GENOMIC EVALUATIONS FOR SINGLE-BREED, MULTI-BREED, ACROSS-BREED, AND CROSSBREED PREDICTIONS

by

YVETTE STEYN

(Under the Direction of Ignacy Misztal)

ABSTRACT

Combining multiple pure breeds or admixed breeds into one evaluation can be detrimental if the accuracy of prediction for one is lower than within-breed. Prediction accuracy was compared when considering SNP effects as different across breeds (non-shared), or the same in all (shared) for 5 simulated breeds. The non-shared approach prevented changes in accuracy, while the shared method only maintained accuracy when the SNP density was high and effective population size was large. Imputation accuracy for crossbred animals differs depending on reference populations. Different reference populations (crossbreds, Jersey, Holstein, or all combined) for the imputation of Holstein-Jersey crossbred genotypes were compared. The best results were achieved with a crossbred reference population. The accuracy and inflation of indirect genomic predictions (IP) for milk yield were evaluated for Holstein-Jersey crossbred animals. Different reference populations were used to calculate SNP effects – ~80k Holstein (HO), ~40k Jersey (JE), ~22k crossbreds (CROSS), Holstein and Jersey combined (JE_HO), or equal proportions of each pure breed and crossbred animals (MIX). While JE, CROSS, and JE_HO gave the same accuracy (0.50), HO and MIX were slightly lower (0.47 and 0.46). An additional method that used breed proportion in combination with SNP effects based on pure

breeds, had the lowest accuracy (0.32). Inflation was best when using the MIX scenario (1.00), and worst when using HO (0.55). Diversity within 20,990 US Holstein cattle was evaluated by using k-means clustering on the genomic relationship matrix. Each of the 5 clusters were traced back for 10 generations - G0 (oldest) to G10 (youngest) to form 5 families (F1 to F5). Allele frequency changes over time was observed for specific SNP based on different criteria – key genes of known importance, markers associated with time, a population diversity parameter (F_{st}), markers that changed the most in the whole population, and markers that have changed differently across families (based on greatest variance and range). Non-parallel changes were observed across families, showing genetic redundancy and divergent selection. The Replicate Frequency Spectrum (RFS) was used to measure the similarity of change across families. Results show that populations have changed differently, supporting the presence of genetic redundancy.
INDEX WORDS: prediction accuracy, effective population size, imputation, genetic redundancy, pleiotropy, epistasis

GENOMIC EVALUATIONS FOR SINGLE-BREED, MULTI-BREED, ACROSS-BREED, AND CROSSBREED PREDICTIONS

by

YVETTE STEYN

BS, University of Pretoria, South Africa, 2008

BS(Hon), University of Pretoria, South Africa, 2009

MS, University of Pretoria, South Africa, 2012

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

GENOMIC EVALUATIONS FOR SINGLE-BREED, MULTI-BREED, ACROSS-BREED, AND CROSSBREED PREDICTIONS

by

YVETTE STEYN

Major Professor:	Ignacy Misztal
Committee:	Daniela A. Lourenco
	Romdhane Rekaya
	Natascha Vukasinovic
	Thomas J. Lawlor

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2021

DEDICATION

I dedicate this to my retirement fund, which took a beating from a few years of zero contribution

ACKNOWLEDGEMENTS

As is often the case in life, success is due to both hard work and a good stroke of luck. I hit the jackpot when I got to meet and talk to my advisor, Dr Ignacy Misztal, when he presented a course in South Africa. He gave me valuable new knowledge and skills, the opportunity to meet amazing people, and a key to a brighter future – both personally and professionally. Thank you for seeing potential in me, encouraging me to be better, and encouraging all of the students to explore life beyond just work. You arranged social gatherings at your home and other interaction opportunities to make us feel like we are part of a group that can be strong together. You made us feel valued.

Dr Daniela Lourenco has been vital in my development as a student and person. You are an all-rounder – talented teacher, researcher, speaker, writer, mentor, inspiration to all young professionals, and interpreter of what exactly was meant in the project brief. Your door was always open and you explained everything well, starting with basic principles. You love what you do and that rubs off on others. You value each student as an individual with their own backstory, interests, hopes, and dreams.

Dr Romdhane Rekaya has been one of the best and most interesting teachers I have ever had. A 4-hour class (that was supposed to be 75 minutes) did not feel long at all. When you see someone is lost, you'll go off topic to make it clearer. We have possibly learned more from the off-topic discussions than the on-topic ones. Thank you for encouraging us to think, question our own biases, refrain from immediately dismissing ideas of others, and know that we must prepare for change.

Dr Natascha Vukasinovic and Dr Dianelys Gonzalex-Pena were kind, helpful, and the right dose of critical during my time at Zoetis. You made sure I was comfortable at Kalamazoo, collected at the airport and extended stay, had everything I needed, and explored what I could there. You made it an enriching experience and gave me insight in the industry in the USA. Dr Thomas Lawlor gave me different, interesting ideas to explore and enjoy. Our project took numerous turns to become something we were both happy with. Thank you for the passionate interest in this topic, sharing papers I did not notice, taking the time for long discussions, and continuous support. Drs Andres Legarra and Zulma Vitezica have been inspirational people to me. Thank you for being approachable, wise, encouraging, and all the postcards you sent from different places in the world.

Thank you, Dr Ivan Pocrnic, for all the academic and personal support. Without you, I would have been more lost in the beginning, and much more stressed throughout. You and I share a common philosophy – share our experience, knowledge, mistakes, and life tricks with other students/colleagues to reduce their struggle and stress. You never wanted anyone to feel out of place, unwanted, or unprepared. I will always be grateful for that. You were an essential member of our group.

Lastly, Adrian Liversage, my partner. You've been a great support throughout the ups and the downs in life and this endeavor. We've endured long periods of being physically apart across continents, and enjoyed months of being around each other 24/7 in our apartment during a pandemic. We share similar views and plans for the future. You've made dramatic changes in your life to build a new life with me. Thank you for all of it. I am excited to explore more of life with you.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Genomic selection is now applied in multiple species and breeds in various countries. The adoption of these methods has greatly improved the accuracy of selection. The improvement in accuracy is especially pronounced in young animals with no records. This has decreased the generation interval and increased selection intensity. All these factors combined has led to a substantial increase in the rate of genetic change and response to selection. Evaluations are typically performed within a single breed. However, multi-breed evaluations are not uncommon. Multi-breed evaluations are particularly appealing where numerically small breeds want to combine data with others to increase the size of their reference population, which is important for prediction accuracy. Chapter 3 investigates the accuracy of evaluations in a multi-breed context when treating SNP markers as shared, or non-shared among breeds.

Across-breed predictions, where the marker effects calculated using one population are used to indirectly predict genetic merit of animals from a different breed or country, is also desirable to expand benefits of genomic selection to even more sectors of the industry. Across-breed predictions are explored in chapters 3 and 5 in both simulated and real data.

Crossbred dairy animals have generally been excluded from evaluations. The increase in popularity of crossbreeding in dairy has made it crucial to determine how to include these animals in evaluations. Many crossbred animal genotypes contain a small number of SNP markers (such as 3K). In general, the genotypes of purebred animals are imputed from the true SNP chip, to a desired number of markers. Imputation accuracy depends on the size of the

training population, as well as the relationship between said population and the target population. Crossbreeding results in a new population that is different from component pure breeds, thus chapter 4 investigated the accuracy of imputation when using purebred or crossbred animals in the training population. Chapter 5 explores how breeding values can be estimated for crossbred animals using marker effects based on different populations.

When few animals are intensively used within a single breed, inbreeding will increase. This leads to inbreeding depression and reduced genetic diversity. Genetic diversity is important for continuous genetic improvement and adaptation to changing environments. Chapter 6 investigates whether the US Holstein population can be divided into genetically more distinct sub-populations based only on the genomic relationship matrix. The chapter further delves into these different groups, how they have changed, and how it can influence breeding values.

CHAPTER 2

LITERATURE REVIEW

The accuracy of genomic predictions is highly influenced by the reference population. The size of the reference population is a crucial contributor (Hayes and Goddard, 2008). This is a challenge for numerically small breeds that do not have adequate resources. A multi-breed evaluation is a desirable option if these breeds could benefit from the reference population of another (Swan et al., 2014) without disadvantaging the other breeds. Although the importance of a large reference population cannot be overstated, it is only one requirement for the successful application of genomic selection. The relatedness between the reference and validation populations is important (Clark et al., 2012), as well as the relatedness of animals within the reference population (Pszczola et al., 2012). In fact, having a smaller reference population that is more related to the target population can be more advantageous than including more, less related individuals (Neyhart et al., 2017, Van den Berg et al., 2020). Accuracies may be low for specific herds if they do not have animals in the reference population (Hayes et al., 2018). The inclusion, or balance of sexes used in the reference population can further influence its accuracy (Lourenco et al., 2015, Van den Berg et al., 2020). Accuracy decays as the generational distance between reference and target populations increases (Hidalgo et al., 2021, Hollifield et al., 2021). Therefore, the reference population must be updated continuously to maintain optimum accuracy.

