sparse genotypes. In *Proceedings*. 26–29.

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen, 2011 Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of animal breeding and genetics = Zeitschrift für Tierzüchtung und Züchtungsbiologie **128**: 409–21.

Groenen, M. M., P. Wahlberg, M. Foglio, H. H. Cheng, H.-J. Megens, *et al.*, 2009 A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res **19**: 510–519.

Gualdrón Duarte, J. L., R. O. Bates, C. W. Ernst, N. E. Raney, R. J. Cantet, *et al.*, 2013 Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC genetics **14**: 38.

Gualdron Duarte, J. L., R. O. Bates, C. W. Ernst, N. E. Raney, R. J. C. Cantet, *et al.*, 2012 Genotype imputation accuracy in an F2 pig cross using high density and low density SNP panels [abstract]. In *ADSA-ASAS Joint Annual Meeting*. Phoenix, W76.

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2009 Genomic Selection Using Low-Density Marker Panels. Genetics **182**: 343–353.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol **42**: 5.

Harmegnies, N., F. Farnir, F. Davin, N. Buys, M. Georges, *et al.*, 2006 Measuring the extent of linkage disequilibrium in commercial pig populations. Anim Genet **37**: 225–231.

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**: 433–443.

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. J. van der Werf, 2012 Accuracy of genotype imputation in sheep breeds. Anim Genet **43**: 72–80.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009b Increased accuracy of artificial selection by using the realized relationship matrix. Genetics research **91**: 47–60.

He, J., and A. Zelikovsky, 2006 MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression. Bioinformatics **22**: 2558–2561.

He, J., and A. Zelikovsky, 2007 Informative SNP selection methods based on SNP prediction. IEEE Trans Nanobioscience **6**: 60–67.

Henderson, C. R., 1984 *Applications of linear models in animal breeding.*. University of Guelph.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos, 2012 Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Science **52**: 654.

Hickey, J. M., and G. Gorjanc, 2012 Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. G3 (Bethesda, Md.) **2**: 425–7.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. F. Wilson, N. Dunstan, *et al.*, 2011 A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genetics selection evolution GSE **43**: 12.

Howie, B., J. Marchini, and M. Stephens, 2011 Genotype Imputation with Thousands of Genomes. G3 **1**: 13.

Howie, B. N., P. Donnelly, and J. Marchini, 2009 A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet **5**.

Huang, L., Y. Li, A. B. Singleton, J. A. Hardy, G. Abecasis, *et al.*, 2009 Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet **84**: 235–250.

Huang, Y., R. O. Bates, C. W. Ernst, J. S. Fix, and J. P. Steibel, 2013 Estimation of U.S. Yorkshire breed composition using genomic data. in preparation .

Huang, Y., J. M. Hickey, M. A. Cleveland, and C. Maltecca, 2012 Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genetics, selection, evolution : GSE **44**: 25.

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.

Jafarikia, M., L. Maignel, S. Wyss, and B. Sullivan, 2010 Linkage Disequilibrium in Canadian Swine Breeds. In *Proceedings*.

Johnston, J., G. Kistemaker, and P. G. Sullivan, 2011 Comparison of different imputation methods. In *Proceedings*.

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen, *et al.*, 2009 The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. Genetics **183**: 1119–1126.

Macleod, I. M., D. M. Larkin, H. A. Lewin, B. J. Hayes, and M. E. Goddard, 2013 Inferring Demography from Runs of Homozygosity in Whole Genome Sequence, with Correction for Sequence Errors. Molecular biology and evolution .

Marchini, J., D. Cutler, N. Patterson, M. Stephens, E. Eskin, *et al.*, 2006 A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet **78**: 437–450.

Marchini, J., and B. Howie, 2008 Comparing algorithms for genotype imputation. Am J Hum Genet **83**: 535–540.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, *et al.*, 2009 Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS One **4**: e5350.

Meuwissen, T., and M. Goddard, 2010 Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics **185**: 623–31.

Meuwissen, T., B. Hayes, and M. Goddard, 2013 Accelerating Improvement of Livestock with Genomic Selection. Annual Review of Animal Biosciences **1**: 221–237.

Meuwissen, T. H., and M. E. Goddard, 2001 Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol **33**: 605–634.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Moser, G., M. S. Khatkar, B. J. Hayes, and H. W. Raadsma, 2010 Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet Sel Evol **42**: 37.

Mrode, R., 2005 *Linear Models for the Prediction of Animal Breeding Values*. CABI Publishing, Oxfordshire, UK, 2 edition.

Nordborg, M., and S. Tavare, 2002 Linkage disequilibrium: what history has to tell us. Trends in Genetics **18**: 83–90.

