Extent of Linkage Disequilibrium, Consistency of Gametic Phase and Imputation Accuracy Within and Across Canadian Dairy Breeds

by

Steven Larmer

A Thesis
presented to
The University of Guelph

In partial fulfillment of requirements
for the degree of
Master of Science
in
Animal Breeding and Genetics

Guelph, Ontario, Canada

ABSTRACT

EXTENT OF LINKAGE DISEQUILIBRIUM, CONSISTENCY OF GAMETIC PHASE AND IMPUTATION ACCURACY WITHIN AND ACROSS CANADIAN DAIRY BREEDS

Steven G. Larmer

University of Guelph, 2012

Advisor:

F.S. Schenkel

Some dairy breeds have too few animals genotyped for within breed genomic selection to be carried out with sufficient accuracy. As such, the level of linkage disequilibrium within each breed as well as consistency of gametic phase across breeds was studied. High correlations of phase (>0.9) were found between all breed pairs at this same SNP density. The efficacy of imputing animals genotyped on lower density (6k and 50k) panels was then explored in order to increase the size of the reference population with 777k genotypes in a cost-effective manner. These results showed high accuracies (>0.92) in all imputation scenarios studies, using both a within breed and a multi-breed reference population for imputation. It was concluded that given the results of both of these studies, pooling breeds into a common reference population for genomic selection should be a viable option for accurate genomic selection in breeds with few genotyped individuals.

ACKNOWLEDGEMENTS

I would like to begin by expressing my gratitude for my advisor, Dr. Flavio Schenkel. The guidance I receieved throughout the process of my research and writing was invaluable in both the quality of this thesis and the knowledge and ability I have accrued over the past two years. I would also like to thank the other members of my committee, Dr. Mehdi Sargolzaei, Dr. Steve Miller and Dr. Filippo Miglior for their time and their insight into this work. In particular I would like to thank Dr. Sargolzaei for his continued technical help and support. Furthermore, I would like to thank Dr. Andy Robinson for serving on my examination committee. I would also like to thank Dr. Ricardo Ventura for his continued help, support and encouragement.

I would like to extend my gratitude to my friends and office-mates that have made the last two years so enjoyable. I have been lucky to have a phenomenal network of friends all across the University of Guelph research community who have been extremely helpful.

I could not have gotten to this point without the love and support of my family, who have always been uncompromisingly supportive, I cannot thank you enough for always being there to listen and help.

I would also like to acknowledge the Natural Science and Engineering Council (NSERC) of Canada, The Canadian Dairy Network, Ayrshire Canada, Guernsey Canada, Semex and L'Alliance Boviteq inc. for the financial support that made this project possible. I'd also like to thank the USDA for the use of genotypes that also allowed this research to be possible.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

LITERATURE REVIEW

1.1. Introduction

The use of genomic selection has revolutionized the way Holstein cattle are selected and cows are bred in Canada and around the world in recent years. To be able to implement similar breeding strategies in breeds other than Holsteins we must find a way to grow the reference population in an effort to improve reliability of genomic selection in these breeds to approach the level it has reached in Holstein cattle. Reliability has been shown to increase linearly with an increase in reference bulls up to approximately 3,500 bulls (VanRaden et al., 2009). In order to increase the number of bulls in the reference population in small population size breeds, one must look at the links between breeds in terms of linkage disequilibrium and gametic phase to determine if breeds can be pooled into a common reference population for selection.

Assessing linkage disequilibrium (LD) is important for accuracy of genomic selection. Determining the extent of LD is crucial in determining appropriate marker density, along with the size of the reference population (Sargolzaei et al., 2008). Determining the span of LD at given marker distances will help determine the minimum distance between markers to effectively cover the entire genome. By uncovering the LD between markers that are common across breeds we can determine if there is a sufficient relationship to be able to use a pooled reference population.

Determining consistency of linkage phase across breeds will allow for further analysis of the potential to use multi-breed programs for genomic selection. If markers

are not in the same phase across two breeds, the ability to use one breed to determine effects of SNP to aid in selection of another becomes less possible. Determining the consistency of phase between all breeds will determine if any breed, even those with already larger reference populations, will benefit from a joint genomic evaluation. This becomes most important in the less numerous breeds, such as the Canadian Guernsey population, as there is no effective method to implement genomic selection with the current reference population size. Some breeds may have more success with this method as they are closer in relation due to a more recent divergence; or, in some cases, the continued use of cross-breeding as is the case with the use of Swedish red sires in the Canadian Ayrshire population. Due to the use of Swedish reds within the Ayrshire population, there is a good chance that these breeds can be used in a joint genomic evaluation, benefitting both breeds by expanding the overall reference population size.

As marker density increases, the extent of LD has also been shown to increase (Farnir et al., 2000). The use of higher density genotype panels allows for this to be exploited, and potentially may allow for more accurate selection to take place. These higher density genotype panels, however, are significantly more expensive per animal, and so one must look at strategies to impute from lower density genotype panels to a higher density in order to effectively grow the reference population of high density genotypes.

Herein, the literature on consistency of linkage phase as well as determining LD over distance and across populations will be reviewed in an effort to better understand the best possible methods for determining the prospect of using genomic selection across breeds to overcome the shortfalls of having smaller reference populations. The literature

on imputation of genotypes is also reviewed to determine the efficacy of imputing from low to high density genotyping platforms.

1.2. Linkage Disequilibrium

The level of Linkage Disequilibrium is critical in the success of fine-scale mapping. Assessing the level of linkage disequilibrium allows for a characterization of the relationship between SNPs present on marker panels and the Quantitative Trait Loci (QTL) that they are in disequilibrium with. For genomic selection as a whole to work, there must be a significant population-wide disequilibrium between the markers and the QTL such that the markers will predict the effects of the QTL across the entire population (Hayes et al., 2009b). This allows for selection on traits normally selected using a progeny test scheme by Marker Assisted Selection (MAS), using SNP to create a relatively accurate proof for young bulls (Schaeffer, 2006). Schaeffer proposed that this system could save the genetics industry millions of dollars in the cost of genetic change, because of the cost of the process of raising and proving bulls in conventional selection programs.

The level of LD at a given point in the genome is directly correlated to the distance between markers. This LD at given distances can then be exploited to determine QTL effects. (Sargolzaei et al., 2008). Significant linkage was found in the same study by Sargolzaei et al. (2008) to extend as the length between markers increased to a certain point. This is significant as common marker panels that are economically viable to test linkage in large populations are at this point limited in size to approximately 50,000 SNP across the entire genome. As the efficiency of genotyping improves with the advent of

larger panels that are low enough in cost to implement on a large scale, the accuracy of selection will increase and may allow for selection across certain populations. In a study conducted in Australian cattle populations it was found that a panel of over 300,000 SNPs would be needed to compute proofs across breeds using Holstein, Jersey and Angus cattle with accuracy greater than 0.85 (de Roos et al., 2008).

There have been few QTL found that have large effects on economic traits of interest. For this reason, a panel using dense markers spaced across the genome capturing the effects of many QTL with small relative effects has been shown to be more effective than attempting to capture only QTL with larger effects (Meuwissen et al., 2001). There are, however, exceptions to this, as the *DGAT1* gene has been shown to have a very significant effect on milk production, as well as milk fat and protein percentages (Weller et al., 2003). There are other genes that have relatively large effects on certain traits across the genome, and as such, there have been numerous models created to weigh both the genes of moderate to large impact with the many, less significant QTL, to create an accurate genomic breeding value (Meuwissen et al., 2001; Roche et al., 2007; VanRaden, 2009; Zhang et al., 2007) . The genome has also been shown to have hotspots of recombination, where recombination events take place more often (Li and Stephens, 2003). This is useful information as recombination leads to a change in the LD pattern over time. This change over time must be taken into account in order to maintain accurate predictions of LD. Hotspots are also useful in determining marker density as a greater recombination rate benefits from a higher marker density as there is a lesser chance of LD being broken down by a recombination event when the marker and QTL are close together (Rafalski and Morgante, 2004). It must be further studied to determine if these

hotspots are consistent across breeds or if there is significant variation in areas of recombination across breeds as well.

It has been shown that linkage disequilibrium decays as markers that are further and further apart are examined (Sebastiani and Abad-Grau, 2007). Studies have shown that there is a significant difference in LD decay across breeds of sheep, with breeds that have a greater within breed genetic diversity showing a larger decay of LD as marker distance increases (Meadows et al., 2008). This could have a positive impact on the less diverse dairy cattle breeds with smaller effective population size such as the Guernsey and Jersey populations. If LD extends over large distances in these breeds, the problems related to having a small reference population could be somewhat overcome by having greater levels of LD over larger distances.

To determine relatedness among breeds for viability of across breed genomic selection, the linkage disequilibrium within each breed must first be determined so that linkage phase can be compared across breeds. The original measure used to determine the extent of LD between two markers is heavily biased by allele frequency. This is the D value of LD. This is calculated as $D = f(AB) - f(A)f(B)$ (Du et al., 2007), where f(AB) is the frequency of haplotype AB and f(A) and f(B) are the allele frequencies of allele A at locus 1 and allele B at locus 2, respectively. To lessen the bias due to allele frequency, more accurate methods have been developed such as D' and r^2, the two most common measures of LD used currently. Both of these measures contrast actual and expected haplotype frequencies to determine the degree to which markers are linked. Both of these methods depend on allele frequency to a certain degree, although r^2 has

been shown to be less biased due to small minor allele frequencies. D' is defined as

$$D' = \sum_{i=1}^{2} \sum_{j=1}^{2} p_i q_j \frac{|D_{ij}|}{D_{max}}$$

Where $D_{max} = \begin{cases} [\min[p_i q_j, (1-p_i)(1-q_j)] \; if \; D_{ij} < 0] \\ [\min[p_i(1-q_j), (1-p_i)q_j] \; if \; D_{ij} \geq 0] \end{cases}$

(Du et al., 2007). If D' is equal to 1, an allele at one locus is in total disequilibrium with an allele at another locus. This will remain the case until a recombination or mutation event takes place between these two markers. The benefit of using D' is that it is a flexible measure in terms of the numbers of alleles that can be used to calculate LD at once. This makes it more viable in studies such as the one done by Farnir et al. (2000), to assess population wide LD using microsatellite markers. Many studies using microsatellites use D' as the primary measure of linkage (Mohlke et al., 2001; Farnir et al., 2000). However, it has also been used in studies where SNPs are used as the primary source of data for determining LD (Sutter et al., 2004). The major drawback of using D' is that it is still somewhat affected by allele frequencies, especially in cases with small sample sizes or low allele frequencies (Du et al., 2007).

r^2 was created to further eliminate bias from certain allele frequencies. It is calculated as $r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$, where $D = f(AB) - f(A)f(B)$.

f(AB),f(A),f(B),f(a), and f(b) are observed frequencies of haplotype AB and alleles A, B, a, and b respectively (Hill and Robertson, 1968). r^2 has been shown to be the least dependent on allele frequency for bi-allelic markers such as SNP (Ardlie et al., 2002), although is not a perfect measure in that regard. $r^2 = 1$ can only occur if all alleles at one

locus are all in perfect disequilibrium with an allele at the other locus, and all allele frequencies are equal. Thus, there are only two possible haplotypes between the 2 markers being examined, and one allele can be perfectly predicted from knowing the other (Laird and Lange, 2011). Hill and Robertson's (1968) measure of r^2 gives a more accurate assessment of the level of LD than D' when attempting to determine one polymorphism given the other, as is the case in marker assisted selection (Chen et al., 2006). The main problem with all measures of LD is still the dependence on allele frequency. Even r^2 carries some bias due to allele frequency. Even with high levels of LD, D values of near zero will be found if minor allele frequency is very low, this can still create a downward bias in the measure of r^2 and an upward bias in D' (Laird and Lange, 2011).

Linkage Disequilibrium has a large effect on the accuracy of genomic estimated breeding values. The greater the level of LD, the more accurately markers can be used to predict the effects of QTL, as there is greater certainty that the correct allele is being predicted at the QTL locus given the allele present at the marker locus. Brito et al. (2011) explored the level of accuracy of direct genomic estimated breeding values (DGEBVs) in a simulated population with the approximate level of LD found in beef cattle populations using ~50,000 markers. Beef cattle have been found to have a lower level of LD at the same distance, when compared to dairy cattle (McKay et al, 2007). With lower levels of LD and larger within-breed diversity in beef than is found in dairy cattle populations, Brito et al. (2011) concluded that a denser marker panel would likely be needed for genomic selection to be possible in beef cattle.

1.3. Linkage Phase

Linkage phase refers to the arrangement of alleles on two homologous chromosomes. If linked alleles are on the same chromosome, they are said to be in coupling phase. Adversely, if the linked alleles are found on different chromosomes, they are said to be linked in repulsion phase. This is valuable for selection programs to determine which allele is to be selected for at the marker locus to have the desired effect on the QTL. Linkage phase can differ from family to family and even more so from population to population (Dekkers and Hospital, 2002). Linkage phase can be determined biologically, by looking at the linkage phase of parents and grandparents of progeny. If parental genotypes are not available, phase can sometimes be statistically determined using a likelihood function as described by Liu (2011). Linkage phase is also determined by using phenotypes on traits that have a single marker tied to them. This is commonly done to use genomics to determine likelihood of disease (Axenovich, 1996). It is performed by examining parental phenotypes across populations and the marker status of both parents and progeny. This is not overly useful in current genomic selection techniques due to the fact that genomic selection assumes that there are many QTL with small effects contributing to phenotypes (Calus, 2010).

Linkage Phase is most effective when it is used to compare populations to determine relatedness and to determine the extent to which markers are consistent in phase across these populations. In general, when looking across populations, the level of useful linkage disequilibrium, that is, areas where the LD is in the same phase, is much lower than when looking at members of the same population. It has also been shown that breeds that are known to be more related based on descent are more likely to be more in

more consistent phase (Maudet et al., 2002). Persistence of phase across populations is also extremely useful in determining the marker density to perform accurate multi-breed genomic selection. Goddard et al. (2006) and Gautier et al. (2007) both reported that for selection to be made across breeds with high accuracy, spacing between markers would need to be no greater than 10kb between adjacent markers. This was previously not available, however, with the advent of the 777k SNP chip, this has become a reality and could allow for genomic selection to be put in place for breeds that originally were thought to have reference populations of insufficient size for genomic selection. Persistency of phase has also been used to estimate the time since breeds diverged as there has been significant evidence showing that phylogenetic similarities correlate with time since breed divergence (de Roos et al., 2008; S. McKay et al., 2008; Nagamine et al., 2008).

Many studies have indicated the value of having a larger number of SNPs to cover the genome for fine mapping, especially when looking across populations (de Roos et al., 2008; Goddard et al., 2006; Sargolzaei et al., 2008). de Roos et al. (2008) concluded that to select across breeds that have diverged a significant time ago, as all cattle breeds have, and still find markers that are in LD with QTL across breeds, approximately 300,000 markers would be needed. Although there has been a significant effect of selection within cattle breeds, there is still a great degree of genetic diversity (The Bovine HapMap Consortium, 2009). It has also been shown that LD levels are low at distances greater than 1 Mb considering the effective population sizes of most breeds, and thus the effective population sizes must have been large recently, and have shrunk significantly due to the advent of artificial selection (The Bovine HapMap Consortium, 2009). It has

also been shown that breeds that have had less selection pressure have more diverse genetic structures even when their effective population size is smaller (Maudet et al., 2002). This indicates that selection has been the major contributor for the lack of diversity among breeds. When looking across breeds this becomes a larger issue as selection pressures have been weighted differently across breeds and may have caused greater breed divergence.