Genomic evaluations essentially work by capturing independent chromosome segments (ICS) (Daetwyler et al., 2010). The number of ICS (Me) is related to the effective population size ($Me=N_e4L$ where N_e is the effective population size and L is the chromosome length in

Morgans) (Stam, 1980). More homogenous breeds will have fewer, longer ICS. Therefore, accuracies will be higher within breeds that show greater genetic similarities among animals.

MULTI-BREED EVALUATIONS

The number of markers required to successfully apply genomic selection must be enough to capture the ICS (Pocrnic et al., 2018). Multi-breed evaluations include animals from multiple breeds or populations. These will have many more M_e, and consequently more markers are expected to be required to achieve the same accuracies as within-breed evaluations (Rahimi et al., 2020, Marjanovic and Calus, 2021).

In general, multi-breed studies have shown small changes in accuracy compared to within-breed, as long as the target breed is included in the reference population and marker density is high enough. Results have been inconsistent, not only across studies, but across traits and methods (Erbe et al., 2012, Kemper et al., 2015, Calus et al., 2018, Van den Berg et al., 2019). Overall, small or no differences have been observed with approximately 50k markers or more, even when increasing markers to over 700K (Su et al., 2012, Raymond et al., 2018a). When improvements did occur in multi-breed studies, the numerically smaller breed tended to benefit more (Olson et al., 2012, Hozé et al., 2014, Jónás et al., 2017). However, when the largest breed is considerably larger than the smaller breed, the larger one will dominate and lead to lower accuracies for the smallest breeds (Van den Berg et al., 2020). In a study with three tropical beef breeds, Hayes et al. (2018) compared the scenario where all breeds and herds were represented in the reference population, to the scenario where each herd in turn was removed from the reference population and used for validation. For the first scenario, increasing from 28K to 728K markers resulted in small improvements in accuracy for female fertility, and including multiple breeds in the reference population did not result in higher accuracies than within-breed

predictions. For the second scenario, even within-breed accuracies approached zero. The accuracy increased substantially when marker density was increased, BayesR was applied, and a multi-breed evaluation was used.

Approaches for multi-breed evaluations include a simple, joint relationship matrix (Pryce et al., 2011, Erbe et al., 2012, Olson et al., 2012), a genomic relationship matrix using breed-wise allele frequencies (Makgahlela et al., 2014), accounting for linkage disequilibrium (LD) (Zhou et al., 2014, Rahimi et al., 2020), using two genomic matrices of which one contains the most important markers and the second the remaining markers (Raymond et al., 2018b, Raymond et al., 2020), or treating breeds as different traits (Olson et al., 2012, Calus et al., 2018, Van den Berg et al., 2020). Accounting for breed-wise allele frequencies did not lead to improvement in accuracies. Treating breeds as different, but correlated traits, generally delivered better results than a single joint matrix, but not necessarily a remarkable difference. A disadvantage of this method, is that animals must be classified into a specific pure breed. Thus, crossbred animals cannot be included. While Rahimi et al. (2020) found improvements in accuracy when accounting for LD, Zhou et al. (2014) found that accounting for LD phasing did not improve accuracy beyond what could be achieved by treating the breeds as different traits.

Wientjes et al. (2016) successfully derived an equation to predict the accuracy of genomic values when combining breeds. Results showed that treating the breeds as different traits in a multi-trait evaluation will deliver similar accuracies when the genetic correlation between populations is 1, and better than within-breed evaluations if the correlation deviates from 1. Input parameters included the number of individuals, heritability from each of the populations in the reference population, the genetic correlation between populations, the effective number of chromosome segments across target and reference populations, and the

proportion of the genetic variance in the predicted population captured by the markers in each of the reference populations.

GENETIC CORRELATION BETWEEN POPULATIONS

The additive genetic correlation between populations are typically estimated by treating the same phenotype in the populations as separate, correlated traits in a multi-trait model. The genomic restricted maximum likelihood (GREML) or Bayesian methods can be used to estimate variance components, from which the correlation can be calculated (Karoui et al., 2012). Wientjes et al. (2017) derived a method of compiling the genomic relationship matrix to accurately estimate variance components. This matrix takes the allele frequencies of each population into consideration as follows:

$$\boldsymbol{G} = \begin{bmatrix} \dfrac{\boldsymbol{Z}_1\boldsymbol{Z}_1'}{\sum 2p_{1j}(1-p_{1j})} & \dfrac{\boldsymbol{Z}_1\boldsymbol{Z}_2'}{\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})}} \\ \dfrac{\boldsymbol{Z}_2\boldsymbol{Z}_1'}{\sqrt{\sum 2p_{1j}(1-p_{1j})}\sqrt{\sum 2p_{2j}(1-p_{2j})}} & \dfrac{\boldsymbol{Z}_2\boldsymbol{Z}_2'}{\sum 2p_{2j}(1-p_{2j})} \end{bmatrix}$$

where $\mathbf{Z_1}$ and $\mathbf{Z_2}$ are centered genotypes within populations as $g_{ijm} - 2p_{jm}$, where g_{ijm} is the allele count of individual *i*, from population *m* at locus *j*, and p_{1j} and p_{2j} are the allele frequencies of marker *j* in populations 1 and 2.

Another method to determine the correlation between two populations was introduced by Duenk et al. (2020) and applied between pure- and crossbred populations in a later study (Duenk et al., 2021). This correlates the additive genetic values of animals when expressed in the genetic background of different populations. The additive genetic value of individual *i* for the trait expressed in the population that *i* belongs to (here population 1) is:

$$\boldsymbol{v}_i^{P1} = \boldsymbol{h}'_{a,j}\boldsymbol{\alpha}^{P1}$$

where $\boldsymbol{v}_i^{P1}$ is the additive value of animal *i* when expressed in population 1, $\mathbf{h_{a,j}}$ is a column vector of additive genotypes (measured as allele counts, minus the mean allele count in the population) of individual *i* at the marker, and $\boldsymbol{\alpha}^{\mathbf{P1}}$ is a column vector of average effects at those markers in population 1. The additive genetic value of individual *i* for population 2 is:

$$\boldsymbol{v}_i^{P2} = \boldsymbol{h}'_{a,j}\boldsymbol{\alpha}^{P2}$$

where $\boldsymbol{v}_i^{P2}$ is the additive value of animal *i* when expressed in population 2, $\boldsymbol{\alpha}^{\mathbf{P2}}$ is a column vector of average effects in population 2. The correlation between these two additive genetic values is the genetic correlation between the populations. Essentially, it is the correlation between indirect genomic predictions (IGP) when using SNP effects based on one population, and the IGP of the same animals when using SNP effects based on the other population.

ACROSS-BREED EVALUATIONS

While multi-breed evaluations refer to a scenario where all breeds are represented in the reference population, across-breed evaluations refer to scenarios where the target breed is not included. Thus, information from one breed, or a collection of breeds, is used to predict the genetic merit of animals in a different breed. Across-breed evaluations would be highly desirable if it improves the accuracy of prediction in breeds that do not have enough genotypes to have their own reliable reference population. The same applies for within-breed but across-country populations. Additionally, it can also be advantageous to large breeds with successfully established genomic evaluations if they want to incorporate a novel trait that is well recorded in a different breed.

In contrast to multi-breed evaluations, across-breed genomic prediction accuracies have consistently been poor, although the extent of accuracy varies (Olson et al., 2012, Van den Berg et al., 2016a, Raymond et al., 2018a, Karaman et al., 2021). Within-breed but across country has

been shown to be more successful than across-breed, but results were not compared to within-breed within-in country (Raymond et al., 2018a). Raymond et al. (2018b) managed to apply across-breed predictions more successfully than using a traditional G-matrix by using multiple genomic relationship matrices that allow more weight on pre-selected significant markers. It also accounted for genetic correlations between breeds as described by Wientjes et al. (2017). Breeds that were not highly correlated to each other did not benefit from this approach. Although better than a single traditional G-matrix, accuracies were still much lower than what was achieved using within-breed or multi-breed evaluations.

A recent study by Meuwissen et al. (2021) concluded that the accuracy from their across-breed evaluation was higher for the Jersey breed than a within-breed evaluation, and similar for Holstein. At first, this may seem to be in stark contrast to all previous research. However, the terminology used in their study is different. The reference population included both Holstein and Jersey animals. Thus, based on our definition, this is a multi-breed scenario for the Jersey and Holstein. These results correspond to previous research where small differences are achieved for participating breeds, and the smaller breed tended to benefit more when any are found. When Australian Red was used as validation population (thus, a true across-breed scenario), accuracies were much lower than that of the other breeds but not close to zero (0.17 to 0.34). No comparison was made to a within-breed Australian Red evaluation. Bayesian methods delivered better results, while increasing markers from 600K to WGS marginally improved the accuracy of prediction for the Australian Red.

REASONS FOR LACK OF IMPROVEMENT IN ACCURACY OF MULTI- AND ACROSS-BREED EVALUATIONS

Capturing LD is considerably more important for the accuracy of genomic predictions compared to capturing only Mendelian sampling (Habier et al., 2013, Ling et al., 2021). Due to random recombination during meiosis, the LD will not persist across breeds, even if they originated from the same historical population. An early simulation study by De Roos et al. (2009) showed promising potential for multi-breed evaluations to increase the accuracy of prediction for the participating breeds as long as marker density is high enough to capture LD in all sub-populations. However, this has not been achieved in real data, even with whole-genome sequencing. In fact, having too many markers have a diluting effect on accuracy (Raymond et al., 2018a). However, when only markers near the QTL are used, prediction accuracies of across-breed evaluations can be improved compared to using all markers on the 50K chip, or randomly selected markers, but will still be lower than within-breed evaluations (Erbe et al., 2012, Van den Berg et al., 2016a). Considerable improvements over sequence data were achieved by using as few as 133 significant markers (Raymond et al., 2018a).