Nsengimana, J., P. Baret, C. S. Haley, and P. M. Visscher, 2004 Linkage disequilibrium in the domesticated pig. Genetics **166**: 1395–1404.

Pérez, P., G. de los Campos, J. Crossa, and D. Gianola, 2010 Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. Plant Genome **3**: 106–116.

Pritchard, J. K., and P. Donnelly, 2001 Case-control studies of association in structured or admixed populations. Theor Popul Biol **60**: 227–237.

Qin, Z. S., S. Gopalakrishnan, and G. R. Abecasis, 2006 An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. Bioinformatics **22**: 220–225.

Ramos, A. M., R. P. M. A. Crooijmans, N. A. Affara, A. J. Amaral, A. L. Archibald, *et al.*, 2009 Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. PLoS One **4**: e6524.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411**: 199–204.

Sargolzaei, M., and F. S. Schenkel, 2009 QMSim: a large-scale genome simulator for livestock. Bioinformatics **25**: 680–681.

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer, 2008 Extent of linkage disequilibrium in Holstein cattle in North America. J Dairy Sci **91**: 2106–2117.

Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet **78**: 629–644.

Steibel, J., M. Wysocki, and J. Lunney, 2009 Assessment of the swine protein-annotated oligonucleotide microarray. Anim Genet .

Strandén, I., and D. J. Garrick, 2009 Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. Journal of dairy science **92**: 2971–5.

Sved, J. A., A. F. McRae, and P. M. Visscher, 2008 Divergence between human populations estimated from linkage disequilibrium. Am J Hum Genet **83**: 737–743.

Team, T. R. D. C., 2011 *R: A language and environment for statistical computing*. Vienna, Austria.

Tenesa, A., P. Navarro, B. J. Hayes, D. L. Duffy, G. M. Clarke, *et al.*, 2007 Recent human effective population size estimated from linkage disequilibrium. Genome Research **17**: 520–526.

Tribout, T., C. Larzul, and F. Phocas, 2012 Efficiency of genomic selection in a purebred pig male line. Journal of animal science **90**: 4164–76.

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**: 609–614.

USDA, 2009a Quaterly Hog and Pigs Report. United States Government Printing Office, Washington DC : 19.

USDA, 2009b Various Market News. United States Government Printing Services, Washington DC **Quaterly**: 60.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of dairy science **91**: 4414–23.

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel, 2011 Genomic evaluations with many more genotypes. Genet Sel Evol **43**: 10.

VanRaden, P. M., C. P. V. Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, *et al.*, 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls.

Vazquez, A. I., D. M. Bates, G. J. M. Rosa, D. Gianola, and K. A. Weigel, 2010 Technical note: an R package for fitting generalized linear mixed models in animal breeding. Journal of animal science **88**: 497–504.

Villa-Angulo, R., L. K. Matukumalli, C. A. Gill, J. Choi, C. P. V. Tassell, *et al.*, 2009 High-resolution haplotype block structure in the cattle genome. BMC Genet **10**: 19.

Vittinghoff, E., S. C. Shiboski, D. V. Glidden, and C. E. McGulloch, 2005 *Regression Methods in Biostatistics : Linear , Logistic , Survival , and Repeated Measures Models*. Springer, New York, NY.

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

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, *et al.*, 2010a Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J Dairy Sci **93**: 5423–5435.

Weigel, K. A., C. P. V. Tassell, J. R. O'Connell, P. M. VanRaden, and G. R. Wiggans, 2010b Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J Dairy Sci **93**: 2229–2238.

Welsh, C. S., T. S. Stewart, C. Schwab, and H. D. Blackburn, 2010 Pedigree analysis of 5 swine breeds in the United States and the implications for genetic conservation. J Anim Sci **88**: 1610–1618.

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.

Wiggans, G. R., T. A. Cooper, P. M. VanRaden, K. M. Olson, and M. E. Tooker, 2012 Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. Journal of Dairy Science **95**: 1552–1558.

Wiggans, G. R., P. M. Vanraden, and T. A. Cooper, 2011 The genomic evaluation system in the United States: past, present, future. Journal of Dairy Science **94**: 3202–3211.

Wolc, A., C. Stricker, J. Arango, P. Settar, J. E. Fulton, *et al.*, 2011 Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Genet Sel Evol **43**: 5.

Zhang, Z., and T. Druet, 2010 Marker imputation with low-density marker panels in Dutch Holstein cattle. J Dairy Sci **93**: 5487–5494.

Zheng, J., Y. Li, G. R. Abecasis, and P. Scheet, 2011 A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genet Epidemiol **35**: 102–110.