The genetic diversity of breeds has been directly linked to their areas of origin, indicating that breeds that diverged more recently were generally closer together geographically. There is also a demonstrated larger difference between taurine and indicine breeds due to a much greater time since divergence (McKay et al., 2007). It has also been shown that there is a significant difference between beef and dairy breeds when contrasted to breeds within dairy or beef. This may be due to divergent selection pressures across these groups (Hayes et al., 2009a; McKay et al., 2007). Microsatellites have been used to assess the genetic variation among many breeds relative to their areas of origin. The results of that study are summarized in this figure 1.1 by (Blott et al., 1998)

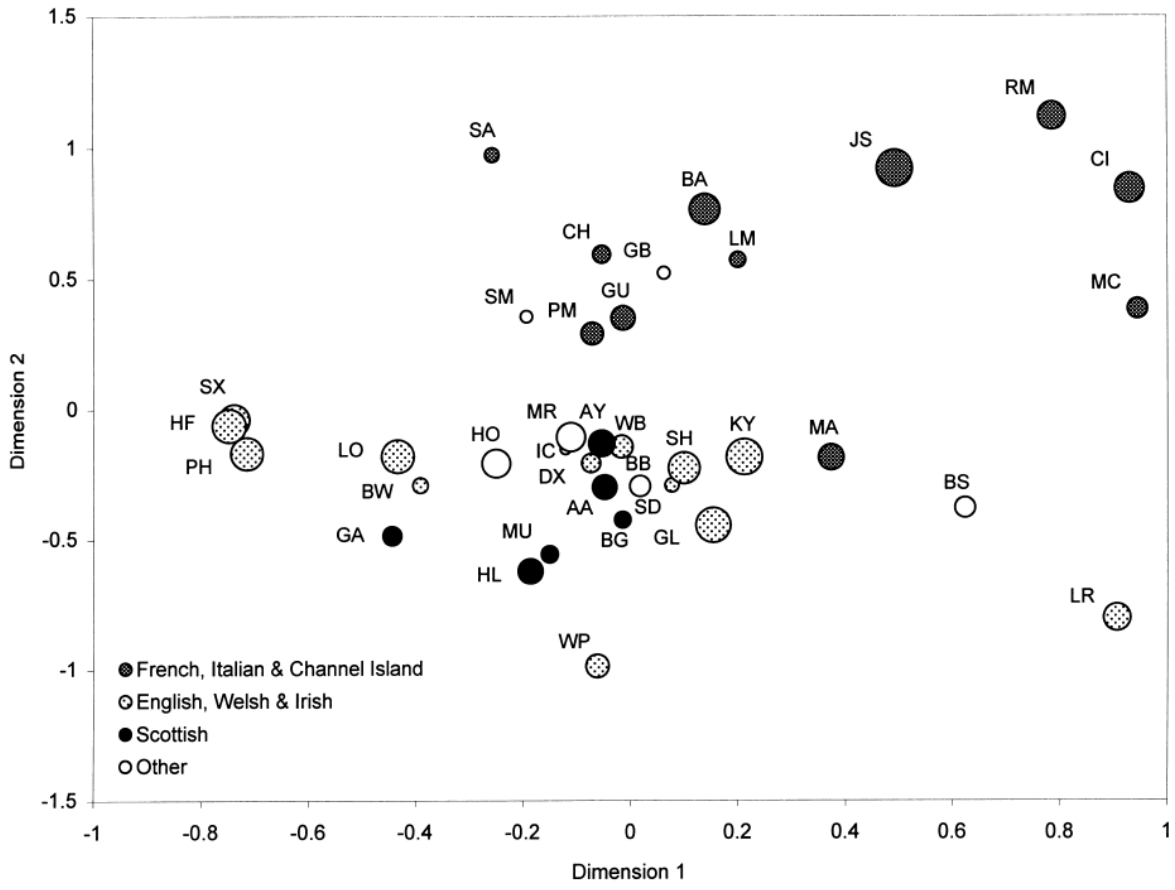

Figure 1.1. – From Blott et al. (1998) – Genetic variations among beef and dairy cattle breeds.

As is shown, there is a pronounced correlation between the area in which the breed originated and the genetic similarities between breeds. The study went further to look at the direct similarities between certain breeds and a tree was developed to show the closest related breeds. This information will be invaluable in determining which breeds are relevant to each other for pooling selection. This can be seen in figure 1.2 from Blott et al., (1998).

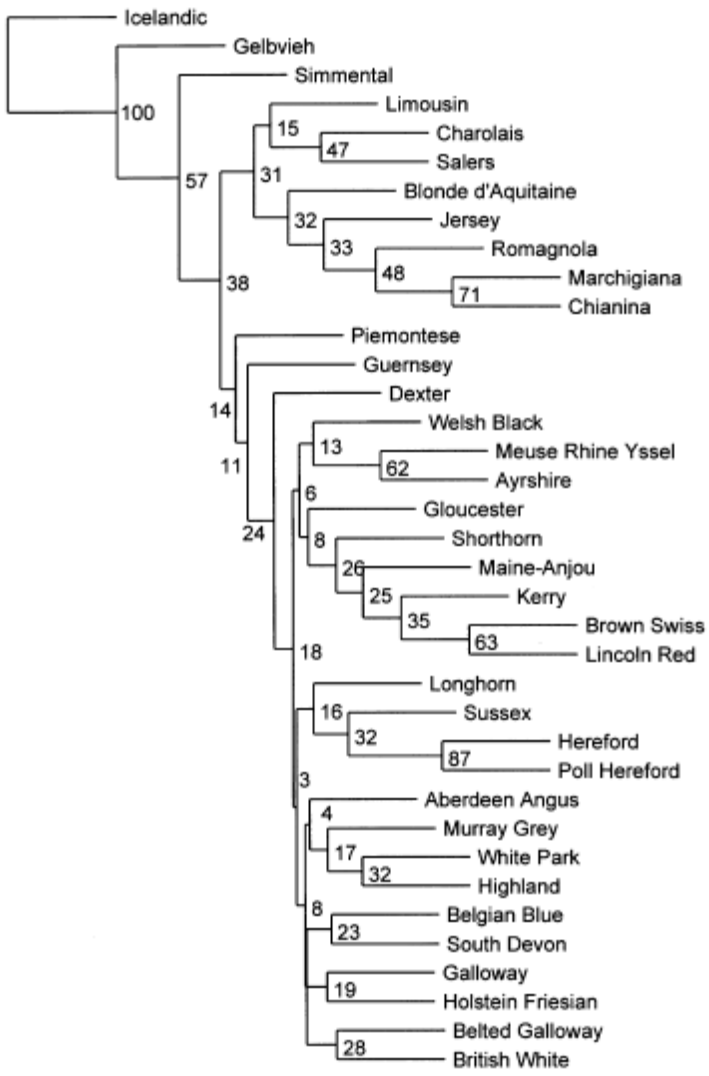

Figure 1.2. – From Blott et al. (1998). Phylogenetic tree of beef and dairy breeds.

Using this type of information, coupled with the observed consistencies in linkage phase, one should be able to determine which breeds can be pooled for selection in an effort to apply genomic techniques for breeds with fewer genotyped animals.

Implementation of genomic selection across breeds is made much more possible by the use of high density genotyping. It was found by Rolf et al. (2010) that when the high density panel was used on chromosome 29, one SNP could be found that was

significantly associated with Warner-Bratzler shear force, an important measure of beef tenderness. That same SNP was found to be significantly associated with beef tenderness across breeds. Without the use of the high density panel, there is not enough consistency across breeds to find such associations. Methods such as the haplotype model may also be implemented in the future to create greater associations with QTL (Boichard et al., 2012) in an effort to find associations that will persist across breeds. The problem with this model is that it can create an even larger number of parameters to be estimated than the already larger number required when using a higher density SNP panel. With a small effective population size as is observed in dairy breeds, this may not be the case, however, because of a small number of relative haplotypes present at certain loci due to a high level of LD. These problems can be aided by the use of data mining methodologies such as that described by Shah and Kusiak (2004) to select only the SNPs in closest association to a trait of interest to limit the number of parameters that need to be estimated.

1.4. Imputation

For effective genomic selection, a large reference population is needed for effective estimation of SNP effects (Goddard, 2009). It has also been found that an increase in relationship between the reference population and evaluated population leads to higher accuracy of SNP effect estimation (Pszczola et al., 2012). Having a large number of animals genotyped on panels with a high enough density for SNP effect estimation can be quite expensive. To reduce costs of having a large genotyped reference population, animals can be genotyped on a select smaller number of SNPs and the remainder of the genotypes are predicted by using information from other animals

genotyped on the high density panel (Druet et al., 2010). This process is called imputation and exploits the linkage between markers on a given haplotype to infer the markers that would be contained between them on a higher density panel. As such, for effective imputation to be carried out, haplotypes must be reconstructed from genotype data (Hickey et al., 2011).

The two primary categories of imputation algorithms are population-based and family-based. Both of these methods are commonly used and a combination of the two is the current method for cattle genotype imputation in Canada. For a population with a large reference set and high degree of relatedness, imputation from ~3k to ~50k has been shown to be very accurate using both family-based and population-based methods with accuracy exceeding 0.90 for population-based methods, such as IMPUTE (Druet et al., 2010; Weigel et al., 2010; Nothnagel et al., 2009), Beagle (Druet et al., 2010; Nothnagel et al., 2009; Calus et al., 2011), DAGPHASE (Zhang et al., 2010) and fastPhase (Weigel et al., 2010, Nothnagel et al., 2009, Calus et al., 2011) as well as family based methods CHROMIBD (Zhang et al., 2010), FImpute (Sargolzaei et al., 2010, Johnston et al., 2011), Chromophase (Daetwyler et al., 2011) and findhap (Johnston et al., 2011). It was found that when family information was available, family-based methods do have an advantage over population-based methods (Zhang et al., 2010). Sargolzaei et al. (2010) found that an increase in family size leads to significantly higher imputation accuracy. When imputing in animals with at least 5 related animals in the reference population from 3k to 50k, greater than 97% of markers were correctly called.

In Canadian dairy populations, imputation is currently carried out from 3k and 6k to 50k and in the past was carried out solely from 3k to 50k for routine calculations of

genomic estimated breeding values as well as genomic enhanced parent averages by the Canadian Dairy Network. This imputation process is carried out using the family and population based imputation algorithm of FImpute (Sargolzaei et al, 2010) using the family-based component of FImpute combined with the population imputation step performed by Beagle (Browning and Browning, 2007). Johnston et al. (2011) found that this combination imputation method yielded imputation accuracy of greater than 0.95 from 3k to 50k in Holstein and Brown Swiss populations.

The 6k panel can be imputed at higher accuracies than the 3k when similar reference population sizes are used due to a greater marker density as well as using the same chemistry to determine genotype calls as the 50k panel, which the 3k panel does not (Wiggans et al., 2012). Sargolzaei et al. (2010) determined that imputation accuracy increases as the density of markers increases on the panel that is to be imputed.

Imputation to a higher density chip such as the Illumina 777k bovine Beadchip has not been regularly carried out. Due to the higher levels of LD found at higher marker densities (Farnir et al., 2000), haplotypes on the 50k panel should be more consistent with the markers that are to be imputed for the higher density panel found within each haplotype. This should lead to a high degree of accuracy when imputing to higher density from 50k. Van Raden et al. (2011) simulated 500,000 markers to be imputed from the 50k population of genotyped Holsteins in the USA. Accuracies of imputation were found to exceed 95% for this imputation step when using a combination of family and population based imputation methods.

Imputed genotypes are then used for genomic evaluation. Wiggans et al. (2012) found that when imputed genotypes from the 3k panel were used for generation of Genomic Predicted Transmitting Ability (GPTA), there was very little difference in SNP effect estimation accuracy. It was also found that GPTAs from imputed and non-imputed 50k genotypes had a correlation greater than 0.95 within the 3 different breeds studied (Holstein, Jersey and Brown Swiss).

One of the largest gains that can be made from imputation of genotypes is imputing genotypes for females in order to increase accuracy of selection (in females). McHugh et al (2011) found that when female information is included in genomic selection schemes, an overall increase of genetic gain of up to 3 times can be seen. This is due to a decrease in generation interval by selecting females earlier as well as an increase in accuracy of selecting the most desirable females for a number of traits.

CHAPTER 2

EXTENT OF LINKAGE DISEQUILIBRIUM AND CONSISTENCY OF PHASE IN FIVE CANADIAN DAIRY BREEDS

2.1. Abstract

Implementation of genomic selection requires a large reference population to accurately estimate SNP effects. In some Canadian dairy breeds, large enough reference populations are not available to estimate SNP effects accurately for the traits of interest. If marker phase is highly correlated across multiple breeds, it may be possible to pool several breeds into one common reference population. This study investigates the amount of linkage disequilibrium (LD) in 5 major dairy breeds using the 50k SNP panel and 3 of the same breeds using the 777k SNP panel. We investigated both the extent of LD at varying distances as well as correlation of pair-wise SNP phase. SNPs were filtered for a minor allele frequency less than 5%. The level of LD was measured using the squared correlation of alleles at 2 loci (r^2), and consistency of SNP phase was correlated using the signed square root of these values. Analysis showed that LD values greater than 0.2 are found in all breeds at distances at or shorter than average pair-wise distances using the 50k panel. Correlations of r values, however, did not reach high levels (<0.9) at these distances. High correlation values of SNP phase between breeds was observed (>0.94) when the average pair-wise distances using the 777k SNP panel were examined. These findings suggest that a multi-breed reference population for genomic selection might be possible using the 777k SNP panel in Canadian dairy breeds.

2.2. Introduction

The advent of genomic selection has been a major breakthrough in many dairy breeds, and continues to improve with the advent of new technologies to improve the accuracy of selection and the reliability of the methods used. To begin genomic selection within a breed or group of breeds, an important step is assessing the level of linkage disequilibrium (LD) (Meuwissen et al., 2001). LD is a measure of the non-random association of alleles that helps us to infer the alleles present at other loci, especially at quantitative trait loci (QTL) that have an effect on phenotypes of interest. Numerous studies have found high levels of LD between adjacent marker pairs as well as LD extending over tens of centimorgan (Farnir et al., 2000,). However, useful LD in Holsteins was only found at distances shorter than 100 kb (Sargolzaei et al., 2008). Useful LD was defined as an r^2 value of greater than 0.3. This LD value was defined based on enough LD both for association studies as well as effective genomic selection. Simulations have shown, however, that accuracy of genomic selection can be as high as 0.85 when the level of LD, based on r^2, is above 0.2 (Meuwissen et al., 2001).

The two factors that affect accuracy of genomic selection that can be controlled are the level of LD between markers and QTL, and the size of the reference population of animals used to estimate SNP effects on phenotypes (Hayes et al., 2009c). The level of LD between markers and QTL can be controlled by utilizing very dense marker panels. Genomic selection, as it has been implemented in the Holstein, Jersey and Brown Swiss populations utilizes the 50k panel. With the advent of new SNP panels, with up to 777,000 markers, the accuracy of genomic selection can be increased. Calus et al. (2008) found that if the r^2 measure is increased from 0.1 to 0.2, accuracy of genomic selection

increased from 0.68 to 0.82 using simulated data. As r^2 values higher than 0.2 can be obtained using the 50k panel (Sargolzaei et al., 2008), the expected gain in accuracy of genomic selection will not be as large when increasing the r^2 measure further using the 777k panel when the reference population size is large. Brito et al (2011) reported that for similar accuracy of genomic selection, approximately half as many animals would be needed in the training population when high density panels were used. Van Raden et al. (2011), using simulation on an existent pedigree, reported that for an already large population (33,414 Holsteins with good pedigree information), an increase in markers from 50,000 to 500,000 yielded a gain in GEBV accuracy of 1.6%. This increase in accuracy is relatively small due to enough linkage to effectively implement genomic selection being present at the marker density found on the 50k panel.

Once the extent of linkage disequilibrium has been maximized, one must then look at growing the reference population in order to increase accuracy. Luan et al (2009) showed an increase in accuracy of ~5% for genomic selection of production traits in Norwegian Red cattle when the reference population grew from 250 to 400 animals. Liu et al. (2011) also saw an increase in the variance of SNP effects as the reference population size grew. A larger variance in SNP effects will allow for more genetic progress to be made per generation as individuals with SNPs of larger effect can be selected upon. One way to increase the reference population size for selection would be to use a breed with more genotyped individuals in the reference population for a breed with fewer candidates. It was found that for Holsteins and Jerseys, one breed could not be directly used as the sole reference population for selection in the other breed using the 50k marker panel, but results showed more promise when both breeds were combined

into a common population (Pryce et al., 2011). For genomic selection to be effective, markers and QTL need to be in the same linkage phase across the populations, and must be consistent from the reference population to the validation population. Gautier et al. (2007) as well as Goddard (2006) both showed that breed phases did not correlate past 10kb. A more practical solution is to combine more than one breed in the reference population in an attempt to grow the population while still capturing the linkage phase present for the breed in which genomic selection is being implemented. This was the ultimate goal of the study by De Roos et al. (2008). For this strategy to work, breeds must still be consistent in phase in order to be pooled into a common reference population. De Roos et al. (2008) estimated that for breeds to be pooled into a common reference population, approximately 300,000 evenly spaced markers would be needed to ensure adequate consistency of phase across breeds in the multi-breed reference population for selection.

The goal of this study was to explore the extent of linkage disequilibrium in 5 Canadian dairy breeds using the 50k SNP panel, as well as 3 breeds using the 777k SNP panel. Consistency of phase was then explored to determine the efficacy of a multi-breed reference population for genomic selection in Ayrshires and Guernseys using both the 50k and 777k SNP panels.

2.3. Materials and Methods

2.3.1. Data

Genotypes from Holstein (n=47,433) Jersey (n=4,517) and Brown Swiss (n=1,566) were taken on proven animals from the North American Collaboration on Genomic Prediction. These genotypes were performed using the Illumina 50k Beadchip (Illumina Inc., San Diego, USA). In addition to this, 351 proven Ayrshire bulls and 60 proven Guernsey bulls were genotyped with the 777k panel from Illumina bovine (Illumina Inc., San Diego, USA). Holstein 777k genotypes (n=1,115) were also obtained from The North American Collaboration on Genomic Prediction.

Due to small population sizes and a small degree of relatedness between animals in the genotyped populations, haplotypes could not be accurately reconstructed to determine the extent of LD in all breeds except for Holsteins genotyped on the 50k panel. Haplotypes for the Holsteins genotyped using the 777k panel were also not reconstructed due to a lack of pedigree information. Holstein genotypes came from a variety of sources and countries, and as such, animal IDs were not consistent with registration numbers and some individuals had no records of parentage whatsoever. To be able to make consistent comparisons across breeds, haplotypes were not reconstructed for Holsteins on the 50k panel either.

SNPs were filtered for minor allele frequencies less than 0.05 to examine only SNP that are segregating in a significant number of animals in the population. SNPs were also excluded if the missing SNP call for that SNP locus was greater than 10%.