Even if LD is fully captured across the populations, QTL properties are important factors affecting the accuracy of selection. The same QTL may have different substitution effects in different populations (Thaller et al., 2003). The predominant factors affecting the correlation of substitution effects across populations are the genetic relatedness between populations, the distribution of allele frequencies at QTL, and various non-additive genetic factors such as dominance, gene-by-gene interactions (epistasis), and genotype by environment interactions (G x E) (Duenk et al., 2020, Legarra et al., 2021). A simulation study by Duenk et al. (2020) showed

that additive genetic correlations between breeds, or populations, below 0.80 are not due to dominance alone. Realistic levels of epistasis decreased this correlation down to 0.45.

When the number of QTL affecting the trait is low (and thus effect size is larger), across-breed evaluations will be more accurate (Van den Berg et al., 2015). When the number of QTL is lower than the number of independent chromosome segments (M_e), Bayesian variable selection models are expected to perform better than GBLUP models (Van den Berg et al., 2015). The minor allele frequency (MAF) of QTL also affects the accuracy of multi-breed evaluations, with accuracies of both within- and multi-breed evaluations decreasing as MAF decreases (Wientjes et al., 2015). This is especially true if the QTL has a large effect. Commercial SNP chips tend to include markers with higher MAF to ensure that markers segregate in a variety of populations (Matukumalli et al., 2009). Many causative QTL may not segregate in all breeds in a multi-breed evaluation (Raven et al., 2014). The greatest potential of sequence data is therefore the ability to capture these low MAF markers and use only prioritized markers in multi- and across-breed evaluation. This has been shown to be more accurate than using all markers (Erbe et al., 2012, Van den Berg et al., 2016a, Raymond et al., 2018a). Imputation errors may hinder this process by not capturing causative QTL, nor markers in close LD with these QTL. The mapping of QTL can be improved by using multi-breed populations, as LD only persists over short regions across breeds, and more ICS are present (Raven et al., 2014, Kemper et al., 2015, Van den Berg et al., 2016b).

CROSSBREED EVALUATIONS

Although across-breed evaluations have been shown to be unsuccessful, including crossbred animals to the purebred reference population has the potential to improve the accuracy of genomic predictions of the component pure breed not included in the reference population.

However, this accuracy is lower than within-breed (Toosi et al. 2010). This was shown in a study by Moghaddar et al. (2014) using genotypes from purebred Merino sheep and Merino crosses – Merino crossed with either Border Leicester, Poll Dorset, or White Suffolk. When the crossbreds were used as a reference population, the accuracy of genomic predictions for the purebreds was higher than an across-breed scenario (using Merino to predict the other breed). Depending on trait, the accuracy of prediction for the Merino breed was highest when using a reference population containing only Merino, or predominantly Merino (>70%). The accuracy of prediction for the crossbred animals was not tested.

Combining crossbred and purebred genotypes has been shown to increase the accuracy of prediction for crossbred animals, both in real pig populations and simulated data (Lourenco et al., 2016). This combination might not improve the accuracy of the purebred pigs, as most of the genomic information is already captured by the purebred animals (Pocrnic et al., 2019). However, in a recent study that simulated a 3-rotational dairy cross, Karaman et al. (2021) found that the accuracy of purebred predictions can be improved with the inclusion of crossbred animals, especially when the breed of origin of alleles are accounted for. Additionally, in pig populations, it has been shown that using crossbred genotypes instead of purebred genotypes can be more advantageous if the focus is on crossbred performance (as opposed to purebred performance), and the correlation between crossbred- and purebred performance is not high (Van Grevenhof and Van Der Werf, 2015). The same conclusion was reached in a recent simulation study by González-Diéguez et al. (2020).

Crossbreeding is a central part of the pig and poultry breeding. While many pure breeds or lines are maintained in breeding programs of nucleus herds, the animal marketed to farmers are crossbred animals that are expected to perform in a commercial environment. This adds the

additional complexity of genotype by environment interactions (G x E) and non-additive genetic factors, such as heterosis and dominance. The same trait measured in purebred and crossbred animals can be considered as different, but correlated traits (Wei and van der Werf, 1994, Wei and van der Werf, 1995). The crossbred model that includes both pure and crossbred animals was proposed by Wei and van der Werf (1994) and extended by Christensen et al. (2014) to accommodate genotypes of both pure- and crossbred animals.

In the beef industry, crossbreeding is often applied to exploit heterosis in rotational and terminal crosses, to use breed complementarity for the creation of synthetic breeds, or to transform a herd from one breed to another (upgrading). Conversely, crossbreeding was relatively uncommon in dairy cattle, with New Zealand being an exception. According to the New Zealand Dairy Statistics, almost 50% of registered dairy cattle during the period 2019 to 2020 were crossbreds. Common reasons breeders want these crossbreds are hybrid vigor, and the combination of desirable traits from component breeds (LIC and DairyNZ, 2020). The improvement in fertility, health, and longevity could offset a loss of production that producers may incur by crossbreeding. (Buckley et al., 2014). Crossbreeding elsewhere, including the US, has become increasingly popular. Reasons include the improvement of fertility and health, changes in consumer demand for different milk products, and the promotion of additional European breeds (VanRaden et al., 2020). An analysis of trends in the breed composition of US Dairy Herd Improvement herds showed that the percentage of dairy cattle reported as crossbred increased from 0.1% to 5.3% from 1990 to 2018 (Guinan et al., 2019). By May 2019, the total number of genotyped dairy cattle in the US exceeded 3 million, of which about 2% were crossbred cattle (VanRaden et al., 2020).

Adding crossbred animals in a traditional pedigree BLUP evaluation improved accuracy of pure- and crossbreds since more phenotypic information could contribute to animals' breeding values through pedigree connections (VanRaden et al., 2007). However, realized relationships captured by the genomic relationship matrix connects all animals to each other. Genomic evaluations in US dairy initially excluded genotypes of crossbred animals. This was because the marker effects estimated within a pure breed may not be appropriate, parent averages are incomplete or incorrect if the breeding value of both parents are not on the same scale, and the imputation accuracy of crossbred genotypes was low (VanRaden et al., 2020). In 2019, the national genomic evaluation in the US extended services to provide genomic values for crossbred animals (Wiggans et al., 2019). VanRaden et al. (2020) successfully estimated breeding values by using breed proportions based on genomic information, and breed-specific SNP effects. Karaman et al. (2021) used genotypes of three dairy breeds for the simulation of a 3-way cross, and breed of origin instead of breed proportion. Combining all breeds and crosses in the same evaluation gave higher accuracies than indirect predictions for crossbred selection candidates with breed specific SNP effects and breed origin of alleles. However, a multibreed evaluation that included crossbreds in the reference population and accounted for breed origin of allele, gave the highest accuracies.

A dairy crossbred system that has become a hot topic over the last decade, is beef-on-dairy. Heifers that are not preferred for the next generation of replacements, can be mated to beef bulls. This will allow the female to have a lactation while producing a calf more suited to beef production. This increases the income generated from the sale of surplus calves. A dairy-beef index was derived to enable breeders in Ireland to select beef bulls that will maximize the value of the calves while maximizing profit from the lactating female (Berry et al., 2019).

Large populations containing multiple different breeds and crosses creates computational challenges. Genomic evaluations require the inverse of the genomic relationship matrix (**G**). Unlike the traditional relationship matrix (**A**), this matrix only contains non-zero elements and must be inverted directly. Once the matrix becomes large, it can be impossible to invert directly with currently available computational resources. Even if it can be inverted, it could take more time than what is practically acceptable for routine evaluations. Two approaches have been developed to overcome this challenge.

A commonly used method to approximate this inverse, is the Algorithm for Proven and Young (APY) (Misztal et al., 2014). This requires the partitioning of the **G** into blocks corresponding to core (c) and non-core (n) animals:

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_{cc} & \boldsymbol{G}_{cn} \\ \boldsymbol{G}_{nc} & \boldsymbol{G}_{nn} \end{bmatrix}$$

The inverse with APY was then obtained with:

$$\boldsymbol{G}_{APY}^{-1} = \begin{bmatrix} \boldsymbol{G}_{cc}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\boldsymbol{G}_{cc}^{-1}\boldsymbol{G}_{cn} \\ \mathbf{I} \end{bmatrix} \boldsymbol{M}_{nn}^{-1} \begin{bmatrix} -\boldsymbol{G}_{nc}\boldsymbol{G}_{cc}^{-1} & \mathbf{I} \end{bmatrix}$$

where $\mathbf{M}_{nn} = \text{diag}(m_{nn,i}) = \text{diag}(g_{ii} - \mathbf{g}_{ic}\mathbf{G}_{cc}^{-1}\mathbf{g}_{ci})$, g_{ii} is the diagonal element of $\mathbf{G}_{nn}$ for non-core animal *i*, and $\mathbf{g}_{ic}$ is a vector of the genomic relationship of non-core animal *i* with all core animals. The selection of core animals is important. While initially the core animals were recommended to be only proven animals, it has been shown that randomly selecting the core animals is better (Fragomeni et al., 2015, Bradford et al., 2017) as long as enough animals are included in the core. A study on multiple sheep breeds in New Zealand also showed that a randomly selected core was best, both in terms of accuracy and bias. However, it was still important for all breeds to be part of the core (Nilforooshan and Lee, 2019). The size of the core should at least be equal to the number of eigenvalues that explain 98% of the variation in the **G**

matrix (Pocrnic et al., 2016a, Pocrnic et al., 2016b). This may be only a small fraction of animals, especially when including only one breed.