The 777k SNP panel from the Ayrshire and Guernsey breeds was filtered to find only the SNPs present in the 50k panel to determine extent of LD at consistent locations with the 50k panel, in order to determine phase consistency between all 5 breeds using the 50k panel. This resulted in 39,127 (out of 43,382) SNPs being considered that were consistent between the 50k and 777k panels.

2.3.2. Determining Extent of Linkage Disequilibrium

Linkage Disequilibrium was determined using r^2 which is the squared correlation of alleles at 2 loci. This was calculated for each pair of loci on each chromosome to determine the LD at close distances as well as the LD decay over distance. r^2 was determined both for the 50k panel for all 5 breeds, as well as the 777k panel for Holsteins, Ayrshires and Guernseys. For computational ease, r^2 was only determined at distances less than 0.5 Mb with the 777k panel, as the 50k panel gives an accurate measure of LD decay beyond this distance. r^2 was calculated as follows:

$$r^2 = \frac{D^2}{f(A)f(a)f(B)f(b)}$$

Where, $D = f(AB) - f(A)f(B)$, and $f(AB)$, $f(A), f(a), f(B), and f(b)$ are observed frequencies of haplotype AB and alleles A, a, B and b, respectively. When haplotypes are not reconstructed, an unbiased estimator of D under Hardy-Weinberg equilibrium is

$$D = \frac{N}{N-1}\left[\frac{4N_{AABB} + 2\left(N_{AABb} + N_{AaBB}\right) + N_{AaBb}}{2N} - f(A)\times f(B)\right]$$ (Lynch and Walsh, 1998),

where N is the total number of individuals, and $N_{AABB,}$ $N_{AABb,}$ $N_{AaBB,}$ and N_{AaBb} are the corresponding number of individuals in each genotypic category (AABB, AABb, AaBB,

and AaBb). Not reconstructing haplotypes may lead to downward bias in the level of LD as marker pairs not in Hardy-Weinberg equilibrium will generally lead to higher absolute D values. Therefore, the measure of LD in this study is expected to be conservative, if any bias exists. The r^2 measure of LD was used because it is less biased by sample size and low allele frequency compared to other popular measures of LD such as D' (Ardlie et al., 2002), although no LD measure is completely allele frequency independent. Data was sorted into groups based on pair-wise marker distance in order to determine the breakdown of LD in each breed as length between markers increased.

2.3.3. Determining Consistency of Phase

For each marker pair with a measure of r^2, the signed r value was determined by taking the square root of the r^2 value and assigning the appropriate sign based on the calculated D value. Data was sorted into groups based on pair-wise marker distance to determine the breakdown in correlation across distance, as well as to be able to assess the correlation of signed r values at the smallest distances possible given the number of SNPs that were investigated. The r values were then correlated between breeds using the PROC CORR procedure in SAS (SAS Institute Inc., Cary, USA). This allows for the measurement of the breakdown in correlation of breed LD as distance between marker pairs increases. Correlations were performed between all pairs of breeds available on a certain SNP panel. This gave us 10 independent breed pairs on the 50k panel and 3 pairs on the 777k panel, those being Ayrshire-Guernsey, Ayrshire-Holstein and Holstein-Guernsey.

2.3.4. Effective Population Size over Time

Linkage Disequilibrium measures combined with marker distance can be used to determine the approximate effective population size (Ne) a number of generations ago. This was done on the 50k panel by Sargolzaei et al. (2008). Here, the historical effective population size in 3 breeds using the high density (777k) panel was explored for 200 generations ago and further, and the 50k panel was used to explore more recent effective population sizes. The Ne was estimated using the expectation for r^2, as described by Sved (1971) assuming no mutation, as follows:

$$E(r^2) = \frac{1}{1+4N_e\, c} \ ,$$

Where c is the recombination distance in morgans between SNP, assumed here to be equivalent to 1.1 centiMorgans per million base pairs (Arias et al., 2009), N_e is the effective population size and r^2 is the average r^2 value at a given distance. Distances were taken as the middle of a calculated range, and the r^2 value for that range was assumed to be the best estimate of the level of LD at that distance. N_e was then calculated at each distance. The age of N_e in generations can be roughly calculated as 1/2c (Hayes et al., 2003). Using the high density SNP panel, historical N_e was explored from 6 generations ago all the way back to 40,000 generations in the past.

2.4. Results

2.4.1. Extent of Linkage Disequilibrium

The r^2 values for all breeds from the 50k SNP chip are presented in Table 2.1 and the trend for each breed is presented in Figure 2.1. The Guernsey and Jersey population had the greatest extent of LD across all distances. LD in Ayrshires and Holsteins was the lowest at short distances and the trend continued for all distances. Trends for all breeds across distances were very similar and linkage decayed at a very similar rate in all breeds. As distances decreased, a smaller number of marker pairs were available to be examined. A very small number of marker pairs being examined at very small distances may have, by chance, led to the decrease in r^2 values as distances decreased below 0.02 Mb.

The r^2 values for the 777k panel for Ayrshires, Holsteins and Guernseys are presented in Table 2.2 as well as in Figure 2.2. These results show a large increase in LD as marker distance approaches zero. An r^2 value of ~0.65-0.7 at a distance of ~0.004Mb was observed, which is the average distance between markers using the 777k panel. If half of this distance is examined, which is the expected maximum distance, on average, between a marker and a QTL, we find an r^2 of ~0.75 in all three breeds. The results at distances similar to those explored in the 50k panel were very consistent with results found in the 50k except for those that were at the shortest distances on the 50k panel. This can be explained by a small number of SNPs being explored at short distances on the 50k panel.

The r^2 values for adjacent SNPs were also explored and are presented in Table 2.3. These values were taken on a per chromosome basis and a weighted average was taken based on the number of adjacent pairs on each chromosome.

2.4.2. Correlation of Linkage Phase

Consistency of phase was explored between all pairs of breeds using the 50k SNP panel. Results are presented in Table 2.4 as well as Figure 2.3. The greatest correlations of signed r values were found between Ayrshires and Holsteins. Correlations of all breeds were similar for equal distances with correlations at 0.07-0.08Mb (the average distance between adjacent markers using the 50k panel) between 0.46 and 0.57. All breed pairs show a strong upward trend up until 0.02Mb.

Consistency of phase was explored between Ayrshires, Guernseys and Holsteins using the 777k SNP panel. Results are presented in Table 2.5 as well as Figure 2.4. Similar results were found when we look at distances explored using the 50k panel. When very short distances with the 777k SNP panel are considered, we discover very high Pearson correlations of signed r values are found, approaching 1 as marker distance approaches 0. Correlations were consistently higher between Ayrshires and Holsteins than either breed paired with Guernsey. This is consistent with the findings when looking at the 50k panel.

2.4.3 Historical N_e

Effective population size (N_e) was calculated for various generations in the past. The results of this are shown in Table 2.6. Looking at the N_e in the most distant past (40,000 generations ago), effective populations were found to be ~8458, 6494 and 8285

animals for Ayrshire, Guernsey and Holstein populations, respectively. Based on an average generation interval of 7 years (Ritz et al., 2002) this corresponds to a time between when *Bos taurus* cattle diverged from *Bos indicus* and when they were later domesticated (Ritz et al., 2002). At the time of domestication of *Bos taurus* (~1428 generations ago) (Bradley et al., 1998), approximate effective population sizes of 1612, 1270, and 1768 were found for Ayrshires, Guernseys and Holsteins, respectively. At the closest measured time to the origin of the Holstein breed (200 generations ago), which occurred ~285 generations ago (Lush et al., 1936), the effective population size of the Holstein breed was found to be ~925 animals. At the most recent measure of effective population size, 6 generations ago, effective populations of approximately 90, 55 and 99 animals were found for the Ayrshire, Guernsey and Holstein breeds, respectively.

2.5. Discussion

2.5.1. Linkage Disequilibrium

The extent of Linkage Disequilibrium is a major factor in the ability to implement genomic selection. The levels of LD found in this study in Holsteins were consistent with that found by Sargolzaei et al. (2008). The level of LD in all other breeds was found to be higher at all measured distances. This is consistent with the smaller effective population sizes found in colored dairy breeds relative to Holsteins by Stackowicz et al. (2011).

Sargolzaei et al. (2008) concluded that 50,000 markers was an appropriate number for association studies with the level of LD found in Holsteins. Other Studies (e.g. Calus, 2008; Meuwissen et al., 2001) have concluded that an r^2 value >0.2 was sufficient for genomic selection to take place. At short distances, all breeds exceeded an r^2 value of 0.2 and all breeds had average adjacent marker pairs above 0.2 as well.

LD measures using the 777k panel at average adjacent marker distances (~ 0.004Mb) are high enough in all breeds to suggest that association studies as well as genomic selection would be possible given a large enough training population. Linkage Disequilibrium values (r^2) exceed 0.6 at these distances due to, on average, a low level of recombination between loci that are closely spaced. Adjacent marker r^2 values were 0.585, 0.635 and 0.588 for Ayrshires, Guernseys and Holsteins respectively. Assuming all QTL are, at most, at half of the distance between adjacent markers, r^2 values climb even higher, and exceed 0.6 for all breeds. Results were consistent between the 50k and 777k populations at given distances for all breed pairs.

2.5.2. Consistency of Phase

The consistency of phase between breeds gives a good idea of how strongly related breeds are, based on how much they have diverged over time. Breeds that are more strongly correlated at all distances diverged more recently than pairs of breeds who are not as strongly correlated over distance. The breed pair of Holstein and Ayrshire is more highly correlated at all distances on both SNP panels. This leads to the hypothesis that these two breeds diverged more recently than other breed pairs, or have undergone a greater degree of admixture since these breeds diverged historically. As phase correlation is higher for this breed pair both at short and long distances between marker pairs, it is likely a combination of these factors that lead to a greater correlation of phase between Holstein and Ayrshire populations. Correlation of gametic phase of these 2 breeds was higher at large distances however. This is due to the breeds sharing more long segments of haplotypes when compared to other breeds. This implies that there has been a greater admixture between these 2 breeds in more recent generations rather than a more recent divergence of the two breeds. This may be due to a more diverse use of sires in the Ayrshire population with some usage of Holstein sires carrying red coat colour.

Breed pairs showed a similar trend to those seen in De Roos et al. (2008), although that study was performed on populations that may have diverged at different points in time than those in this study. Correlations using the 50k panel at average adjacent distance are not high enough to allow for pooling of breeds into a multi-breed reference population for accurate genomic selection. When short distances on the 777k panel are examined, a strong correlation of phase across breeds, which is greater than 0.9 in all breed pairs, is found. With high correlations such as the ones found in this study,

one can assume that markers and QTL phases are strongly associated across breeds. Using a large number of markers, Rolf et al. (2010) found that when the high density panel was used on chromosome 29 of several beef breeds, one SNP could be found that was significantly associated with beef tenderness across breeds, which could not be found on the 50k panel. Assuming that there are many marker QTL relationships such as this one across the genome, we can assume that the implementation of genotyping with the high density panel will allow for across breed genomic selection.

2.5.3. Effective Population Size

Exploring effective population size over generations allows us to gain insight into the evolutionary history of these breeds. Sargolzaei et al. (2008) estimated a current effective population size(N_e) under 100 in Holsteins. This same result was found in this study for all three breeds studied 6 generations ago. The Guernsey population studied here had an especially low effective population size at all generations but N_e was particularly low at the most recent generations studied. This may be due to a smaller number of available sires having been collected for artificial insemination in recent years. When less recent effective population sizes are examined, much larger effective population sizes are found. Around the time of the advent of the Holstein breed as we know it today, there were approximately 925 animals in that effective population, suggesting a much more diverse population than the current world population of Holsteins. At the time of domestication of *Bos taurus* approximately 1,428 generations ago (Bradley et al., 1998), based on an average of the 3 breeds studied here, there were approximately 1,550 animals in the effective *Bos taurus* population. This implies that at this point in time there was either few *Bos taurus* animals present in the world, or that the

domesticated population that formed our current populations was selected from a certain region, where only approximately this many animals were selected, or a larger group of highly related animals was selected to domesticate.

2.5.4. Implications

For effective genomic selection to be carried out, the major limiting factors have been found to be the extent of linkage disequilibrium and the size of the reference population used to estimate SNP effects (Hayes et al., 2009c). A small reference population causes an inability to accurately estimate the many SNP effects that add up to a genomic estimated breeding value. The goal of this study was to examine both of these major problems by looking at increased marker density compared to what is currently commercially used in Canada. By increasing marker density from 50,000 to 777,000 SNPs the extent of linkage disequilibrium is increased between adjacent SNPs. Calus et al. (2008), as well as Meuwissen et al. (2001) found that with an r^2 value of 0.2, GEBV accuracies of 0.8 could be attained with a large reference population.

With significantly increased r^2 values due to a higher density SNP panel, the accuracy of genomic selection can potentially be increased even further with the same number of reference animals. By studying the correlation of linkage phase between breeds, one can also hope to find a way to pool breeds into a common reference population for training of marker effects. The results of this study show extremely strong correlations at close distances using the HD SNP panel. de Roos et al. (2008) concluded that accuracy of genomic selection could be increased by combining breeds if all breeds that GEBVs were to be calculated for were included in the reference population. Toosi et

al. (2010) found that similar strategies can be used to create accurate GEBVs for a crossbred population. For selection across breeds to be possible, however, it was found that a large number of markers would be needed. de Roos et al. (2008) found that at least 300,000 markers would be needed to implement an across breed reference population structure. The results of this study validate what was found by de Roos et al. (2008), and show that with a large number of markers, correlation of phase is very strong at close distances among dairy breeds. This will potentially allow all animals to be pooled into a common reference population to determine SNP effects. This assumes that QTLs will have similar effects on traits of interest across breeds. This is known to be true for some of the larger effect QTL that have been found. Thaller et al. (2003) found that DGAT1 has a similar effect across breeds. Using a higher density marker panel should also allow genomic predictions to be more accurate over successive generations. Higher r^2 values, as have been seen in this study at high density will lead to a lower number of recombination events between adjacent markers, and accordingly between markers and QTL over time. Therefore, with less recombination, there will be greater consistency between markers and QTL over generations and marker effects will more accurately estimate the true effects of QTL over many generations. This is only true with large reference populations where enough marker effects can be estimated without introducing noise from over-paramaterization of models leading to an overall decrease in accuracy.

A follow up study needs to be performed to calculate GEBVs using the 777k SNP panel, in order to determine the accuracy gained by using the higher density SNP panel as the size of the reference population changes. A multi-breed reference population also needs to be tested using the HD panel to ensure that QTL effects are consistent across

breeds. This study would be most useful for the Ayrshire and Guernsey populations that cannot currently generate accurate GEBVs for most traits with the current reference population sizes available in Canada.

Another major concern with the HD SNP panel is the cost of genotyping both to farmers and to the industry as a whole. The most common solution to large genotyping costs of dense marker panels has been the advent of imputation. The 50k panel has been efficiently imputed from 3,000 SNP markers. In Holsteins, Sargolzaei et al. (2010) found that this can be done with extremely high accuracy, especially when there is a strong pedigree structure, and family wise imputation can be used. Accurate imputation from the 3k panel to the 777k panel seems impossible without genotyped parents (Sargolzaei et al., 2012). We will likely, then, need to impute from the 50k panel to the 777k panel. A study needs to be performed to determine the accuracy of population based imputation methods from the 6k or 50k panel to the 777k SNP panel.

2.6 Conclusions

Guernsey and Jersey breeds had greater linkage disequilibrium measures between markers at all distances, indicating smaller effective population sizes both presently and in the past. At pair-wise distances, useful LD (r^2>0.2) was found in all breeds when measured using the 50k panel indicating that with a large enough reference population, genomic selection could be implemented within each of these five breeds. On the 50k panel, Pearson correlations were not high enough at pair-wise distances between all breed pairs to pool breeds into a common reference population, however, the correlation between the Ayrshire and Holstein breeds was higher than all other breed pairs across distance, indicating a greater level of relatedness between these breeds.

The extent of linkage disequilibrium in Ayrshire, Guernsey and Holstein breeds was extremely high when looking at adjacent markers using the high density (777k) SNP panel. Linkage decayed rapidly as distance increased, however, useful linkage was found spanning to ~0.07Mb. Correlations of phase between breeds were extremely high, and approached 0.95 at distances corresponding to adjacent SNPs using the HD panel. The use of the 777k SNP panel, although more costly, can be expected to lead to higher accuracy of genomic estimated breeding values. Using the 777k panel can allow the use of a multi-breed reference population to be a possibility, which could both increase accuracy of selection as well as make the advent of GEBVs possible in breeds that have a small reference population of genotyped animals, such as Ayrshires and Guernseys.

Table 2.1. Average r^2 values and number of marker pairs (in brackets) for five breeds at given distance ranges on the 50k SNP panel

Distance (Mb)	Ayrshire	Holstein	Jersey	Brown Swiss	Guernsey
0.02-0.03	0.2809 (6184)	0.2705 (8619)	0.3351 (6576)	0.2967 (7013)	0.3429 (5467)
0.03-0.04	0.2464 (5093)	0.2417 (6870)	0.2975 (5279)	0.2602 (5625)	0.2962 (4450)
0.04-0.05	0.2288 (4576)	0.2273 (6268)	0.2764 (4834)	0.2457 (5092)	0.2834 (4024)
0.05-0.06	0.2008 (4696)	0.1995 (6318)	0.2631 (4762)	0.2216 (5211)	0.2573 (4117)
0.06-0.07	0.1908 (4639)	0.1891 (6279)	0.2509 (4842)	0.2112 (5138)	0.2417 (4047)
0.07-0.08	0.1831 (4728)	0.1972 (6414)	0.2414 (4866)	0.1982 (5220)	0.2353 (4099)
0.08-0.09	0.1705 (4650)	0.1683 (6320)	0.2253 (4740)	0.1866 (5112)	0.2247 (3982)
0.09-0.1	0.1685 (4733)	0.1652 (6360)	0.2153 (4842)	0.1808 (5206)	0.2142 (4091)
0.1-0.2	0.1356 (45651)	0.1344 (61934)	0.1887 (46804)	0.1583 (50661)	0.1844 (39496)
0.2-0.3	0.1110 (44858)	0.1124 (60926)	0.1616 (45790)	0.1359 (49499)	0.1574 (38698)
0.3-0.4	0.1017 (44496)	0.1036 (60913)	0.1479 (45722)	0.1263 (49374)	0.1457 (38463)
0.4-0.5	0.0957 (44365)	0.0981 (60081)	0.1383 (45253)	0.1221 (48735)	0.136 (38204)
0.5-0.6	0.0936 (44006)	0.0935 (60159)	0.1320 (45182)	0.1164 (48771)	0.1293 (38013)
0.6-0.7	0.0889 (43751)	0.0896 (59563)	0.1263 (44796)	0.1142 (48463)	0.1257 (37730)
0.7-0.8	0.0856 (43670)	0.0862 (59588)	0.1236 (44581)	0.1108 (48477)	0.1222 (37714)
0.8-0.9	0.0825 (43500)	0.0846 (59175)	0.1176 (44449)	0.1078 (48138)	0.1199 (37517)
0.9-1.0	0.0818 (43454)	0.0827 (59229)	0.1144 (44260)	0.1064 (47980)	0.1162 (37429)
>1.0	0.0172 (19784411)	0.0136 (26970288)	0.0171 (20081920)	0.0185 (21872059)	0.0346 (17053834)

Table 2.2. Average r^2 values for Ayrshires, Guernseys and Holsteins at given distances using 777k SNP panel

Distance (Mb)	Ayrshire	Guernsey	Holstein
0-0.0025	0.7028	0.7549	0.7071
0.0025-0.005	0.6216	0.6743	0.6186
0.005-0.0075	0.5634	0.6159	0.5568
0.0075-0.01	0.5182	0.5725	0.5101
0.01-0.02	0.4415	0.4953	0.4282
0.02-0.03	0.3581	0.4113	0.3409
0.03-0.04	0.3071	0.3600	0.2878
0.04-0.05	0.2715	0.3233	0.2508
0.05-0.06	0.2441	0.2953	0.2231
0.06-0.07	0.2230	0.2736	0.2023
0.07-0.08	0.2066	0.2569	0.1854
0.08-0.09	0.1924	0.2434	0.1720
0.09-0.1	0.1810	0.2318	0.1607
0.1-0.2	0.1459	0.1943	0.1259
0.2-0.3	0.1158	0.1628	0.0976
0.3-0.4	0.1044	0.1483	0.0866
0.4-0.5	0.0976	0.1384	0.0797

Table 2.3. Average r^2 between adjacent SNP by breed and SNP panel

SNP panel	Breed	Average r^2	Average distance between adjacent SNP (Mb)
50k	Ayrshire	0.216	0.096
	Holstein	0.215	0.082
	Jersey	0.274	0.095
	Brown Swiss	0.238	0.091
	Guernsey	0.265	0.103
777k	Ayrshire	0.585	0.005
	Holstein	0.589	0.005
	Guernsey	0.635	0.005

Table 2.4. Pearson correlations between gametic phase of 10 breed pairs[1] on the 50k SNP panel

Distance (Mb)	AY-HO	AY-JE	AY-BS	AY-GU	HO-JE	HO-BS	HO-GU	JE-BS	JE-GU	BS-GU
0.02-0.03	0.79	0.73	0.74	0.74	0.72	0.73	0.73	0.72	0.75	0.72
0.03-0.04	0.73	0.67	0.69	0.68	0.66	0.68	0.68	0.65	0.69	0.67
0.04-0.05	0.69	0.63	0.63	0.63	0.60	0.62	0.60	0.60	0.64	0.60
0.05-0.06	0.65	0.57	0.57	0.59	0.59	0.57	0.57	0.54	0.60	0.55
0.06-0.07	0.60	0.50	0.50	0.52	0.49	0.52	0.51	0.50	0.54	0.51
0.07-0.08	0.57	0.51	0.47	0.49	0.50	0.47	0.48	0.46	0.52	0.46
0.08-0.09	0.52	0.43	0.41	0.44	0.42	0.43	0.41	0.40	0.47	0.39
0.09-0.1	0.50	0.41	0.41	0.42	0.39	0.37	0.40	0.37	0.43	0.39
0.1-0.2	0.35	0.27	0.26	0.27	0.25	0.25	0.25	0.24	0.32	0.24
0.2-0.3	0.20	0.11	0.10	0.11	0.11	0.10	0.10	0.10	0.18	0.08
0.3-0.4	0.12	0.07	0.05	0.07	0.05	0.03	0.06	0.05	0.12	0.03
0.4-0.5	0.09	0.03	0.03	0.04	0.03	0.01	0.06	0.04	0.09	0.01
0.5-0.6	0.08	0.01	0.01	0.02	0.03	0.01	0.04	0.02	0.07	0.01
0.6-0.7	0.07	0.01	0.01	0.02	0.01	0.01	0.03	0.01	0.07	-0.01
0.7-0.8	0.07	0.00	0.01	0.01	0.01	0.00	0.02	0.00	0.06	0.01
0.8-0.9	0.04	0.00	0.00	0.01	0.02	0.00	0.03	0.01	0.07	0.01
0.9-1.0	0.05	0.01	0.01	0.01	0.01	0.01	0.02	0.00	0.05	0.00
1.0-10.0	0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01	0.00

[1] AY - Ayrshire, HO - Holstein, JE – Jersey, BS – Brown Swiss, GU - Guernsey

Table 2.5. Pearson correlations of signed r value between 3 breed pairs for Ayrshire, Holstein and Guernsey at given distances using the 777k SNP panel.

Distance (Mb)	Ayrshire-Holstein	Ayrshire-Guernsey	Holstein-Guernsey
0-0.0025	0.97	0.95	0.96
0.0025-0.005	0.96	0.94	0.94
0.005-0.0075	0.94	0.92	0.92
0.0075-0.01	0.93	0.90	0.90
0.01-0.02	0.90	0.86	0.86
0.02-0.03	0.85	0.79	0.80
0.03-0.04	0.80	0.73	0.74
0.04-0.05	0.75	0.68	0.69
0.05-0.06	0.71	0.62	0.63
0.06-0.07	0.67	0.58	0.59
0.07-0.08	0.63	0.53	0.54
0.08-0.09	0.59	0.49	0.50
0.09-0.1	0.56	0.46	0.47
0.1-0.2	0.41	0.30	0.31
0.2-0.3	0.23	0.14	0.15
0.3-0.4	0.14	0.07	0.08
0.4-0.5	0.11	0.04	0.06

Table 2.6. Effective Population size for Ayrshire, Guernsey and Holstein breeds for a given number of generations ago (ngen)

Distance (Mb)	Ayrshire	Guernsey	Holstein	ngen
0.0000125	8458	6494	8285	40000[a]
0.0000625	3100	2495	3184	8000
0.00035	1612	1270	1768	1428[b]
0.00075	1280	964	1465	667
0.0025	764	514	925	200[c]
0.0055	440	306	441	91
0.0085	327	216	318	59
0.015	216	147	210	33
0.025	159	108	153	20
0.055	107	69	109	9
0.085	90	55	99	6

[a] *Bos taurus* ancestral population

[b] Approximate first domestication of *Bos taurus*

[c] Close to origin of Holstein breed (~2000 years ago)

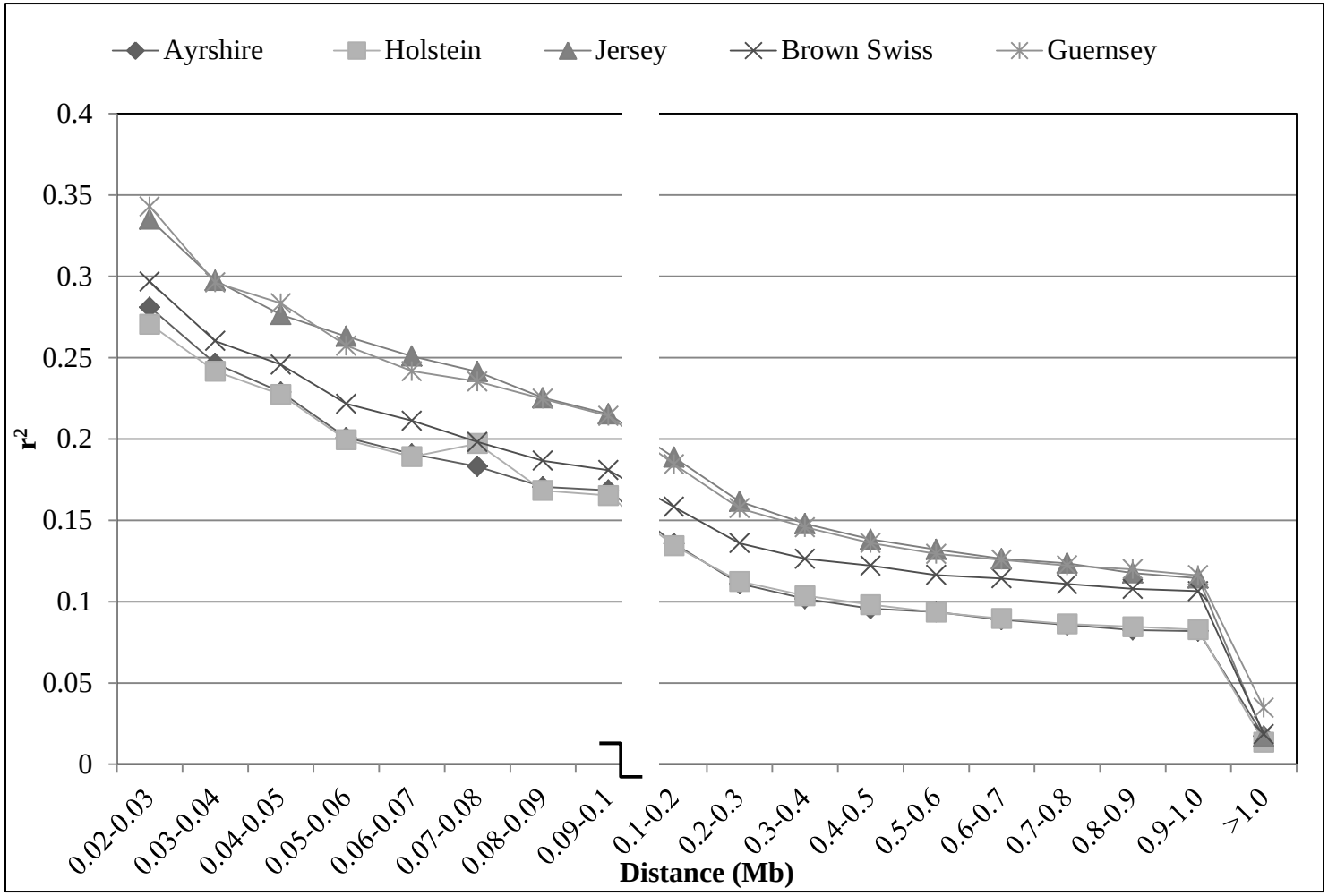

Figure 2.1. Average r^2 values at given distances for five breeds using the 50k SNP panel.

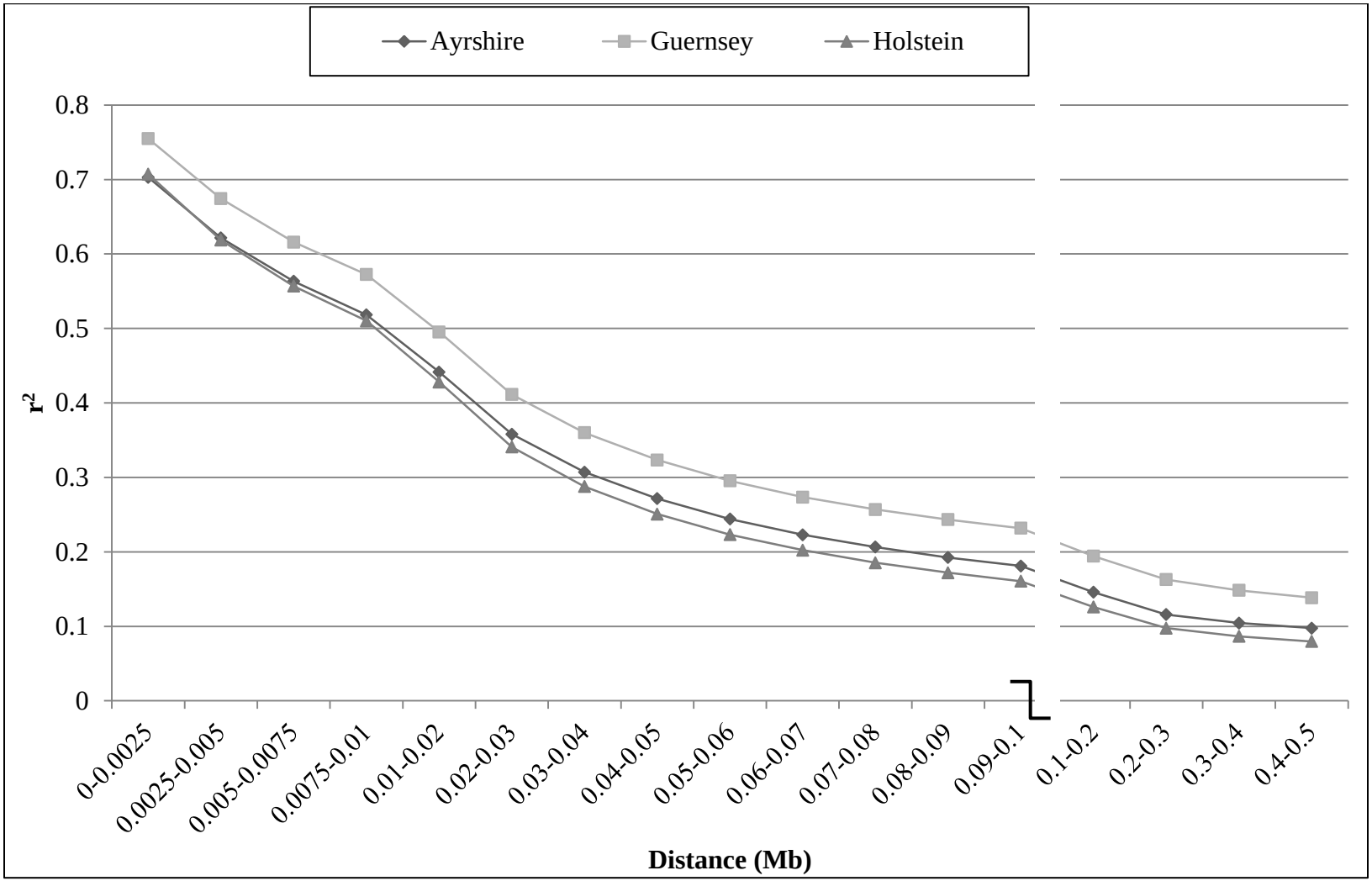

Figure 2.2. Average r^2 values for Ayrshires, Guernseys and Holsteins at given distances using the 777k SNP panel

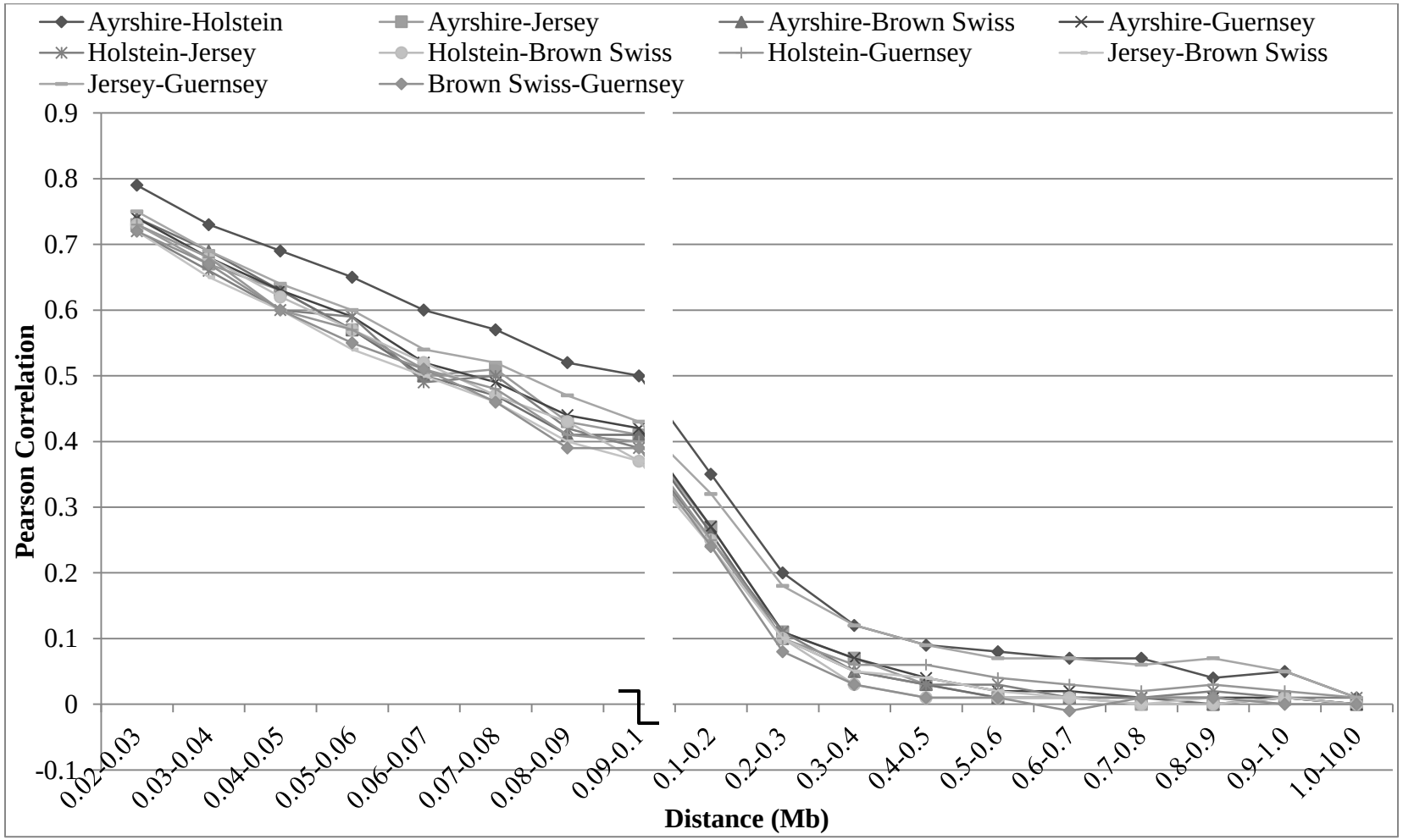

Figure 2.3. Pearson correlations of signed r values at given distances for ten breed pairs using the 50k SNP panel.