However, the dimensionality will be higher for populations that include many breeds. The M_e would be much higher and LD will be poorer. If the number of eigenvalues that explain 98% of the variance is too large, even the inversion of only core animals would not be possible, or require too much time. Mäntysaari et al. (2017) developed a different method of inverting **G** by using a T matrix. This obtains the exact inverse instead of an approximation while greatly reducing computational cost. It was successfully applied on a large Irish population with 41 different breeds (pure and crosses).

WITHIN-BREED POPULATIONS

Genomic evaluations are typically applied within-breed. In contrast to multi-breed evaluations, within-breed populations are more homogeneous, with greater levels of inbreeding. Fewer, larger M_e are expected to be present and N_e will be lower. Thus, the number of markers required to capture LD is lower. Since capturing LD is the main component to the accuracy of genomic evaluations (Habier et al., 2013), this is a great advantage. However, the lack of genetic diversity can be detrimental for long term improvement.

There can still be distinct sub-populations within a breed, especially when they have been separated for many generations. The founder effect can occur from a bottleneck where a subset of a population is isolated from another (Mayr, 1954). This is the case when animals of an existing breed are imported to establish a local population in a different country. This founder population can undergo considerable genetic differentiation from the population of origin due to reduced genetic variation, changes in allele frequencies, and genetic drift. In fact, the resulting genetic changes over time can even lead to speciation (Templeton, 1980). Additionally, breeding

objectives may differ between countries, leading to selection pressure on different subsets of genes. Upgrading mating strategies further incorporate different genetics into the new local population. Therefore, prediction accuracy may be low for some, or all groups, if animals of all countries are combined in the same population.

Genetic variance is expected to reduce within a population over generations of selection. This is referred to as the Bulmer effect (Bulmer, 1971). Strong selection on traits that are highly heritable, will lead to better LD and a greater reduction in genetic variance (Walsh and Lynch, 2018). This was observed in pigs where the reduction in heritability was greater for growth traits than fitness traits (Hidalgo et al., 2021). While strong selection can produce a desirable uniform product, the reduction in genetic variation can be concerning. A lack in diversity will reduce a population's ability to adapt to change (Markert et al., 2010), which is a growing concern in the face of climate change and everchanging consumer preferences. Increased inbreeding has been shown to lead to inbreeding depression in dairy populations for both production and reproduction (Bjelland et al., 2013). Identifying and maintaining genetic diversity within a population is important.

Animals that have the same genetic merit and similar phenotype can still differ genetically. Genetic redundancy produces more genetic variants (Nowak et al., 1997) than is needed to express the same phenotype. Most traits of economic importance are highly polygenic, with each gene having small contributions to the phenotype, hence the infinitesimal model (Fisher, 1918, Bulmer, 1971, Turelli, 2017). The breeding value itself is simply the sum of the contributions of all genes/markers. The number of possible combinations of genes that still result in the same sum, is infinite. Different sub-populations in the same breed can respond differently to achieve the same breeding objective, showing unparallel changes in allele frequencies over

time (Barghi et al., 2019). Additionally, additive genetic variance is not the only kind of genetic variance. Pleiotropy, epistasis, and dominance can further allow genetic diversity. Liu et al. (2019) found that up to 70% of variance can be attributed to trans-chromosomal effects through peripheral genes that impact the expression of core genes. In the US Holstein population, the percentage of epistatic effects that were inter-chromosomal varied from 1.9% to 84.2%, depending on trait (Prakapenka et al., 2021). This is encouraging for continuous improvement through selection.

CONCLUSION

The accuracy of genomic predictions is predominantly dependent on the ability of the markers to capture LD, the size of the reference population, and the relatedness between the reference and validation population. Multi-breed and across-breed genomic predictions are not straightforward. Linkage disequilibrium does not persist across populations and QTL properties differ. More diverse populations, such as a multi-breed or crossbred population, require more markers to capture LD. Even when LD is captured, other non-genetic factors can reduce the ability to predict across breeds. When all breeds are represented in the reference population, the accuracy of prediction does not vary greatly from those obtained from separate within-breed evaluation. When a pure breed is not included in the reference population, but crossbred animals that contain this breed are, the accuracy of prediction is lower than what would have been achieved within-breed, but higher than what would be achieved in across-breed predictions. Changes are inconsistent across breeds, traits, and methods. When benefits were present, it was usually for the smaller breed. Even within-breed, distinct sub-populations can be present, especially across countries. Genetic redundancy, epistasis, and pleiotropy help maintain substantial genetic variation within a breed, even between animals with the same breeding value.

REFERENCES

Barghi, N., R. Tobler, V. Nolte, A. M. Jakšić, F. Mallard, K. A. Otte, M. Dolezal, T. Taus, R. Kofler, and C. Schlötterer. 2019. Genetic redundancy fuels polygenic adaptation in Drosophila. PLoS biology 17(2):e3000128.

Berry, D. P., P. R. Amer, R. D. Evans, T. Byrne, A. R. Cromie, and F. Hely. 2019. A breeding index to rank beef bulls for use on dairy females to maximize profit. Journal of Dairy Science 102(11):10056-10072.

Bjelland, D. W., K. A. Weigel, N. Vukasinovic, and J. D. Nkrumah. 2013. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. Journal of Dairy Science 96(7):4697-4706.

Bradford, H. L., I. Pocrnić, B. O. Fragomeni, D. A. L. Lourenco, and I. Misztal. 2017. Selection of core animals in the Algorithm for Proven and Young using a simulation model. Journal of Animal Breeding and Genetics 134(6):545-552.

Buckley, F., N. Lopez-Villalobos, and B. J. Heins. 2014. Crossbreeding: implications for dairy cow fertility and survival. animal 8(s1):122-133.

Bulmer. 1971. The effect of selection on genetic variability. The American Naturalist 105(943):201-211.

Calus, M., M. Goddard, Y. Wientjes, P. Bowman, and B. Hayes. 2018. Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. Journal of dairy science 101(5):4279-4294.

Christensen, O. F., P. Madsen, B. Nielsen, and G. Su. 2014. Genomic evaluation of both purebred and crossbred performances. Genetics Selection Evolution 46(1):23.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetics Selection Evolution 44(1):4.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185(3):1021-1031.

De Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183(4):1545-1553.

Duenk, P., P. Bijma, M. P. L. Calus, Y. C. J. Wientjes, and J. H. J. van der Werf. 2020. The Impact of Non-additive Effects on the Genetic Correlation Between Populations. G3 Genes|Genomes|Genetics 10(2):783-795.

Duenk, P., P. Bijma, Y. C. J. Wientjes, and M. P. L. Calus. 2021. Predicting the purebred-crossbred genetic correlation from the genetic variance components in the parental lines. Genetics Selection Evolution 53(1):10.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 95(7):4114-4129.

Fisher, R.A. 1918. The correlation between relatives on the supposition of Mendelian inheritance.

Fragomeni, B., D. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. Lawlor, and I. Misztal. 2015. Hot topic: use of genomic recursions in single-step genomic best linear

unbiased predictor (BLUP) with a large number of genotypes. Journal of dairy science 98(6):4090-4094.

González-Diéguez, D., L. Tusell, A. Bouquet, A. Legarra, and Z. G. Vitezica. 2020. Purebred and Crossbred Genomic Evaluation and Mate Allocation Strategies To Exploit Dominance in Pig Crossbreeding Schemes. G3 Genes|Genomes|Genetics 10(8):2829-2841.

Guinan, F. L., H. D. Norman, and J. W. Dürr. 2019. Changes occurring in the breed composition of US dairy herds. Interbull Bulletin (55):11-16.

Habier, D., R. L. Fernando, and D. J. Garrick. 2013. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics 194(3):597-607.

Hayes, B. J., N. J. Corbet, J. M. Allen, A. R. Laing, M. R. McGowan, G. Fordyce, R. Lyons, and B. M. Burns. 2018. Towards multi-breed genomic evaluations for female fertility of tropical beef cattle. Journal of Animal Science 97(1):55-62.

Hayes, B. J. and M. Goddard. 2008. Prediction of breeding values using marker-derived relationship matrices. Journal of animal science 86(9):2089-2092.

Hidalgo, A. M., J. W. M. Bastiaansen, M. S. Lopes, B. Harlizius, M. A. M. Groenen, and D.-J. de Koning. 2015. Accuracy of Predicted Genomic Breeding Values in Purebred and Crossbred Pigs. G3: Genes|Genomes|Genetics 5(8):1575-1583.