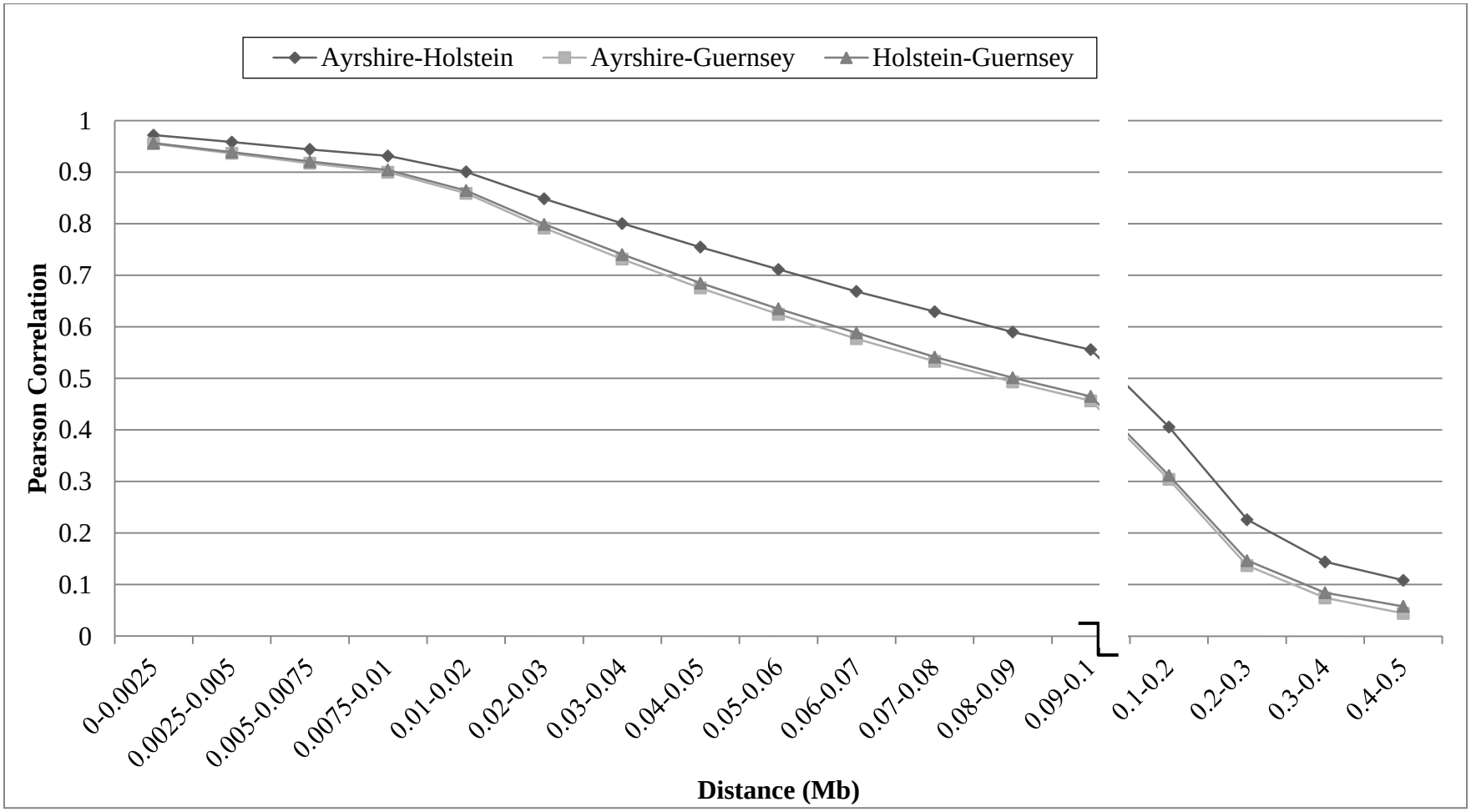

Figure 2.4. Pearson correlations of signed r values between Ayrshires, Holsteins and Guernseys at given distances using the 777k SNP panel.

CHAPTER 3

IMPUTATION FROM LOW TO HIGH DENSITY USING WITHIN BREED AND MULTI-BREED REFERENCE POPULATIONS IN HOLSTEIN, GUERNSEY, AND AYRSHIRE CATTLE

3.1. Abstract

Genomic selection in dairy cattle requires dense marker panels to effectively estimate the effects of quantitative trait loci that are in linkage disequilibrium with markers on these dense marker panels. Dense panels are often expensive and so imputation is often carried out to infer markers on a dense panel from animals genotyped with a sparser panel. The advent of the 777k high density (HD) SNP panel may allow for genomic selection to take place across breeds. Due to the high cost of this panel, the accuracy of imputation from lower density marker panels (6k or 50k) was examined both within breed and using a mutli-breed reference population in Holstein, Ayrshire and Guernsey. Both population-based and a combined family and population-based methods were tested. Imputation from 6k to 777k was also examined using 1 or 2-step approaches. Beagle V3.3 and FImpute programs were used for carrying out imputation. High Density genotypes were taken for Holstein, Ayrshire and Guernsey (n=1115, 531 and 60, respectively). Animals were then split by year of birth into reference and imputation groups. For imputation animals, markers on the HD panel were filtered to only include those SNPs found on the lower density genotyping platforms (6k or 50k). Imputation was carried out for a number of scenarios both within breed and using a multi-breed reference population. Imputation accuracies were then calculated as the proportion of correct SNPs

out of all those that were filled in by the different imputation methods and the different imputation software. Computation time was also explored to determine the efficiency of the different methods and software for imputing both from 6k and 50k to HD. Very high imputation accuracy (>0.97) was found for all breeds when imputation was carried out with FImpute both using population-based as well as combined family and population-based imputation. Accuracy for Guernsey and Ayrshire was slightly lower when using the population-based imputation method employed by Beagle. Imputing using a multi-breed reference group by combining the reference populations had a very small effect on imputation accuracy for 50k to HD. This effect was larger in breeds with less genotyped animals in the reference population. When imputing from 6k to HD, it was found that a two-step method gave higher accuracies for all breeds both within breed and when all breeds were used in the reference population. It was also found that adding information from other breeds was detrimental to imputation accuracy when imputing from 6k both in one and two-step procedures. Computing time was significantly greater when using Beagle imputation software, with all comparable procedures being 9-13 times less efficient compared to the FImpute software. Overall it was found that imputation could be effectively carried out to fill in missing markers for both the 6k and 50k panels to reduce genotyping costs while growing the size of the population of HD genotyped animals.

3.2. Introduction

Genomic selection in dairy cattle uses information from dense marker panels to estimate effects of Quantitative Trait Loci (QTL) that are at or are in Linkage Disequilibrium (LD) with markers on these dense panels (Goddard and Hayes, 2007). These dense marker panels, however, are expensive and can be a major limitation to the number of animals available to be genotyped. The advent of imputation has significantly aided this problem, making it possible for many animals to be genotyped with lower density marker panels and imputed to higher density for genomic selection. In Canada, imputation takes place regularly from approximately 3,000 and 6,000 Single Nucleotide Polymorphisms (SNPs) to a panel of nearly 50,000 SNPs. This is done by using the FImpute program v1 (Sargolzaei, 2010) using only the family-based component of the algorithm, followed by population imputation with Beagle. Similar strategies are carried out in the United States and around the world using a variety of methods using both family-based imputation methods as well as those relying on population-based imputation (Johnston et al., 2011). For a large reference population, imputation from ~3k to ~50k has been shown to be accurate using both family-based and population-based methods with accuracy exceeding 0.9 for population-based methods as IMPUTE (Druet et al., 2010, Weigel et al., 2010, Nothnagel et al., 2009), Beagle (Druet et al., 2010, Nothnagel et al., 2009, Calus et al., 2011), and fastPhase (Weigel et al., 2010, Nothnagel et al., 2009, Calus et al., 2011) as well as combined family and population based methods CHROMIBD (Zhang et al., 2010), FImpute (Sargolzaei et al., 2010, Johnston et al., 2011) and DAGPHASE (Zhang et al., 2010). It was found that when family information is available, family-based methods do have an advantage over population-based methods (Zhang et al., 2010) More

notably, it was found that a combination of FImpute v1 and Beagle could yield greater than 95% accuracy when combined (Johnston et al., 2011).

It has been found that for a limited reference population size, such as that of Ayrshires and Guernseys, a denser SNP panel is required to increase the accuracy of genomic selection (Hayes et al., 2009c). A number of these animals are currently genotyped with a panel encompassing approximately 777,000 SNP. The cost of this panel is high still, and creates an even greater problem in creating a suitably large reference population for generating accurate genomic breeding values. Imputation from lower density panels to the 777k panel would significantly decrease genotyping costs. Van Raden et al. (2011) found that 500,000 markers could be accurately imputed from approximately 50,000 markers (>95%) using a combination of family and population based methods. This was in a Holstein population with strong family information available. In a population with limited family information, using a family-based method may not yield a significant increase in accuracy of imputation. Calus et al. (2011) found that Beagle outperformed a family-based multivariate mixed model method for imputation when a high density marker panel was simulated (more markers simulated surrounding the marker to be imputed in the same span).

The goal of this study is to investigate the accuracy of imputation from 6k or 50k to 777k using real Ayrshire, Guernsey and Holstein data. Both a population-based (Beagle V3.3) and family and population-based (FImpute) method will be tested. The effects of a larger reference population and the effect of having direct ancestors genotyped will be examined. In addition, imputation from an even lower density panel (6k) to the high density panel using a 2 step approach will be assessed.

3.3. Materials and Methods

3.3.1. Data

High Density (777k) genotypes from Illumina bovine HD Beadchip (Illumina Inc., San Diego, USA) for Holstein, Ayrshire and Guernsey (n=1115, 351 and 60, respectively) were used. The Holstein data consisted of both bulls and cows from the North American Collaboration for Genomic Prediction, whereas the Ayshire and Guernsey data set consisted of only bulls with Canadian official proofs (also from the North American Collaboration for Genomic Prediction). The data was examined to determine if alleles with low minor allele frequency had an effect on imputation accuracy. This was done by imputing both using all possible SNP, and by removing SNP to a sub-set where markers shown to be in high LD (~300,000 markers) with one another were removed. No effect on imputation accuracy was found, so the entire set of approximately 777,000 markers was used for imputation purposes. However, markers mapped to the X chromosome as well as the pseudo-autosomal region (chromosomes 30 and 31) were excluded due to inconsistencies in the marker maps. This left a set of 735,293 SNPs to be examined.

To perform imputation, animals were divided up into 2 distinct groups, reference and imputation animals. This sorting was done by birth year in an effort to capture as many sires of imputation animals in the reference population as possible, creating a practical scenario that mimics how imputation is carried out for routine genomic evaluations. In Ayrshires and Guernseys animals born in 2000 or after were included in the imputation population. This gave reference populations of 211 and 41 and imputation

groups of 140 and 19 for Ayrshire and Guernsey, respectively. Holsteins born in 2004 or later were included in the imputation group, creating populations of 892 and 223 for reference and imputation. There were a number of Holsteins that had no known birthdate, these were included in the reference population. Number of animals who have a genotyped sire, dam, maternal grand sire and paternal grand sire was also examined and presented in Table 3.1.

3.3.2 Mimicking Low-Density Marker Panels

Imputation was carried out from both the 6k and 50k Illumina SNP panels. Both of these panels markers are almost entirely included on the HD SNP panel. This being the case, the HD panel was then filtered to erase all SNPs not contained in whichever low density genotyping panel was being investigated. This was performed for all animals in the imputation population to mimic not having been genotyped on the high density SNP chip. This left 39,946 and 6,556 SNP to be considered for imputation on the 50k and

6k panels, respectively. It should also be noted, that all 6,556 SNP in the 6k chip are present in the 50k chip as imputation from 6k was also carried out as a two-step procedure (from 6k to 50k and then from 50k to HD) in an effort to improve overall imputation accuracy.

3.3.3. Imputation Scenarios

A number of imputation scenarios were carried out in this study. The primary goal was to evaluate the effect of family vs. population based imputation both within breed and with a multi-breed reference population using the HD SNP panel. This was done by imputing all breeds individually as well as combining all populations for imputation both

in Beagle V3.3 (population-based) and in FImpute (family and population-based). We also wanted to determine accuracy of imputation when comparing population only imputation in both Beagle and FImpute. This was performed in FImpute by removing all pedigree information before imputation. All the above scenarios were only carried out for the 50k SNP panel. Imputation accuracy for the 6k panel was determined either as a single step (from 6k to HD) or using a two-step procedure (imputing from 3k to 50k followed by imputation from the imputed 50k panel to 777k markers), as a two-step has been found in other preliminary studies to increase imputation accuracy rather than imputing directly from the 6k panel to high density. This was carried out both within breed and with a multi-breed reference using FImpute with family information included. For all scenarios, using both imputation programs, the effect of having family information in the pedigree was examined. Imputation accuracy rates between animals who had genotyped parents in the reference population, and those that did not were compared. Additionally, to determine the amount of information that could be gained from a multi-breed reference population, imputation was carried out using entirely animals from another breed (Holsteins) as the reference population to impute animals in other breeds (Guernseys and Ayrshires). To determine the amount of useful information this provided, imputation was also carried out randomly, using minor allele frequencies. This was averaged over 20 iterations to determine the average accuracy of random imputation from allele frequency alone.

3.3.4. Imputation of Missing Markers - FImpute

3.3.4.1. Family Based Imputation

The family based imputation algorithm used by FImpute as described by Sargolzaei (2010). Family-based imputation with FImpute is a 3 step procedure. In the first step, parent or progeny information is used to fill in missing genotypes where it can be done with a high degree of certainty (long stretches of consistent markers are found on the low density marker panel). The second step is to reconstruct marker haplotypes as described in detail by Sargolzaei et al. (2008). Haplotypes are reconstructed iteratively. This is done by considering parent information first to determine phase at a marker locus when at least one parent is homozygous at that loci. When phase is still unknown, the nearest partially informative heterozygous marker is used along with linkage information between those 2 markers to infer haplotype probabilities further. Partially informative markers flanking the marker in question are then found and once again, linkage information is used to estimate haplotype probabilities. This iteration is repeated until the sum of squares of haplotype probabilities is sufficiently small. Haplotypes of progeny are then matched to haplotypes of parents and untyped loci are then filled in.

3.3.4.2. Population Based Imputation

When there is no family information present for an individual or the pedigree file is excluded, FImpute uses population imputation. Population imputation is carried out by FImpute using overlapping windows to reconstruct haplotypes. FImpute, unlike most population imputation software, assumes that all animals are related to some degree and uses these overlapping windows to find segments of haplotype that are consistent

between individuals having come from a common ancestor. The windows are large at first to find segments of haplotype that come from more recent ancestors. The window walks along each chromosome finding large segments consistent with reference animals, overlapping by 75% of the window size each step (This overlap can be modified to optimize accuracy and computing time). After each chromosome has been completed with large windows, the same process is repeated numerous times with smaller and smaller windows to capture consistent haplotypes from less recent ancestors. This procedure is repeated until all markers have successfully been filled in. When multiple haplotypes are found at a certain window size, haplotype frequency in the reference population is used to determine the most likely haplotype, and fills that haplotype into the imputed animal's genotype.

3.3.5. Imputation of Missing Markers - Beagle

The Beagle imputation method is described in full detail in Browning and Browning (2007). Beagle uses a "localized haplotype-cluster" model to perform imputation on missing genotype markers. This algorithm uses only local haplotype data in order to capture markers in tight LD with one another. It uses an underlying Hidden Markov Monte-Carlo (HMM) approach to determine transition probabilities from one "node" to another based on overall haplotype counts. In this case, a node is a collection of haplotypes that have the same allele at a certain locus. So, at a given node, the probability of the following allele, or the probability of moving to a certain child node is determined. The sums of all of these probabilities for the entire pathway from the root node to the terminal node give the probabilities of each unique haplotype. Beagle first uses a phasing algorithm to determine haplotype phase for each individual. This is done by first

constructing the local haplotype clusters, and then sampling a number of haplotypes for each individual from the HMM. The sampled haplotypes are then used to reconstruct the local haplotype cluster. This is repeated over 10 iterations to achieve a high level of phasing accuracy while maintaining computational efficiency.

3.3.6. Calculation of Imputation Accuracy

For this study, imputation accuracy was measured using a ratio of correct call to overall call rate. Any markers that were left as missing were not included in the calculation of accuracy. This meant for Beagle that accuracy was always the same as correct call rate, as all markers are filled in the Beagle imputation algorithm. If a haplotype in the imputation population is not detected in the reference population in FImpute, all markers within this haplotype are filled in as missing. These missing calls do not contribute towards the measures of accuracy reported. Missing calls in FImpute comprised a very small percentage of all marker calls and should not bias the overall results or the comparison to Beagle to any extent. It should also be noted that the most probable call at any locus was considered in all cases to be the allele present. That is to say, no threshold was set on allele probability for it to be included in predictions of imputation accuracy. Markers were filtered originally from the HD genotype panel, so to calculate correct call rate, the imputed genotype was compared to the complete HD genotype and the amount of correct calls, incorrect calls and those called as missing were calculated. Correct calls are those in which the call after imputation exactly matches that of the original HD genotype, excluding those markers that were present on the low density marker panel. This is a slightly downward biased measure of imputation accuracy when compared to the allelic R^2, as described by Browning and Browning (2009), which

takes into account the correlation between heterozygous and homozygous calls given that if one allele at a locus is correct there will still be valuable information available at that locus for further studies.

3.3.7. Comparison of Accuracy between Animals with and without Genotyped Parents

The difference in accuracy realized between groups with and without one parent genotyped was determined in the Ayrshire population. This was done by creating separate validation groups, those with and those without having their sire genotyped. It should be noted that all genotyped Ayrshires were male, so there were no genotyped dams. Realized differences in accuracy and correct call rate were then determined. Accuracies were calculated for each individual animal in both groups and were used as a new data set for analysis. The Proc ANOVA procedure is SAS (SAS Institute, Inc., Cary, USA) was used to determine if there was a significant effect of having a sire genotyped in all scenarios for the Ayrshire breed. The number of animals in each group is shown in Table 3.1.

3.3.8. Computing Time

The amount of time used imputing in each scenario with each algorithm was also recorded. This was done on a per chromosome basis as due to computational requirements, a different number of chromosomes were run in parallel in each scenario. Measuring computational efficiency is important when determining the efficacy of each algorithm especially when it is to be applied to larger and larger data sets for routine genomic evaluations.

3.4. Results

3.4.1. Imputation Accuracy

First, the difference between population vs. family and population based imputation when imputing from the 50k SNP chip to the HD (777k) chip using FImpute was examined. The results of this imputation are presented in Table 3.2. There was very little difference seen in terms of imputation accuracy between these 2 methods when FImpute was used with or without pedigree information. Accuracy as well as correct call rate was within one percent in all scenarios when comparing these 2 methods.

Secondly, the difference between the population-based imputation performed by Beagle compared to imputation performed by FImpute when pedigree information was omitted (population imputation only) was considered. These results are presented in Table 3.3. There is very little difference in imputation accuracy or correct call rate between the two methods when population-based imputation is carried out on the Holstein population. There is a difference, however, when one looks at the imputation accuracy between these 2 methods for Guernsey and Ayrshire, with FImpute out-performing Beagle in both cases. Imputation accuracy is higher by nearly 1% for the Ayrshire population and nearly 2% for Guernsey. As reference population size decreases, it seems that FImpute performs better. It should be noted that the missing rates in all scenarios were less than 0.1 percent, and as such did not create any bias in comparing the imputation algorithms.

The use of all breeds as a reference population for imputation using both Beagle and FImpute (with and without pedigree information) was assessed. These results are

presented in Table 3.4, along with imputation accuracy when reference populations were comprised within a single breed. For both breeds with smaller reference populations (Ayrshire and Guernsey) an increase in imputation accuracy as well as correct call rate was observed when information from other breeds was included. This difference was greater when imputation was carried out with Beagle. The largest difference was seen for Beagle in the Guernsey population when information from Holstein and Ayrshire were included. Including other breeds led to over a 1% increase in imputation accuracy. FImpute also gained from using information from other breeds, however the difference was smaller. The Ayrshire data set gained ~0.3% in accuracy from adding Holsteins and Guernseys when family and population based imputation was carried out and ~0.1% without pedigree information. Guernsey data gained slightly less with increases in accuracy of 0.2% and 0.1% with and without pedigree information, respectively even though there were no genotyped parents of any of the Guernsey individuals in the imputation population.

Imputation accuracy was also determined when imputing from 6k to HD. These results are presented in Table 3.5. Imputing in two steps yields a higher accuracy than when imputation is carried out directly from 6k to the high density platform. This is seen both within breed and using a multi-breed reference population. Adding information from other breeds has a detrimental effect on imputation accuracy in both the one-step and two-step imputation methods from 6k to HD. Imputation accuracy is, however, generally lower when imputing from 6k to the HD panel, especially in breeds with smaller reference populations than Holsteins.

To determine the amount of information that could be gained from a multi-breed reference population for genomic selection, imputation accuracy was also measured when using a reference population comprised entirely of animals from another breed (in this case, Holsteins). In addition, accuracy when imputation was carried out based solely on minor allele frequency was explored to determine the difference between random imputation and imputation using only Holsteins in the reference population for both Ayrshires and Guernseys. These results are presented in Table 3.6. Increases in correct call rates of 15% and 11% were seen for Ayrshires and Guernseys, respectively.

To determine to what extent accuracy was lost in each step of the two-step imputation process, the imputation accuracy for the first step only was measured. These results are presented in Table 3.9. Measured accuracies for this step closely reflect the accuracy found for the entire 2-step process indicating that most incorrect calls occur during the first step of 2-step imputation. The marker density of the panel which animals that are to be imputed are genotyped on has a much larger effect than the density of the high density panel in imputation studies.

3.4.2. Determination of Family Effect

Imputation accuracies were measured on a per animal basis and used to determine if there was a significant effect of having or not having a genotyped sire. Table 3.7 shows the results of this analysis. Mean correct call rate and accuracy are presented as well as the difference between these two groups. In addition Table 3.7 presents the corresponding P values from the ANOVA used to determine if there was a significant effect of having sire genotyped. Data was transformed and ANOVA was performed on the log score of 1

minus the accuracy value for each individual. In all scenarios the P-value was below 0.05 and the effect of having a sire genotyped was deemed significant.

3.4.3. Computing Time

The results for overall computing time as well as computing time per chromosome for all imputation scenarios are presented in Table 3.8. The largest differences in computing were seen between Beagle and FImpute, with FImpute being 9-13 times more computationally efficient in all comparable scenarios. There was also a difference in computing time within the FImpute imputation program when pedigree information was included or omitted. A slight increase in computational efficiency was seen when pedigree information was excluded. When imputation from 6k was examined, there was an increase in computational efficiency when the two-step procedure was carried out due to the algorithm not having as many possible haplotype blocks to consider in each imputation step.

3.5. Discussion

3.5.1. Imputation Accuracy

Family-based imputation algorithms have been shown to be extremely effective in dairy cattle populations with a high level of pedigree information and a large reference population (Sargolzaei et al., 2010). The results presented in this study, however, show that a population-based imputation algorithm can be just as accurate in all sizes of reference population when imputation animals are genotyped with a dense enough panel. The number of animals with genotyped ancestors are seen in Table 3.1, The imputation animals within the Guernsey breed had no genotyped ancestors, while Holsteins and Ayrshires had some ancestors genotyped both maternally and paternally. There was no difference seen in accuracy between family and population based algorithms in any case, however, as the population based algorithm in FImpute is still able to detect ancestors without pedigree information based on long conserved segments of haplotypes. The family and population based algorithm used in this study did, however, default to population imputation in the absence of pedigree information, which was the case in a large proportion of the imputation scenarios. This was especially true in Guernseys, where no sires of imputation animals were present in the reference group. Previous studies showing an advantage to family-based imputation methods have focused on imputation from either the 3k panel or the 6k panel to the 50k SNP chip. When imputing from 50k to HD shorter segments of haplotype conserved over many generations can be found when compared to the lower density panels and thus can be accurately imputed using only a population-based algorithm without knowing any pedigree information.

There may be an advantage in computational efficiency when family imputation is employed as haplotypes can be more quickly identified when ancestors are known.

Many population-based imputation methods exist and are commonly used for a large number of purposes, including boosting power of association studies, fine-mapping, imputation of untyped or non-SNP variation, among others (Marchini and Howie, 2010). The Beagle imputation method has been used extensively in human genome studies. In this study we compared Beagle to the population algorithm in FImpute, which assumes some level of relatedness between all animals. FImpute performed more effectively in most scenarios, especially those in which a smaller reference population was available. Beagle, however, had a greater gain from multi-breed imputation. Beagle considers all animals unrelated, and could capture haplotypes that were common between breeds because of this. Although animals across breeds are generally unrelated, there remain segments of haplotype conserved between breeds from before those breeds diverged. This can be seen by a highly consistent persistency of gametic phase across breeds at short distance as seen in Gautier et al. (2007).This was also observed in the second chapter of this thesis, where all breed pairs have highly correlated gametic phases. FImpute is able to take this into account. When haplotype window size is very small, FImpute considers reference animals and imputation animals to have a common ancestor many generations ago, in this case, before breeds had diverged. This explains the small gain still realized by the FImpute software when more than one breed was included in the reference population.

The population structure of the animals used in this study may have had an influence on the high imputation accuracies that were realized. Due to the populations of Ayrshires

and Guernseys used being comprised of solely proven Canadian sires, a high degree of relatedness exists between all animals studied within each breed. A high degree of relatedness was found by Berry and Kearney (2011) to have a positive correlation with genotype concordance rate, a measure of imputation accuracy. Due to the extremely heavy use of artificial insemination leading to high selection intensity in almost all Canadian dairy cattle, this should not have a large influence on how these results are to be interpreted when applied to imputation for genomic selection in Canadian dairy cattle. It should be noted, however, that, when imputing animals from a less related population, slightly lower imputation accuracies should be expected. This is especially true when considering imputation from the 6k panel, where less historical linkage is relied on and more recent LD patterns are examined.

In an effort to grow the reference population for smaller breeds in a fast and economically viable manner, the accuracy of imputing from the 6k panel directly to the HD as well as doing so in two imputation steps was examined. The results ranged from 0.88-0.97, however, for Guernseys and Ayrshires (the breeds with greater need of a larger reference population) accuracy values were generally between 0.91-0.95. These accuracies are comparable with many population based imputation accuracies from 3k-50k, reported by Daetwyler et al. (2011) in which only very modest decreases in accuracy of genomic breeding values were found using these imputed genotypes. Thus, imputation from 6k to HD may be feasible in breeds with smaller reference populations and will become more and more useful as the high density reference set grows in these populations. Multi-breed imputation from very sparse panels to high density was detrimental when compared to within breed reference in all scenarios. Gautier et al.

(2007) showed that significant consistency of phase is only present at distances shorter than 0.1kb. Average marker pair distance on the 6k panel is much longer (~0.5kb). Similar results were found in chapter 2 of this thesis, where consistency of gametic phase was determined to be very high (>0.94) at very short distances explored on the HD panel. Gametic phase correlations were not sufficiently high (<0.9) when exploring pair-wise distances found on the 50k panel. This leads to a lack of conserved haplotypes between breeds, and very little useful information can be captured from animals outside of the breed, along with some incorrect information being filled in due to haplotype frequencies differing significantly across breeds at this distance as well. The accuracy of imputation carried out in one or two steps using the 6k panel was also examined. A slight increase in accuracy was found when we carried out imputation in two steps, from 6k to 50k and then from 50k to 777k. When imputing from 6k directly to HD, there are a large number of possible haplotypes present in the reference population given any short haplotype segment from the sparse panel. This leads both to a decrease in accuracy as well as an increase in computing time when comparing these two scenarios to imputation from the 50k panel.

To gain further insight into how helpful information from other breeds would be both for imputation as well as genomic selection, the Guernsey and Ayrshire populations were imputed using only Holsteins as the reference population. Compared to a base-line level of the average accuracy of random imputation using only allele frequencies over 20 replicates, there was a significant increase in correct call rate. This is an indication that there are many conserved segments of haplotype between Holstein and both of the other breeds. These segments will be useful for both imputation and are the reason multi-breed

imputation accuracy is higher than within breed when imputing from 50k to HD. Having shared haplotype segments between the breeds will also potentially lead to an increase in genomic selection accuracy as any QTL within the regions of consistent linkage across breeds will potentially increase accuracy of estimated SNP effects. Adversely, areas where linkage is not maintained to a high degree may have a negative effect on accuracy. If a SNP is in the opposite phase with a QTL across breeds the SNP effect will be inaccurately estimated as reference animals in each breed will have opposing effect.

3.5.2. Computing Time

One important consideration when evaluating the application of imputation to routine genomic evaluations is computing time. Computing time for imputation is determined by the reference population size, number of animals to impute as well as the density of the high and low density marker panels to which the imputation algorithm is being applied. Imputation carried out with FImpute more accurate and ten times more efficient when compared to Beagle for identical scenarios. This advantage in computational efficiency is quite significant considering that FImpute also imputes more accurately in most scenarios. For large data sets, FImpute is able to impute much more efficiently by only considering a subset of the reference population haplotypes. It also is more efficient at capturing long segments of consistency between reference and imputation animals which fills in large portions of the genome, leaving fewer haplotypes to be filled in by considering allele and haplotype frequencies. Nothnagel et al (2009), as well as Marchini and Howie (2010), found Beagle to be among the more computationally efficient methods for population imputation. Given the results of this study we can assume that FImpute is easily among the best methods for imputation in terms of computing time.

3.5.3. Comparison of Results to Other Studies

Van Raden et al. (2011) simulated 500,000 markers for all Holsteins genotyped in the United States with the 50k SNP panel (33,414). An imputation accuracy of >95% was found when imputing from 50k to the simulated marker set. After imputation, it was found that an increase in genomic reliability of 1.6% could be added to an already high genomic reliability with a very large reference population genotyped with the 50k SNP panel. With slightly higher imputation accuracy found in this study and a lower genomic reliability to begin with due to a smaller population, much larger gains in GEBV reliability can be expected in the Ayrshire and Guernsey populations studied here. With a comparable number of samples to the Ayrshire population in this study, using a population that was comprised of both related and unrelated humans, Nothnagel et al. (2009) reported similar accuracies using the Beagle imputation algorithm. The imputation carried out in the human genome was done from 24,185 SNPs to 586,217 SNPs on the Affymetrix SNP panels. These marker densities in both the low and high density scenarios are similar to what was examined in this study. Due to the high cost of genotyping, there is a limited population of animals genotyped with panels of the density used in this study. For this reason, imputation from low or moderate density SNP panels to a panel of this high density has not been carried out in real livestock data. The imputation accuracy was, however, expected to be higher when imputing from 50k to 777k than has been recorded in the many studies imputing from 3k or 6k to the 50k SNP panel. A higher level of linkage is present between SNPs at the average distance between markers on the 50k panel and so there will be greater association with the SNPs in between markers that are to be filled in when imputing to high density.

3.5.4. Implications

The results of this study look promising for application in genomic selection. Using the high density marker panel should allow for more accurate genomic selection in breeds with smaller reference populations for the training of SNP effects by using a multi-breed reference set. Imputation of animals accurately to the HD panel will also allow for these populations to incorporate more animals, both male and female, in an economic manner. Daetwyler et al. (2011) found little decrease in accuracy of genomic breeding values from imputed marker panels with an average imputation accuracy of ~92%. In all situations, there was a method to impute with at least this level of accuracy for all scenarios studied here. A relatively accurate GEBV could be generated for every animal in this study, given there is a sufficiently large reference population for estimating SNP effects. GEBVs generated from the HD panel may also have an advantage in successive generations when compared to those estimated using the 50k SNP panel. With closer linkage of markers to QTL, resulting in greater linkage disequilibrium measures, there will be less frequent recombination events between a QTL and the nearest marker. This will result in less change of marker effects over time, as markers will consistently be estimating the same QTL effects over more successive generations. Further studies need to be completed to determine the loss in accuracy of GEBVs when the reference population for training SNP effects is largely comprised of imputed animals.