Hidalgo, J., D. Lourenco, S. Tsuruta, Y. Masuda, V. Breen, R. Hawken, M. Bermann, and I. Misztal. 2021. Investigating the persistence of accuracy of genomic predictions over time in broilers. Journal of Animal Science 99(9).

Hollifield, M. K., D. Lourenco, M. Bermann, J. T. Howard, and I. Misztal. 2021. Determining the stability of accuracy of genomic estimated breeding values in future generations in commercial pig populations. Journal of Animal Science 99(4).

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, and P. Croiseau. 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. Journal of Dairy Science 97(6):3918-3929.

Jónás, D., V. Ducrocq, S. Fritz, A. Baur, M.-P. Sanchez, and P. Croiseau. 2017. Genomic evaluation of regional dairy cattle breeds in single-breed and multibreed contexts. Journal of Animal Breeding and Genetics 134(1):3-13.

Karaman, E., G. Su, I. Croue, and M. S. Lund. 2021. Genomic prediction using a reference population of multiple pure breeds and admixed individuals. Genetics Selection Evolution 53(1):46.

Karoui, S., M. J. Carabaño, C. Díaz, and A. Legarra. 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genetics Selection Evolution 44(1):39.

Kemper, K. E., C. M. Reich, P. J. Bowman, C. J. vander Jagt, A. J. Chamberlain, B. A. Mason, B. J. Hayes, and M. E. Goddard. 2015. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genetics Selection Evolution 47(1):29.

Legarra, A., C. A. Garcia-Baccino, Y. C. J. Wientjes, and Z. G. Vitezica. 2021. The Correlation of Substitution Effects Across Populations and Generations in the Presence of Non-Additive Functional Gene Action. bioRxiv:2020.2011.2003.367227.

Ling, A. S., E. H. Hay, S. E. Aggrey, and R. Rekaya. 2021. Dissection of the impact of prioritized QTL-linked and -unlinked SNP markers on the accuracy of genomic selection. BMC Genomic Data 22(1):26.

Liu, X., Y. I. Li, and J. K. Pritchard. 2019. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177(4):1022-1034.e1026.

Livestock Improvement Corporation and Dairy NZ. 2019. New Zealand dairy statistics 2019-2020. https://d1r5hvvxe7dolz.cloudfront.net/media/documents/NZ_Dairy_Statistics_2019-20_WEB_FINAL.pdf Accessed: Sept 20, 2021.

Lourenco, D., S. Tsuruta, B. Fragomeni, C. Chen, W. Herring, and I. Misztal. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. Journal of animal science 94(3):909-919.

Lourenco, D. A., B. O. Fragomeni, S. Tsuruta, I. Aguilar, B. Zumbach, R. J. Hawken, A. Legarra, and I. Misztal. 2015. Accuracy of estimated breeding values with genomic information on males, females, or both: an example on broiler chicken. Genetics Selection Evolution 47(1):56.

Makgahlela, M., I. Strandén, U. Nielsen, M. Sillanpää, and E. Mäntysaari. 2014. Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. Journal of dairy science 97(2):1117-1127.

Mäntysaari, E. A., R. D. Evans, and I. Strandén. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. Journal of Animal Science 95(11):4728-4737.

Marjanovic, J. and M. P. L. Calus. 2021. Factors affecting accuracy of estimated effective number of chromosome segments for numerically small breeds. Journal of Animal Breeding and Genetics 138(2):151-160.

Markert, J. A., D. M. Champlin, R. Gutjahr-Gobell, J. S. Grear, A. Kuhn, T. J. McGreevy, A. Roth, M. J. Bagley, and D. E. Nacci. 2010. Population genetic diversity and fitness in multiple environments. BMC Evolutionary Biology 10(1):205.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLOS ONE 4(4):e5350.

Mayr, E. 1954. Change of genetic environment and evolution.

Meuwissen, T., I. van den Berg, and M. Goddard. 2021. On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL. Genetics Selection Evolution 53(1):19.

Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. Journal of Dairy Science 97(6):3943-3952.

Moghaddar, N., A. A. Swan, and J. H. Van Der Werf. 2014. Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. Genetics Selection Evolution 46(1):58.

Neyhart, J. L., T. Tiede, A. J. Lorenz, and K. P. Smith. 2017. Evaluating Methods of Updating Training Data in Long-Term Genomewide Selection. G3 Genes|Genomes|Genetics 7(5):1499-1510.

Nilforooshan, M. A. and M. Lee. 2019. The quality of the algorithm for proven and young with various sets of core animals in a multibreed sheep population. Journal of Animal Science 97(3):1090-1100.

Nowak, M. A., M. C. Boerlijst, J. Cooke, and J. M. Smith. 1997. Evolution of genetic redundancy. Nature 388(6638):167-171.

Olson, K., P. VanRaden, and M. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science 95(9):5378-5383.

Pocrnic, I., D. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2018. Limited dimensionality of genomic information and effective population size. Page 32 in Proc. Proceedings of the World Congress on Genetics Applied to Livestock Production.

Pocrnic, I., D. A. Lourenco, C.-Y. Chen, W. O. Herring, and I. Misztal. 2019. Crossbred evaluations using single-step genomic BLUP and algorithm for proven and young with different sources of data. Journal of animal science 97(4):1513-1522.

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. The dimensionality of genomic information and its effect on genomic prediction. Genetics 203(1):573-581.

Pocrnic, I., D. A. L. Lourenco, Y. Masuda, and I. Misztal. 2016b. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. Genetics Selection Evolution 48(1):82.

Prakapenka, D., Z. Liang, J. Jiang, L. Ma, and Y. Da. 2021. A Large-Scale Genome-Wide Association Study of Epistasis Effects of Production Traits and Daughter Pregnancy Rate in U.S. Holstein Cattle. Genes 12(7):1089.

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic

selection using a multi-breed, across-country reference population. Journal of Dairy Science 94(5):2625-2630.

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

Rahimi, S. M., A. Rashidi, and H. Esfandyari. 2020. Accounting for differences in linkage disequilibrium in multi-breed genomic prediction. Livestock Science:104165.

Raven, L.-A., B. G. Cocks, and B. J. Hayes. 2014. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. BMC Genomics 15(1):62.

Raymond, B., A. C. Bouwman, C. Schrooten, J. Houwing-Duistermaat, and R. F. Veerkamp. 2018a. Utility of whole-genome sequence data for across-breed genomic prediction. Genetics Selection Evolution 50(1):27.

Raymond, B., A. C. Bouwman, Y. C. J. Wientjes, C. Schrooten, J. Houwing-Duistermaat, and R. F. Veerkamp. 2018b. Genomic prediction for numerically small breeds, using models with pre-selected and differentially weighted markers. Genetics Selection Evolution 50(1):49.

Raymond, B., Y. C. J. Wientjes, A. C. Bouwman, C. Schrooten, and R. F. Veerkamp. 2020. A deterministic equation to predict the accuracy of multi-population genomic prediction with multiple genomic relationship matrices. Genetics Selection Evolution 52(1):21.

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genetics Research 35(2):131-155.

Su, G., R. F. Brøndum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (~ 54,000) and high-density (~ 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of dairy science 95(8):4657-4665.

Swan, A., D. Brown, H. Daetwyler, B. Hayes, M. Kelly, N. Moghaddar, and J. Van der Werf. 2014. Genomic evaluations in the Australian sheep industry. in Proc. 10th World Congress of Genetics Applied to Livestock Production.

Templeton, A. R. 1980. The theory of speciation via the founder principle. Genetics 94(4):1011-1038.

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

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

Turelli, M. 2017. Commentary: Fisher's infinitesimal model: A story for the ages. Theoretical Population Biology 118:46-49.

Van den Berg, I., D. Boichard, B. Guldbrandtsen, and M. S. Lund. 2016a. Using Sequence Variants in Linkage Disequilibrium with Causative Mutations to Improve Across Breed Prediction in Dairy Cattle: A Simulation Study. G3: Genes|Genomes|Genetics.

Van den Berg, I., D. Boichard, and M. S. Lund. 2016b. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genetics Selection Evolution 48(1):83.

Van den Berg, I., I. MacLeod, C. Reich, E. Breen, and J. Pryce. 2020. Optimizing genomic prediction for Australian Red dairy cattle. Journal of Dairy Science.

Van den Berg, I., T. H. E. Meuwissen, I. M. MacLeod, and M. E. Goddard. 2019. Predicting the effect of reference population on the accuracy of within, across, and multibreed genomic prediction. Journal of Dairy Science 102(4):3155-3174.

Van den Berg, S., M. P. L. Calus, T. H. E. Meuwissen, and Y. C. J. Wientjes. 2015. Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. BMC Genetics 16(1):146.

Van Grevenhof, I. E. and J. H. Van Der Werf. 2015. Design of reference populations for genomic selection in crossbreeding programs. Genetics Selection Evolution 47(1):14.

VanRaden, P. M., M. E. Tooker, T. C. S. Chud, H. D. Norman, J. H. Megonigal, I. W. Haagen, and G. R. Wiggans. 2020. Genomic predictions for crossbred dairy cattle. Journal of Dairy Science 103(2):1620-1631.

VanRaden, P. M., M. E. Tooker, J. B. Cole, G. R. Wiggans, and J. H. Megonigal. 2007. Genetic Evaluations for Mixed-Breed Populations. Journal of Dairy Science 90(5):2434-2441.