High imputation accuracies both within breed and using a multi-breed reference population as found in this study may have significant implications outside of the dairy breeds. Being able to gain information from animals outside of the target breed when imputing to a high density panel may be extremely helpful for both beef genomics and

possibly have a large effect in other species. The beef population has a much greater number of breeds than does the dairy population in Canada. This breed diversity can be a major hurdle in implementing genomic selection for beef cattle. Being able to work across breeds, when using high density genotyping, will allow for all beef breeds to gain a large amount of information quickly. If other species have similar consistency of gametic phase across breeds as has been found in dairy cattle, combining breeds in those species for imputation as well as genomic selection should likely be advantageous. As reference populations grow, less and less will be gained from a multi-breed reference population for imputation, however, results have shown that there may still be some value to a combined population for genomic selection. As genotyping costs remain high, imputation is an invaluable tool to gain information on select animals while keeping costs for producers low.

Having increased accuracy of imputation in all scenarios, when a sire is included in the genotyped reference population gives valuable insight about what animals should be genotyped with the high density panels in the future to gain the most from large-scale imputation for genomic selection. Key ancestors to the general population should be continually genotyped in an effort to most accurately impute the greatest number of animals. This includes continuing to genotype key proven sires, as they will have highest degree of relatedness to the entire population and thus will lead to the greatest gain in accuracy of imputation and, subsequently, in accuracy of genomic selection itself.

3.5.5. Future Studies

The next step in determining the validity of genomic selection within and across breeds with the high density SNP panel is to estimate GEBV accuracies. This will be done both as single-breed evaluations as well as using all breeds as a reference group for SNP effect estimation. Based on the results of this study, we should see gains in GEBV accuracies after validation when a multi-breed reference structure is implemented. This follow-up study will also help to determine how many animals will be needed in the reference population for sufficient accuracy using the HD SNP panel. Imputed genotypes should also be tested both as reference and validation animals to determine the accuracy lost when a significant number of imputed animals are included in either of those groups.

With a large number of SNPs, there is a risk of models being severely over-paramaterized. This problem increases significantly when genomic selection moves towards denser and denser SNP panels. Moser et al. (2009) tested various methods for genomic selection, including a method that involved only using select SNPs that had larger effects. This method was less accurate, however, in the aforementioned study, only ~7,000 total SNPs were considered. There may be merit in methods that select only SNP of moderate to large effect when we consider genomic selection on very dense marker panels. Calus et al. (2008) also showed a difference in the performance of certain models at different SNP densities. There may be greater merit to a different method for QTL effect estimation when marker density increases. Calus et al. (2008) showed greater genomic accuracies at higher density with a haplotype model when trait heritability was high and with a SNP effect model when trait heritability was lower.

In the future, the methods and results discussed in this study could be applied for imputation of genome sequence. As technology continues to evolve and genotyping costs continue to decrease, it may become viable in the near future to sequence enough animals across both beef and dairy breeds to efficiently and accurately impute lower density genotypes to a full sequence, creating a large spectrum of new genomic tools.

Imputing to full sequence data will allow for direct measures of markers present at QTL. This will allow for more accurate estimation of QTL effects both within and across breeds. Direct measurement of causative mutations affecting phenotypes will also help reduce over-parameterization of genomic selection models, as any marker that is not a QTL of an economically important trait can be removed for the analysis of that trait. With full sequence data there will also be greater persistence of accuracy across generations as recombination will be a less important factor affecting the relationship between a trained marker effect and a phenotype in successive generations.

3.6 Conclusions

Results of the study discussed herein indicate that imputation from lower density 50k SNP panel to the bovine 777k SNP panel is viable for dairy breeds with at least 60 animals genotyped on the high density platform. As reference population size grew, imputation accuracy increased both when imputing from 6k and 50k to 777k. For breeds with fewer animals genotyped, combining information from other breeds can modestly improve imputation accuracy from the 50k panel to the 777k panel but not when imputing from 6k to 777k. FImpute was more efficient and more accurate than Beagle to impute high density genotypes both from 6k and 50k panels. A two-step method for imputing from 6k to 777k was more accurate and more efficient as well. This could have significant implications when imputing to even higher densities in the future as a multi-step approach may need to be developed to impute efficiently. Imputation accuracy was significantly higher in all scenarios when comparing animals with genotyped sires in the reference population to those whose sires are ungenotyped or not included in the reference population.

Table 3.1. Number of imputation animals with genotyped ancestors in the reference population.

	Holstein (n=223)	Ayrshire (n=164)	Guernsey (n=19)
Sire	83	84	0
Dam	25	0	0
Maternal Grand Sire[1]	21	13	0
Paternal Grand Sire[1]	60	54	0

[1]All animals with genotyped maternal grand sire and paternal grand sire also have sire genotyped

Table 3.2. Imputation accuracy for family and population (Fam+Pop) vs. population only based (Pop) imputation for 3 breeds using FImpute.

Imputation Scenario	Missing[1] (%)	Imputed[2] (%)	Correct Call[3] (%)	Incorrect Call[4] (%)	Accuracy[5]
Guernsey (Pop)	0.08	99.92	97.18	2.74	0.97
Guernsey (Fam+Pop)	0.08	99.92	97.18	2.74	0.97
Ayrshire (Pop)	0.01	99.99	97.98	2.01	0.98
Ayrshire (Fam+Pop)	0.02	99.99	98.00	1.99	0.98
Holstein (Pop)	0.00	100.00	99.23	0.77	0.99
Holstein (Fam+Pop)	0.00	100.00	99.23	0.76	0.99

[1] SNP that are left as missing after imputation due to unobserved haplotypes in the reference population
[2] SNP that are filled in after the imputation process
[3] SNP whose genotype call after imputation matches their high density genotype
[4] SNP whose genotype call after imputation do not match their high density genotype
[5] proportion of correct calls within imputed SNP

Table 3.3. Imputation accuracy for population based imputation using FImpute and Beagle software for 3 breeds.

Imputation Scenario	Missing[1] (%)	Imputed[2] (%)	Correct Call[3] (%)	Incorrect Call[4] (%)	Accuracy[5]
Guernsey Beagle	0.00	100.00	95.37	4.63	0.95
Guernsey FImpute	0.08	99.92	97.18	2.74	0.97
Ayrshire Beagle	0.00	100.00	97.16	2.84	0.97
Ayrshire FImpute	0.01	99.99	97.99	2.01	0.98
Holstein Beagle	0.00	100.00	99.30	0.70	0.99
Holstein FImpute	0.00	100.00	99.23	0.77	0.99

[1] SNP that are left as missing after imputation due to unobserved haplotypes in the reference population
[2] SNP that are filled in after the imputation process
[3] SNP whose genotype call after imputation matches their high density genotype
[4] SNP whose genotype call after imputation do not match their high density genotype
[5] proportion of correct calls within imputed SNP

Table 3.4. Imputation accuracy for FImpute and Beagle using single and multi-breed reference populations.

Imputation Scenario	Missing[1] (%)	Imputed[2] (%)	Correct Call[3] (%)	Incorrect Call[4](%)	Accuracy[5]
Guernsey Beagle (Single)	0.00	100.00	95.37	4.63	0.95
Guernsey Beagle (Multi-breed)	0.00	100.00	96.85	3.15	0.97
Guernsey FImpute (Single)	0.08	99.92	97.18	2.74	0.97
Guernsey FImpute (Multi-breed)	0.00	100.00	97.42	2.58	0.97
Ayrshire Beagle (Single)	0.00	100.00	97.16	2.84	0.97
Ayrshire Beagle (Multi-breed)	0.00	100.00	97.72	2.28	0.98
Ayrshire FImpute (Single)	0.01	99.99	98.00	1.99	0.98
Ayrshire FImpute (Multi-breed)	0.00	100.00	98.23	1.77	0.98
Holstein Beagle (Single)	0.00	100.00	99.30	0.70	0.99
Holstein Beagle (Multi-breed)	0.00	100.00	99.28	0.72	0.99
Holstein FImpute (Single)	0.00	100.00	99.23	0.76	0.99
Holstein FImpute (Multi-breed)	0.00	100.00	99.22	0.77	0.99

[1] SNP that are left as missing after imputation due to unobserved haplotypes in the reference population
[2] SNP that are filled in after the imputation process
[3] SNP whose genotype call after imputation matches their high density genotype
[4] SNP whose genotype call after imputation do not match their high density genotype
[5] proportion of correct calls within imputed SNP

Table 3.5. Imputation accuracy from 6k to HD (777k) for 3 breeds using one and two step imputation procedures as well as single and multi-breed reference populations.

Imputation Scenario	Missing[1] (%)	Imputed[2] (%)	Correct Call[3] (%)	Incorrect Call[4] (%)	Accuracy[5]
Guernsey 1-Step (Multi-breed)	0.00	100.00	88.92	11.08	0.89
Guernsey 1-Step (Single)	0.01	99.99	91.89	8.10	0.92
Guernsey 2-Step (Multi-breed)	0.00	100.00	91.96	8.04	0.92
Guernsey 2-Step (Single)	0.01	99.99	93.21	6.77	0.93
Ayrshire 1-Step (Multi-breed)	0.01	99.99	93.96	6.03	0.94
Ayrshire 1-Step (Single)	0.04	99.96	94.81	5.15	0.95
Ayrshire 2-Step (Multi-breed)	0.02	99.98	94.42	5.56	0.94
Ayrshire 2-Step (Single)	0.03	99.97	94.71	5.26	0.95
Holstein 1-Step (Multi-breed)	0.00	100.00	96.92	3.07	0.97
Holstein 1-Step (Single)	0.00	100.00	97.11	2.89	0.97
Holstein 2-Step (Multi-breed)	0.00	100.00	97.29	2.71	0.97
Holstein 2-Step (Single)	0.00	100.00	97.37	2.63	0.97

[1] SNP that are left as missing after imputation due to unobserved haplotypes in the reference population
[2] SNP that are filled in after the imputation process
[3] SNP whose genotype call after imputation matches their high density genotype
[4] SNP whose genotype call after imputation do not match their high density genotype
[5] proportion of correct calls within imputed SNP

Table 3.6. Accuracy of imputation when imputing randomly or using Holsteins only (across breed) in the reference population.

Imputation Scenario	Missing[1] (%)	Imputed[2] (%)	Correct Call[3] (%)	Incorrect Call[4] (%)	Accuracy[5]
Guernsey - Random imputation	N/A	N/A	64.89	N/A	N/A
Guernsey - Holstein reference	0.01	99.99	75.19	24.80	0.75
Ayrshire - Random imputation	N/A	N/A	62.22	N/A	N/A
Ayrshire - Holstein reference	0.01	99.99	77.89	22.11	0.78

[1] SNP that are left as missing after imputation due to unobserved haplotypes in the reference population
[2] SNP that are filled in after the imputation process
[3] SNP whose genotype call after imputation matches their high density genotype
[4] SNP whose genotype call after imputation do not match their high density genotype
[5] proportion of correct calls within imputed SNP

Table 3.7. Effect of having a genotyped parent (sire) in the reference population in different scenarios[1] in the Ayrshire breed with p-values from ANOVA comparing accuracy of imputation between animals with and without genotyped sires.

		Sire Genotyped (n=84)		Sire Not Genotyped (n=80)				
		Correct Call[3] (%)	Acc[4]	Correct Call[3] (%)	Acc[4]	Δ Correct Call[3] (%)	Δ Acc[4]	p-value
FImpute	Fam+Pop[2]	98.34	0.983	97.92	0.979	0.42	0.004	0.044
	Pop[2]	98.36	0.983	97.81	0.978	0.55	0.005	0.004
Beagle	Pop[2]	97.61	0.976	96.88	0.969	0.73	0.007	0.005
FImpute	Fam+Pop (Multi-breed)[2]	98.56	0.986	98.13	0.982	0.43	0.004	0.039
	Pop (Multi-breed)[2]	98.47	0.985	97.88	0.979	0.59	0.006	0.003
Beagle	Pop (Multi-breed)[2]	98.12	0.981	97.42	0.974	0.70	0.007	0.001
FImpute	6k 1-step[5]	95.78	0.958	94.12	0.941	1.66	0.017	0.002
	6k 2-step[5]	95.65	0.957	94.09	0.941	1.55	0.016	0.002
	6k 1-step (Multi-breed)[5]	95.14	0.952	92.98	0.930	2.16	0.022	<0.001
	6k 1-step (Multi-breed)[5]	95.52	0.956	93.54	0.936	1.98	0.020	<0.001

[1]Multi-breed – All three breeds in the multi-breed reference population, Fam+Pop – both family and population based imputation algorithms were used, Pop – Pedigree information was excluded and only population based imputation was carried out,
6k 1-step – imputation was carried out directly from 6k to 777k, 6k 2-step – imputation was carried out from 6k to 50k and then from 50k to 777k.
[2]Imputed from 50k to HD
[3] SNP whose genotype call after imputation matches their high density genotype
[4] proportion of correct calls within imputed SNP
[5]Imputation for 6k was carried out using a combination of family and population imputation

Table 3.8. Total and per chromosome imputing time[1] for all scenarios.[2]

		Computing Time	# of Jobs[3]	Time/Chr
Guernsey	Fam+Pop	0:01:21	10	0:00:26
	Pop	0:01:04 (0:09:25)	10 (10)	0:00:21 (0:03:02)
	Multi-breed	0:37:46 (36:47:10)	10 (3)	0:12:11 (3:48:20)
	Holstein reference	0:22:37	10	0:07:18
	6k 1-step	0:01:41	10	0:00:33
	6k 2-step	0:01:15	10	0:00:24
	6k 1-step (Multi-breed)	1:03:26	10	0:20:28
	6k 2-step (Multi-breed)	0:38:58	10	0:12:34
Ayrshire	Fam+Pop	0:06:41	10	0:02:09
	Pop	0:06:29 (2:38:04)	10 (5)	0:02:05 (0:25:30)
	Multi-breed	0:37:46 (36:47:10)	10 (3)	0:12:11 (3:48:20)
	Holstein reference	0:22:37	10	0:07:18
	6k 1-step	0:11:12	10	0:03:37
	6k 2-Step	0:06:25	10	0:02:04
	6k 1-step (Multi-breed)	1:03:26	10	0:20:28
	6k 2-Step (Multi-breed)	0:38:58	10	0:12:34
Holstein	Fam+Pop	0:26:10	10	0:08:26
	Pop	0:21:47 (10:26:19)	10 (5)	0:07:02 (1:41:01)
	Multi-breed	0:37:46 (36:47:10)	10 (3)	0:12:11 (3:48:20)
	6k 1-step	0:49:18	10	0:15:54
	6k 2-Step	0:29:39	10	0:09:34
	6k 1-step (Multi-breed)	1:03:26	10	0:20:28
	6k 2-Step (Multi-breed)	0:38:58	10	0:12:34

[1]Time for Beagle is in brackets

[2]Multi-breed – All three breeds in the multi-breed reference population, Holstein reference – exclusively Holsteins were used in a reference population for Guernseys or Ayrshires, Fam+Pop – both family and population based imputation algorithms were used, Pop – Pedigree information was excluded and only population based imputation was carried out, 6k 1-step – imputation was carried out directly from 6k to 777k, 6k 2-step – imputation was carried out from 6k to 50k and then from 50k to 777k

[3]Number of jobs for Beagle in brackets

Table 3.9. Accuracy of imputation from 6k to 50k during the first step of 2 step imputation from 6k to HD (777k).

	Missing[2] (%)	Imputed[3] (%)	Correct Call[4] (%)	Incorrect Call[5] (%)	Accuracy[6]
Ayrshire	0.01	99.99	95.22	4.77	0.95
Ayrshire (Multi-breed)[1]	0.00	100.00	95.24	4.76	0.95
Guernsey	0.01	99.99	94.12	5.86	0.94
Guernsey (Multi-breed)	0.00	100.00	93.40	6.61	0.93
Holstein	0.00	100.00	97.76	2.24	0.98
Holstein (Multi-breed)	0.00	100.00	97.70	2.30	0.98

[1]Multi-breed – All three breeds in the multi-breed reference population
[2] SNP that are left as missing after imputation due to unobserved haplotypes in the reference population
[3] SNP that are filled in after the imputation process
[4] SNP whose genotype call after imputation matches their high density genotype
[5] SNP whose genotype call after imputation do not match their high density genotype
[6] proportion of correct calls within imputed SNP

CHAPTER 4

GENERAL DISCUSSION, CONCLUSIONS, AND IMPLICATIONS:

4.1. Main Research Findings:

1. Linkage disequilibrium levels in all breeds, as measured in the least biased manner possible (r^2), were found to be consistent with those found in other studies at average pair-wise distances at the density measured on the 50k SNP panel, and were high enough for genomic selection to take place.
2. When LD was measured on a higher density genotyping panel (777k), r^2 measures were found to be significantly higher, and should contribute to a greater degree of accuracy of genomic selection.
3. Linkage phase was found to be extremely consistent between breeds at short distances measured on the high density panel. Markers should also be consistent in phase with QTL at these distances.
4. Imputation was carried out accurately (>92%) from both 6k and 50k to the high density (777k) panel within breed and using a multi-breed reference population for Holsteins, Ayrshires and Guernseys.
5. Some marker haplotypes were conserved at the density observed on the 50k panel, allowing for increased accuracy of imputation when a multi-breed reference population was considered. At densities examined, imputing from 6k upwards, there was no increase in accuracy, most likely because there were not consistent haplotypes across breeds.

4.2. General Discussion

Linkage disequilibrium measures give a strong indication of the degree to which imputation will be successful. As marker density increases between the low density panel used for imputation to high density from 6k to 50k, the degree to which markers are linked to one another increases. This increase in linkage between markers allows for more accurate prediction of the marker alleles that are present at one locus, given the marker alleles present at the nearest locus, or several loci surrounding the marker locus to be predicted by imputation. Also, as we increase marker density, there is a much greater likelihood that markers will be in the same gametic phase across breeds. This leads to more information being useful across breeds, and leads to the modest increases in accuracy seen when imputing using a multi-breed population from the 50k to 777k compared to a decrease in accuracy when imputing from a lower density (6k) where gametic phase correlations across breeds are low.

With a high degree of correlation between breeds at high density (777k) as well as high imputation accuracies within-breed and using a multi-breed reference population comprised of the 3 breeds studied, strength is added to the hypothesis that multi-breed genomic selection is a distinct possibility to increase accuracy for breeds with smaller populations of genotyped individuals. A high correlation of gametic phase at short distances combined with small increases in accuracy when imputing in a multi-breed population shows that there are consistent haplotypes across breeds that will be useful in the training of SNP effects for genomic selection. Given that imputation accuracies are high in all scenarios studied in this thesis, both from 6k and 50k to the 777k SNP panel, the reference population for genomic selection with the 777k panel can be grown in a

manner that increases accuracy while decreasing costs to both producers as well as the dairy industry when compared to genotyping a large number of animals using the 777k panel.

As even higher density panels and/or genome sequences become more common, the 777k panel may become used as a low density panel to impute to these even higher density panels/sequences. As linkage disequilibrium measures are extremely high on this panel, imputation from this panel to higher density panels should be extremely accurate. Imputation of this type in breeds with few genotyped individuals will likely gain significantly from a multi-breed reference population as gametic phase is high between all breed pairs at the marker density found on the 777k panel and so haplotypes will be highly conserved between breeds making genotypes from each breed more informative for each other breed.

4.3. Suggestions for Future Studies

The results of the studies included in this thesis give strength to the hypothesis that genomic selection becomes possible across breeds by utilizing the high density genotyping panel. To confirm this hypothesis, a follow up genomic validation study must be carried out to determine accuracy of genomic predictions using a reverse validation as carried out by Liu et al. (2011). To be able to perform a validation of this nature, however, methods must be created to allow for genomic selection to be carried out using such a large number of markers. This potentially involves the use of new methods to calculate genomic breeding values, or implementing certain methods to select only SNPs with significant effects for each trait to reduce the number of parameters to be estimated, such as those described by Shah and Kusiak (2004). An assessment of which breed pairs will truly benefit the most from one another will also be necessary, although the results of this thesis imply that outside of the breed pair of Holstein and Ayrshire, all other breed pairs are nearly identical in terms of overall gametic phase consistency, which is high in all instances.

There are not a large number of studies investigating imputation from lower density panels to the high density panels. Similar studies to this one need to be carried out in other dairy and beef breeds to ensure that the trends in imputation accuracy found are consistent for other populations that may be less related than the ones studies here.

Imputation studies need also to be carried out from all densities to genome sequence as more individuals become fully sequenced. There have been multiple initiatives to increase the number of sequenced animals around the world recently. If

animals can effectively be imputed to entire genome sequence it will be much easier to identify causal variants effecting phenotypes and select directly on QTL.

4.4. Conclusions

As linkage disequilibrium measures are significantly higher at a greater density, with a large enough reference population to avoid the problems associated with over-parameterization, more accurate genomic breeding values can be predicted using a higher density SNP panel. Furthermore, the high consistency of gametic phase across breeds at high marker density should allow for multi-breed genomic evaluations to be carried out more accurately. This shall lead to an increase in accuracy of genomic selection, especially for those breeds with fewer genotyped individuals. This conclusion assumes that markers are also in consistent phase with QTL, and that these QTL are similar in size of the effect on a given trait phenotype across breeds.

High imputation accuracy in all breeds from both 6k and 50k genotyping panels leads to the conclusion that a large reference population for implementing genomic selection with the 777k SNP panel can be created in a manner that is more economically feasible for the entire dairy industry. It is also concluded that there are conserved haplotypes between breeds as seen from increased accuracy from a multi-breed reference population when imputing from 50k to 777k.

REFERENCES

Ardlie, K., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews: Genetics, 3*(4), 299-309.

Arias, J., Keehan, M., Fisher, P., Coppieters, W., & Spelman, R. (2009). A high density linkage map of the bovine genome. *BMC Genetics, 10*(1), 18-29.

Axenovich, T. I. (1996). Prediction of linkage phase by parental phenotypes. *Genetic Epidemiology, 13*(3), 271-283.

Berry, D. P., & Kearney, J. F. (2011). Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. *Animal, 5*(08), 1162-1169.

Blott, S. C., Williams, J. L., & Haley, C. S. (1998). Genetic relationships among european cattle breeds. *Animal Genetics, 29*(4), 273-282.

Boichard, D., Guillaume, F., Baur, A., Croiseau, P., Rossignol, M. N., Boscher, M. Y., et al. (2012). Genomic selection in french dairy cattle. *Animal Production Science, 52*(3), 115-120.

Bradley, D.G., Lothus, R.T., Cunningham, P., & MacHugh, D.E. (1998). Genetics and domestic cattle origins. *Evolutionary Anthropology, 6*, 79-86.

Brito, F., Braccini, J., Sargolzaei, M., Cobuci, J., & Schenkel, F. (2011). Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. *BMC Genetics, 12*(1), 80-89.

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

Browning, B. L., & Browning, S. R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *The American Journal of Human Genetics, 84*(2), 210-223.

Calus, M. P. L. (2010). Genomic breeding value prediction: Methods and procedures. *Animal, 4*(2), 157-164.

Calus, M. P. L., Veerkamp, R. F., & Mulder, H. A. (2011). Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework. *Journal of Animal Science, 89*(7), 2042-2049.

Calus, M. (2008). Accuracy of genomic selection using different methods to define haplotypes. *Genetics, 178*, 553.

Chen, Y., Lin, C., & Sabatti, C. (2006). Volume measures for linkage disequilibrium. *BMC Genetics, 7*(1), 54-61.

Daetwyler, H. D., Wiggans, G. R., Hayes, B. J., Woolliams, J. A., & Goddard, M. E. (2011). Imputation of missing genotypes from sparse to high density using long-range phasing. *Genetics, 189*(1), 317-327.

de Roos, A. P. W., Hayes, B. J., Spelman, R. J., & Goddard, M. E. (2008). Linkage disequilibrium and persistence of phase in holstein-friesian, jersey and angus cattle. *Genetics, 179*(3), 1503-1512.

Dekkers, J. C. M., & Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews. Genetics, 3*(1), 22-32.

Druet, T., Schrooten, C., & de Roos, A. P. W. (2010). Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *Journal of Dairy Science, 93*(11), 5443-5454.

Du, F. X., Clutter, A. C., & Lohuis, M. M. (2007). Characterizing linkage disequilibrium in pig populations. *International Journal of Biological Sciences, 3*(3), 166-178.

Farnir, F., Coppieters, W., Arranz, J. J., Berzi, P., Cambisano, N., Grisart, B., et al. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome Res, 10*, 220-227.

Gautier, M., Faraut, T., Moazami-Goudarzi, K., Navratil, V., Foglio, M., Grohs, C., et al. (2007). Genetic and haplotypic structure in 14 european and african cattle breeds. *Genetics, 177*, 1059-1070.

Goddard, M., Hayes, B., Mcpartlan, H., & Chamberlain, A. (August 13-18 2006). Can the same genetic markers be used in multiple breeds? *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brazil, CD-ROM Communication no. 22-16*

Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics, 124*(6), 323-330.

Goddard, M. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica, 136(2),* 245-257.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009c). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science, 92*(2), 433-443.

Hayes, B. J., Chamberlain, A. J., Maceachern, S., Savin, K., McPartlan, H., MacLeod, I., et al. (2009a). A genome map of divergent artificial selection between *bos taurus* dairy cattle and *bos taurus* beef cattle. *Animal Genetics, 40*(2), 176-184.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., Savin, K., van Tassell, C. P., Sonstegard, T. S., et al. (2009b). A validated genome wide association study to breed cattle adapted to an environment altered by climate change. *PLoS ONE, 4*(8), e6676.

Hayes, B. J., Visscher, P. M., McPartlan, H. C., & Goddard, M. E. (2003). Novel multilocus measure of linkage disequilibrium to estimate past effective population size. *Genome Research, 13*(4), 635-643.

Hickey, J., Kinghorn, B., Tier, B., Wilson, J., Dunstan, N., & van der Werf, J. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. *Genetics Selection Evolution, 1,* 43-55.

Hill, W.G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical Applied Genetics, 38,* 226-231.

Johnston, J., Kistemaker, G., Sullivan, P.G. (2011). Comparison of different imputation methods. *Interbull Open Meeting. Stavanger, Norway.*

Laird, N. M., & Lange, C. (2011). The general concepts of gene mapping: Linkage, association, linkage disequilibrium and marker maps *Statistics for Biology and Health,* 67-86.

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics, 165*(4), 2213-2233.

Liu, Z., Seefried, F. R., Reinhardt, F., Rensing, S., Thaller, G., & Reents, R. (2011). Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. *Genetics Selection Evolution, 43*, 19.

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., & Meuwissen, T. H. E. (November 2009). The accuracy of genomic selection in norwegian red cattle assessed by cross-validation. *Genetics, 183*(3), 1119-1126.

Lush, J.L., Holbert, J.C., & Willham, O.S. (1936). Genetic history of the Holstein-Fresian cattle in the United States. *Jounal of Heredity, 27*, 61-72.

Lynch, M. and Walsh, B. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Inc., Massachusetts, USA. pp 98-99.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics, 11,* 499-511.

Maudet, C., Luikart, G., & Taberlet, P. (2002). Genetic diversity and assignment tests among seven french cattle breeds based on microsatellite DNA analysis. *Journal of Animal Science, 80*(4), 942-950.

Mc Hugh, N., Meuwissen, T. H. E., Cromie, A. R., & Sonesson, A. K. (2011). Use of female information in dairy cattle genomic breeding programs. *Journal of Dairy Science, 94*(8), 4109-4118.

McKay, S. D., Schnabel, R. D., Murdoch, B. M., Matukumalli, L. K., Aerts, J., Coppieters, W., et al. (2007). Whole genome linkage disequilibrium maps in cattle. *BMC Genet, 8*, 74-85.

McKay, S., Schnabel, R., Murdoch, B., Matukumalli, L., Aerts, J., Coppieters, W., et al. (2008). An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. *BMC Genetics, 9*(1), 37-45.

Meadows, J., Chan, E., & Kijas, J. (2008). Linkage disequilibrium compared between five populations of domestic sheep. *BMC Genetics, 9*(1), 61-70.

Meuwissen, T., Hayes, B., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics, 157*(4), 1819-1829.

Mohlke, K. L., Lange, E. M., Valle, T. T., Ghosh, S., Magnuson, V. L., Silander, K., et al. (2001). Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in finns. *Genome Research, 11*(7), 1221-1226.

Moser, G., Tier, B., Crump, R., Khatkar, M., & Raadsma, H. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genetics Selection Evolution, 41*(1), 56-71.

Nagamine, Y., Nirasawa, K., Takahashi, H., Sasaki, O., Ishii, K., Minezawa, M., et al. (2008). Estimation of the time of divergence between japanese mishima island cattle

and other cattle populations using microsatellite DNA markers. *Journal of Heredity, 99*(2), 202-207.

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. and Franke, A. (2009). A comprehensive evaluation of SNP genotype imputation. *Human Genetics, 125 (2).* 163-171.

Pryce, J. E., Gredler, B., Bolormaa, S., Bowman, P. J., Egger-Danner, C., Fuerst, C., et al. (2011). Short communication: Genomic selection using a multi-breed, across-country reference population. *Journal of Dairy Science, 94*(5), 2625-2630.

Pszczola, M., Strabel, T., Mulder, H. A., & Calus, M. P. L. (2012). Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of Dairy Science, 95*(1), 389-400.

Rafalski, A., & Morgante, M. (2004). Corn and humans: Recombination and linkage disequilibrium in two genomes of similar size. *Trends in Genetics, 20*(2), 103-111.

Ritz, L.R., Glowatzki-Mullis, M., MacHugh, D.E., & Gaillard, C. (2002). Phylogenetic analysis of the tribe Bovini using microsatellites. *Animal Genetics, 31*(3), 178-185.

Roche, J.R., Macdonald, K. A., Burke, C. R., Lee, J. M., & Berry, D. P. (2007). Associations among body condition score, body weight, and reproductive performance in seasonal-calving dairy cattle. *Journal of Dairy Science, 90*(1), 376-391.

Rolf, M. M, McKay, S.D., Mcclure, M.C., Decker, J.E., Taxis, T.M., Chapple, R.H., Vasco, D.A., Gregg, S.J., Kim, J.W., Scnabel, R.D., & Taylor, J.F.(2010). How the next

generation of genetic technologies will impact beef cattle selection. *Beef Improvement Federation Research Symposium & Annual meeting. Colombia, Missouri, USA.*

Sargolzaei, M., Chesnais, J., & Schenkel, F.S. (2011). Accuracy of imputed 50k genotypes from 3k and 6k chips in dairy cattle breeds using FImpute. *Plant and Animal Genome XX. San Diego, California, USA.*

Sargolzaei, M., Chenais, J.P., Schenkel, F.S. (2010). Accuracy of a family-based genotype imputation algorithm. *GEB Open Industry Session. Saint-Hyacynthe, Quebec, Canada.*

Sargolzaei, M., Schenkel, F. S., Jansen, G. B., & Schaeffer, L. R. (2008). Extent of linkage disequilibrium in holstein cattle in north america. *Journal of Dairy Science, 91*(5), 2106-2117.

Schaeffer, L. R. (2006). Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics, 123*(4), 218-223.

Sebastiani, P., & Abad-Grau, M. (2007). Bayesian estimates of linkage disequilibrium. *BMC Genetics, 8*(1), 36-48.

Shah, S. C., & Kusiak, A. (2004). Data mining and genetic algorithm based gene/SNP selection. *Artificial Intelligence in Medicine, 31*(3), 183-196.

Sutter, N. B., Eberle, M. A., Parker, H. G., Pullar, B. J., Kirkness, E. F., Kruglyak, L., et al. (2004). Extensive and breed-specific linkage disequilibrium in canis familiaris. *Genome Res, 14*, 2388-2396.

Stachowicz, K., Sargolzaei, M., Miglior, F., & Schenkel, F.S. (2011). Rates of inbreeding and genetic diversity in canadian holstein and jersey cattle. *Journal of Dairy Science,94*(10), 5160-5175.

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology, 2*, 125-141.

Thaller, G., Krämer, W., Winter, A., Kaupe, B., Erhardt, G., & Fries, R. (2003). Effects of DGAT1 variants on milk production traits in german cattle breeds. *Journal of Animal Science, 81*(8), 1911-1918.

The Bovine HapMap Consortium. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science, 324*(5926), 528-532.

Toosi, A., Fernando, R. L., & Dekkers, J. C. M. (2010). Genomic selection in admixed and crossbred populations. *Journal of Animal Science, 88*(1), 32-46.

VanRaden, P. M., O'Connell, J. R., Wiggans, G. R., & Weigel, K. A. (2011). Genomic evaluations with many more genotypes. *Genetics, Selection, Evolution, 43*(1), 10-20.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., et al. (2009). Invited review: Reliability of genomic predictions for north american holstein bulls. *Journal of Dairy Science, 92*(1), 16-24.

Weigel, K. A., Van Tassell, C. P., O'Connell, J. R., VanRaden, P. M., & Wiggans, G. R. (2010). Prediction of unobserved single nucleotide polymorphism genotypes of

jersey cattle using reference panels and population-based imputation algorithms. *Journal of Dairy Science, 93*(5), 2229-2238.

Weller, J. I., Golik, M., Seroussi, E., Ezra, E., & Ron, M. (2003). Population-wide analysis of a QTL affecting milk-fat production in the israeli holstein population. *Journal of Dairy Science, 86*(6), 2219-2227.

Wiggans, G. R., Cooper, T. A., VanRaden, P. M., Olson, K. M., & Tooker, M. E. (2012). Use of the illumina Bovine3K BeadChip in dairy genomic evaluation. *Journal of Dairy Science, 95*(3), 1552-1558.

Zhang, Z., & Druet, T. (2010). Marker imputation with low-density marker panels in dutch holstein cattle. *Journal of Dairy Science, 93*(11), 5487-5494.

Zhang, Z., Todhunter, R. J., Buckler, E. S., & Van Vleck, L. D. (2007). Technical note: Use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. *Journal of Animal Science, 85*(4), 881-885.