Walsh, B. and M. Lynch. 2018. Evolution and selection of quantitative traits. Oxford University Press.

Wei, M. and J. H. J. van der Werf. 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. Animal Science 59(3):401-413.

Wei, M. and J. J. H. van der Werf. 1995. Genetic correlation and heritabilities for purebred and crossbred performance in poultry egg production traits. Journal of Animal Science 73(8):2220-2226.

Wientjes, Y. C., M. P. Calus, M. E. Goddard, and B. J. Hayes. 2015. Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genetics Selection Evolution 47(1):42.

Wientjes, Y. C. J., P. Bijma, J. Vandenplas, and M. P. L. Calus. 2017. Multi-population Genomic Relationships for Estimating Current Genetic Variances Within and Genetic Correlations Between Populations. Genetics 207(2):503-515.

Wientjes, Y. C. J., P. Bijma, R. F. Veerkamp, and M. P. L. Calus. 2016. An Equation to Predict the Accuracy of Genomic Values by Combining Data from Multiple Traits, Populations, or Environments. Genetics 202(2):799-823.

Wiggans, G., P. VanRaden, and E. Nicolazzi. 2019. Extending genomic evaluation to crossbred dairy cattle: US implementation. Interbull Bulletin (55):46-49.

Zhou, L., M. S. Lund, Y. Wang, and G. Su. 2014. Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. Journal of Animal Breeding and Genetics 131(4):249-257.

CHAPTER 3

GENOMIC PREDICTIONS IN PUREBREDS WITH A MULTI-BREED GENOMIC RELATIONSHIP MATRIX[1]

---

[1]Y. Steyn, D.A.L. Lourenco, I. Misztal. 2019. *Journal of Animal Science*. 97 (11): 4418-4427.

ABSTRACT

Combining breeds in a multi-breed evaluation can have a negative impact on prediction accuracy, especially if SNP effects differ among breeds. The aim of this study was to evaluate the use of a multi-breed genomic relationship matrix (**G**), where SNP effects are considered to be unique to each breed, i.e., non-shared. This multi-breed **G** was created by treating SNP of different breeds as if they were on non-overlapping positions on the chromosome although in reality they were not. This simple setup may avoid spurious IBS relationships between breeds and automatically considers breed-specific allele frequencies. This scenario was contrasted to a regular multi-breed evaluation where all SNP were shared, i.e., the same position, and to single-breed evaluations. Different SNP densities (9k and 45k), and different effective population sizes (Ne) were tested. Five breeds mimicking recent beef cattle populations that diverged from the same historical population were simulated using different selection criteria. It was assumed that QTL effects were the same over all breeds. For the recent population, generations 1 to 9 had approximately half of the animals genotyped, whereas all animals in generation 10 were genotyped. Generation 10 animals were set for validation; therefore, each breed had a validation group. Analysis were performed using single-step GBLUP (ssGBLUP). Prediction accuracy was calculated as correlation between true (T) and genomic estimated (GE) BV. Accuracies of GEBV were lower for the larger Ne and low SNP density. All three evaluation scenarios using 45k resulted in similar accuracies, suggesting that the marker density is high enough to account for relationships and linkage disequilibrium with QTL. A shared multi-breed evaluation using 9k resulted in a decrease of accuracy of 0.08 for a smaller Ne and 0.12 for a larger Ne. This loss was mostly avoided when markers were treated as non-shared within the same **G** matrix. A **G** matrix

with non-shared SNP enables multi-breed evaluations without considerably changing accuracy, especially with limited information per breed.

INTRODUCTION

Genomic evaluations have become common in animal breeding due to the possibility of improving the rate of genetic gain (Schaeffer, 2006). Traditionally, evaluations are done within pure breeds. There is an interest in multi-breed evaluations in cattle and sheep where there are many breeds and crosses, as well as chickens and pigs with many lines. Combining breeds increases the training population, which can potentially enhance accuracy of genomic predictions (Hayes and Goddard, 2008) and is reasonably simple. It also allows the sharing of resources such as funding, specialists, and infrastructure, which is especially attractive and practical for small breeds or countries. This simplicity may come at the expense of accuracy for some or all breeds. Many studies have struggled to find an advantage of using multi-breed evaluations, sometimes obtaining a slightly increased accuracy but often unchanged or slightly decreased (Hayes et al., 2009; Erbe et al., 2012; Olson et al., 2012; Makgahlela et al., 2014; Calus et al., 2018).

The linkage disequilibrium (LD) between markers and QTL differs among breed and does not persist across breeds (De Roos et al., 2009). Capturing LD contributes more to the accuracy of prediction than tracking relationships through markers (Habier et al., 2013), making it essential to have dense enough markers to capture LD in diverse populations. This extra genotyping cost could defy the cost-saving strategies of sharing other resources. Avoiding a loss of accuracy with low density markers would be beneficial.

The cause of persistent or non-persistent LD is due to independent chromosome segments (ICS). Genomic selection in the absence of QTL identification is based on the estimation of these segments (Goddard, 2009). Larger segments means that more markers will be in LD with the QTL

and therefor fewer SNP are required. The number of ICS (Me) is expected to be 4NeL, where Ne is the effective population size and L is the chromosome length (Stam, 1980). Populations with a smaller Ne will have fewer, larger ICS and therefore require fewer SNP markers to obtain the same accuracy as those with larger Ne (Pocrnic et al., 2018). Combining different breeds or lines together in a single evaluation should increase genetic diversity and Ne, requiring more SNP to trace all segments and avoid loss of accuracy. In fact, Pocrnic et al. (2019) showed that about 30% of the chromosome segments were independent between two different pig lines, which caused a reduction in prediction accuracy across the lines.

Even when markers and QTL are in LD, the QTL and allele substitution effects can differ among breeds (Thaller et al., 2003). Differences in QTL minor allele frequencies (MAF) also affects accuracy of prediction (Wientjes et al., 2015) and differs in populations due to selection and genetic drift.

The objective of this study was to evaluate the accuracy of multi-breed genomic prediction when using the same SNP effects for all breeds, and when obtaining breed-specific SNP effects by treating the markers as non-shared in populations with different Ne and genotyping density.

MATERIALS AND METHODS

Simulated data

Five different breeds were simulated using QMSim (Sargolzaei and Schenkel, 2009) from a historical population that was randomly mated for 1,000 generations. The historic population started with 10,000 animals, decreased to 1,000 at generation 500 to create LD and reached 9,000 animals at generation 1,000. Different number of founders using different selection criteria were selected from this population to create the breeds (i.e., recent populations), or distant lines. Each dam had only one progeny.

Within each breed, animals were randomly mated without artificial selection for 40 generations after which half of the animals in the last generation were used to initiate the selected population. Selection based on high EBV for a single trait was applied for 10 generations using different mating designs and proportion of replacement males and females. Some breeds differed in their number of initial animals. This led to slightly different breed sizes, which created different effective population sizes (Ne). The trait heritability was 0.30 and phenotypes were from a normal distribution with a mean of 0 and variance of 1. Two simulations were done where one had double the number of initially selected animals compared to the other to create breeds with a larger Ne. The summary of parameters used for each breed are presented in Table 3.1.

The genome was simulated assuming 29 chromosomes of varying length, resulting in a total genome length of 23 Morgans, with 1,000 QTL evenly distributed among them. The QTL effects were sampled from a Gamma distribution with a shape parameter of 0.4. After quality control of genotypes, 45,000 segregating SNP with minor allele frequency > 0.05 were retained for analysis. There was a mutation rate for markers and QTLs of 2.5×10^{-5} per generation per locus. The QTL effects were assumed to be the same over all the breeds but the QTLs had different frequencies in each population. This caused a difference in variance explained for each breed. The largest QTL variance in generation 10 of each breed did not exceed 0.02. The average LD measured as the pooled square of the correlation between markers (R^2) was between 0.14 and 0.18 within breed. Few QTL became fixed in the breeds, using a MAF of <0.001 as criteria.

The genotyped animals consisted of all animals in generation 10, and 600 randomly selected animals from each previous 9 generations. The difference in the size of the 10^{th} generation led to slightly different numbers of genotyped animals per breed (Table 1.1). Two different SNP

densities were used in this study – 45k and a subset of 9k. The 9k SNP panel was created by selecting each 5th marker from the full 45k panel.

The simulation was replicated four times and the entire process is visually explained in Fig 3.1, including the total number of animals (genotyped or not) in the 10 generations that had undergone selection. The different selection criteria for the breeds and the resulting number of animals and Ne in the data/pedigree and genotype file for both small and large Ne are presented in Table 3.1. A principle component analysis (PCA) based on 45k was done to visualize the separation between breeds. The Ne for each breed was calculated using the following formula by Wright (1931):

$$Ne_T = \frac{4N_m N_f}{N_m + N_f}$$

where N_m and N_f are the number of breeding males and females in each generation. This method assumes no selection (Table 3.1). Various other methods exist to calculate Ne.

Model and Analyses

A single-trait animal model was fitted for traditional pedigree-based and genomic evaluations:

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

where **y** is a vector of simulated phenotypes, μ is an overall mean, **u** is a vector of additive genetic effects, **Z** is an incidence matrix relating **y** to the effects in **u**, and **e** is a vector of random residuals.

Single-step genomic BLUP (ssGBLUP) was used with BLUPF90 software (Misztal et al., 2014) for analyses of all breeds, both separately and together to obtain genomic estimated breeding values (GEBV). Single-step GBLUP is simple to apply, avoids double counting, accounts for pre-selection on Mendelian sampling, and allows the inclusion of genotyped and non-genotyped

animals in the same evaluation. In ssGBLUP the inverse of the pedigree relationship matrix ($\mathbf{A}^{-1}$), regularly used in BLUP, is replaced with the inverse of the realized relationship matrix ($\mathbf{H}^{-1}$) as demonstrated by Aguilar et al. (2010). This $\mathbf{H}^{-1}$ combines $\mathbf{A}^{-1}$ with the inverse of the genomic relationship matrix ($\mathbf{G}^{-1}$):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix}$$

In this case, **G** was obtained using the formula $\mathbf{G} = \frac{\mathbf{MM'}}{2\sum p_i(1-p_i)}$ where **M** is a centered matrix of marker content adjusted for allele frequencies and p_i is the allele frequency for SNP i (VanRaden, 2008). The pedigree-based relationship matrix between genotyped animals is referred to as $\mathbf{A}_{22}$. To reduce bias due to the different genetic level of genotyped and non-genotyped animals, **G** is tuned using the constant α, where α is $\frac{1}{n^2}(\sum_i \sum_j \mathbf{A}_{22(i,j)} - \sum_i \sum_j \mathbf{G}_{(i,j)})$ (Vitezica et al., 2011) and n is the number of animals. To avoid singularity problems, **G** was multiplied by 0.95 and $\mathbf{A}_{22}$ by 0.05 before combining them.

For validation purposes, phenotypes of the validation animals (generation 10) were removed before estimating GEBV. The resulting GEBV of generation 10 were correlated to the true breeding value (TBV) to obtain the accuracy of prediction. A Pearson correlation was used and bias was estimated as the regression coefficient when regressing TBV over GEBV.

The GEBVs were estimated by within- and multi-breed evaluations using 9k and 45k SNP markers. The multi-breed evaluations were done based on two different assumptions: a) animals from different breeds shared the same SNP effects (shared); b) animals from different breeds did not share the same SNP effects (non-shared). To achieve the first assumption, genotypes for animals from all breeds overlap and are stacked together prior to the evaluation. In the second assumption, SNP are considered non-overlapping. The SNP file is manipulated to achieve this.

The SNP of the same breed are in the same columns in the file. The SNP of different breeds appear in different columns in the SNP file to create a non-shared scenario. Breed A is treated as if it had SNP markers from position 1 to 45k (or 1 to 9k) and missing markers for the remaining 180k (or 36k) SNP. Breed B is treated as if its markers start at position 45,001 to 90k (or 9,001 to 18k) with all other markers as missing. The same pattern continues for breeds C, D and E. This multiplies the number of columns in the SNP file by the number of breeds, even though all animals have SNP in the same physical position on the chromosome. Therefore, in the non-shared scenario, the newly created SNP file will have n x m markers, where n is the number of markers evaluated and m is the number of breeds. The treatment of the SNP file is graphically explained in Fig 3.2. The shared scenario results in a genomic relationship matrix in which all animals are related to each other (i.e., a dense matrix). The non-shared scenario results in a genomic relationship matrix where all animals within a breed are related, but animals from different breeds are not. This assumes a correlation of 0 between breeds. All missing SNP are ignored to calculate allele frequency, which allows each breed to be centered around its own frequencies. This is graphically explained in Fig 3.3.

An across-breed analyses was also performed to determine whether the breeds are genetically distant. If breeds are closely related, it would be expected that the direct genomic values (DGV) of one breed can be predicted based on SNP solutions obtained from another breed, with similar accuracy as in single-breed evaluations. For these analysis, SNP effects for one breed were computed based on GEBV using the POSTGSF90 software (Misztal et al., 2014) with the following formula:

$$\hat{\mathbf{a}} = \lambda \mathbf{D}\mathbf{M}'\mathbf{G}^{-1}(\text{GEBV})$$

where $\hat{\mathbf{a}}$ is a vector of estimated SNP effects, λ is the ratio of SNP to additive genetic variance, **D** is a diagonal matrix of weights (standardized variances) for SNP and **M** is a matrix of centered genotypes for each animal (VanRaden, 2008). Based on SNP effects, DGV for validation animals were calculated by PREDF90 (Misztal et al., 2014) as the sum of SNP effects weighted by the genotype content.

In across-breed evaluations, the SNP effects estimated based on one breed were used to calculate DGV of its own validation population and that of the other breeds. For example, SNP effects estimated using only the training population of Breed A were used to calculate the DGV of the validation population of Breed A and then the validation populations of Breed B, Breed C, Breed D, and Breed E separately.

Accuracy was computed as the Pearson correlation between TBV and GEBV or DGV, whereas bias was calculated as the regression coefficient when regressing TBV on GEBV. This shows the over- or under-dispersion of GEBV.

RESULTS AND DISCUSSION

The genetic distance across the breeds using 45k can be observed in the PCA plots. Fig 3.4 is a 3-dimensional plot showing the first three principal components. Breed A and E showed an overlap although Breed E was more variable. This could be because Breed E underwent negative assortative mating but founders in generation 0 of these two breeds were selected using similar criteria. The variance explained by the first five principal components averaged over replicates for the smaller Ne were 3.65%, 3.21%, 2.80%, 2.40% and 0.56%, respectively. The values for the larger Ne were 2.38%, 2.08%, 1.84%, 1.50% and 0.39%, respectively. The number of eigenvalues explaining 98% (EIGEN98) of the genomic variation for each breed and all combined are presented in Table 3.2. Considerably more eigenvalues were needed to explain 98% of the genomic

variation in the combined evaluation, reflecting a more diverse population. The value is smaller than the sum of the required EIGEN98 for each individual breed, meaning the breeds are not completely independent. This was expected since they originated from the same historical population that had the same QTL effects.

Very poor predictive ability was observed across-breed. Table 3.3 presents the within-breed predictive ability in the diagonal and across-breed in the off-diagonal when using 45k in the smaller Ne scenario. The other SNP densities or Ne showed the same trend. The DGV within breed had an average accuracy of 0.70 ± 0.02 when using the scenario with a smaller Ne and 45k, and 0.66 ± 0.01 with a larger Ne. Whereas, the average accuracy across breed was 0.11 ± 0.03 with a smaller Ne with 45k and 0.07 ± 0.02 with a larger Ne. This simulation results correspond to other studies (Hayes et al., 2009; Kizilkaya et al., 2010; Pryce et al., 2011; Olson et al., 2012; Kachman et al., 2013, Zhou et al. 2014) including Raymond et al. (2018), who found poor predictive ability in dairy cattle across-breed and across-country, even when using whole-genome sequence (WGS) information.

The lack of predictive ability in across breed further showed that breeds were genetically different. Within more homogenous breeds, larger ICS will be present and SNP estimates will capture these segments (Goddard et al., 2011). Across breed, animals will share shorter segments, which is more difficult to estimate accurately. Therefore, information from one breed is limited for another even when the true SNP effects are the same (Khansefid et al. 2014). Correlations between estimated SNP effects of different breeds in this study were all lower than 0.05, regardless of the Ne.

Table 3.4 shows accuracies for breeds A-E with a smaller Ne, 9k and 45k SNP information when each breed was considered separately, when all breeds shared SNP effects, and when the

SNP information for each breed was treated as non-shared. With 9k SNP, i.e., 80% of the SNP masked, the accuracies with analyses for each breed separately were on average 0.05 lower than with 45k SNP. The 9k SNP are not enough to fully account for the genomic information provided by larger SNP panels, although the difference was relatively small. In a study by Luan et al. (2009), masking 75% of SNP reduced the realized accuracy in Norwegian Red Cattle by 0.02. They postulated that this could partly be due to SNP markers clustering together with high LD in some regions of the chromosome.

The accuracies with 45k SNP remained stable no matter how the evaluation was set up. Sharing SNP effects using 9k decreased the accuracy compared to single-breed analyses by an average of 0.08 when the Ne is smaller. Thus with a limited number of SNP, sharing SNP among several breeds is not ideal.

Table 3.5 shows the results obtained with a larger Ne. All accuracies were lower than obtained in a smaller Ne, as expected from the increase of Me (Daetwyler et al., 2010; Pocrnic et al., 2016). The drop in accuracy from single-breed to shared SNP was 0.12 in the larger Ne, instead of 0.08. Sharing SNP comes with a larger accuracy penalty when the population is more diverse. Accuracies still remained stable when SNP were treated as non-shared, even with a drastically lower SNP density or when Ne was larger.

Even though, on average accuracies were the same, very small differences were observed for some breeds in both 45k and 9k evaluations. Such small changes can be attributed to the scaling of **G** before being combined with $\mathbf{A}_{22}$. A simulation study by Vitezica et al. (2011) found that bias exists in estimation when no adjustments are made to account for the fact that the genetic level of genotyped animals is different from that of the whole population, especially with strong selection. The **G** must be combined with a constant α, which is equal to the average difference between all

elements of $\mathbf{A}_{22}$ and **G** to account for this difference. In the non-shared SNP scenario all animals appear in these matrices, and therefore, the constant used for scaling is overall, instead of breed-specific. Such adjustments could be done in theory, however, in practice, any partial change done in specific portions of **G** may result in non-positive definiteness. It is important to note that these differences in accuracy are often in the third decimal and thus negligibly small. They may be bigger in reality where data structures are more complex and incomplete. In practical studies the effect of scaling was limited, indicating weak selection on any individual trait with multi-trait selection, but the scaling had some effect on biases and inflation of GEBV (Chen et al., 2011). Table 3.6 shows the average bias in all scenarios. When using 9k SNP, GEBV were clearly inflated when SNP were shared, especially when Ne was smaller. The bias was essentially unchanged when SNP were non-shared. The 45k scenario led to negligibly small changes, regardless of method used.

Olson et al. (2012) reported that sharing three breeds –Holstein, Jersey, and Brown Swiss using about 44k SNP- reduced accuracy compared to single-breed analyses using US data. This suggests that larger populations may benefit from more SNPs while smaller do not require so many, however, a point is reached where denser markers are not useful. It has been shown that marker densities more than 50k generally do not show a remarkable increase in accuracy of multi-breed evaluations, even when using approximately 600k (Erbe et al., 2012) or 700k (Su et al., 2012; Hozé et al., 2014).

In this study, accuracy was depressed with SNP sharing across breeds only with 9k but not with 45k SNP information. When 9k SNP were shared, there was not enough information to estimate the ICS in the multi-breed population. Pocrnic et al. (2018) showed that combining breeds increases Ne, which requires more SNP to accurately trace all chromosome segments segregating in the population. The Me as proposed by Stam (1980) can be approximated as the number of

eigenvalues explaining 98% of the variance of **G** (EIGEN98; Pocrnic et al., 2016). The EIGEN98 in this study for each breed separately were approximately 3.6k and 5k for the smaller and larger Ne, respectively (Table 3.2). These EIGEN98 do not correspond to the prediction of Stam (1980). This indicates that the Ne is lower than estimated by the formula of Wright (1931). There are various methods of estimating Ne, each has particular justifications or merit. Although the simulated scenario with larger Ne is not double that of the smaller one, it is still larger. When the breeds were pooled together and SNP were shared, EIGEN98 were 13k for smaller Ne and 18k for larger Ne, meaning 9k SNP were not enough to estimate all the chromosome segments segregating in the combined population.

The observed drop of 0.08 for the smaller Ne and 0.12 for the larger reflects a loss in accuracy that is dependent on the proportion of available SNP and segregating ICS in the population under genomic selection. When breeds are pooled together but SNP are not shared, the genomic information for each breed behaves independently and the chromosome segments are assumed to be segregating only within breed. This reflects the reality of multi-breed evaluations if crossbreds are not included. The inclusion of crossbred animals when using multi-breed **G** with non-shared SNP is discussed later.

A question arises whether sharing SNP at 45k level will reduce accuracy for larger data sets. For example, the Irish national evaluation uses a shared SNP model for over 40 breeds (Mantysaari et al., 2017) with a 54k chip. A 14-breed evaluation that includes the Simmental breed also uses a shared SNP model with less than 3k SNP (Golden et al., 2018). If each breed has Me close to 10,000, even with 50% common segments, the combined Me assuming unrelated breeds could be close to 200,000 and 70,000, respectively, much larger than 18,000 reported in this study. Therefore, the impact of SNP sharing can be larger in real populations than observed here.

Pocrnic et al. (2019) looked at accuracy of genomic prediction in GBLUP when only a fraction of the largest eigenvalues of **G** were retained. With less phenotypic information, they found that little accuracy was gained when more than 10% of the total number of EIGEN98 were considered. With more phenotypic information, the accuracy reached a plateau when about 50% of the eigenvalues were considered. As more SNPs were needed to estimate more eigenvalues, the study suggests larger data benefits from more SNP.

Multi-breed evaluations are used in the livestock improvement industry and some studies have shown increases in accuracy, even with the assumption of shared SNP effects. Literature where benefits were found showed that they were mostly small and inconsistent over traits, breeds and methods (Hayes et al., 2009; Karoui et al., 2012; Olson et al., 2012; Makgahlela et al., 2013b; Hozé et al., 2014; Jónás et al., 2017). Olson et al. (2012) also treated three dairy breeds as different traits in a multi-trait evaluation, which slightly increased the accuracy and prevented the largest breed from dominating the smaller breeds. However, this slight improvement did not justify the increased computational demand. Similar findings were made by Makgahlela et al. (2013b). In another study, Makgahlela et al. (2013a) adjusted **G** for breed-specific allele frequencies of a multi-breed evaluation of Nordic Red cattle and Lourenco et al. (2016) did the same for a pig population. They all found that breed-specific **G** did not improve the validation accuracy. Zhou *et al.* (2014) accounted for LD phasing and breed-specific SNP-effects by using weights in externally created within- and between-breed G-matrix blocks. Accuracies were not improved further compared to a multi-trait approached. Khansefid *et al.* (2014) created different **G** for breeds, or combinations of breeds, and combined them with off-diagonals between breeds set to zero. Depending on the method of measuring accuracy and the grouping of animals, accuracies changed for some breeds and traits.

In general, realistic simulation of multi-breed data is a difficult topic as a more comprehensive simulation would involve dominance and epistasis, and consequently QTL with different substitution effects (Spelman et al., 2002). In this simulation, QTL substitution effects were set to be equal among breeds. This assumes a correlation of one, however, this is not the case in practice (Wientjes et al., 2017). Even though this assumption is unrealistic, the SNP markers had different frequencies and their effects still differed among the breeds.

Scaling G when sharing or not sharing SNP effects

Aside from explicit scaling of relationships as in Vitezica et al. (2011), alternative options include using base population allele frequencies for each breed (Strandén and Christensen, 2011), or in Single-Step Bayesian Regression, fitting fixed effect for each breed (Hsu et al., 2017). Scaling by base population frequencies is less obvious in cases when populations are heterogeneous, i.e., parents are missing across generations. With shared SNP effects, a recently developed option is fitting "metafounders" – a special form of unknown parent groups – to each combination of breed and generation (Legarra et al., 2015). The advantage of metafounders depends on the quality of estimates of parameters for such a model.

Non-shared SNP effects and crossbred animals

When sharing SNP, all breeds and crossbred animals are considered together naturally. When SNPs are separate across breeds, considering crossbreds is more difficult. One possibility is to consider only F1 and use phasing to ascertain which alleles came from sire and dam (Xiang et al., 2016). In this case, F1 would have twice as many SNP as the parents. A more complex possibility for any crossbred would be to use weighted genotypes based on breed proportions. The use of crossbreds would be justified if they increase accuracy of the purebreds. In a study in pigs

involving two parental lines and crossbreds, use of shared SNPs resulted in same accuracy, and addition of crossbreds did not improve purebred accuracy (Pocrnic et al., 2019).

Implications

The use of non-shared SNP as in this study, i.e., by creating a SNP file with a separate block for each breed may lead to large files if many breeds are considered. Possible solutions include compressed files where empty fields are not explicitly stored, or separate genotype files per breed and breed combinations. Using preselected markers based on importance for the traits can also drastically reduce the genotype file while even improving the accuracy of prediction (Erbe et al., 2012; Van den Berg et al., 2016; Raymond et al., 2018). Preselected SNPs for each breed separately can be combined in a single evaluation using this non-shared SNP method. Additionally, if a fraction of SNP has similar effect size among breeds, a combined **G** can be constructed that considers shared and non-shared SNP. Overall, the use of non-shared SNP may make the most sense when the number of genotyped animals is unequal among breeds. In such case, sharing SNP favors the largest breeds, possibly at the cost of lower accuracy for the smaller breeds.

CONCLUSION

Sharing SNP effects among breeds is an easy way to perform genomic evaluation of multiple breeds and crossbreds, but can result in reduction of accuracy as well as biases if scaling is incorrect and the number of markers is not sufficient. Remedies include increasing the number of SNP and using appropriate scaling. Use of non-shared SNP per breed can avoid reduction of accuracy even when marker density is low, however, accommodating crossbreds more complex than F1 may be difficult. A decision on whether to share or not to share SNP effects can be made based on validation and on computing viability.

REFERENCES

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93: 743-752 doi:10.3168/jds.2009-2730

Calus, M., M. Goddard, Y. Wientjes, P. Bowman, and B. Hayes. 2018. Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. J. Dairy Sci. 101: 4279-4294

Chen, C.-Y., I. Misztal, I. Aguilar, A. Legarra, and W. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci 89: 2673-2679 doi:10.2527/jas.2010-3555

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021-1031 doi:10.1534/genetics.110.116855

De Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183: 1545-1553 doi:10.1534/genetics.109.104935

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95: 4114-4129 doi:10.3168/jds.2011-5019

Falconer, D., and T. Mackay. 1996. Introduction to quantitative genetics. 4 ed. Pearson, Essex, UK.

Goddard, M. E. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136: 245-257 doi:10.1007/s10709-008-9308-0

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128:409-421. doi:10.1111/j.1439-0388.2011.00964.x

Golden, B. L., M. L. Spangler, W. M. Snelling, and D. J. Garrick. 2018. Current Single-step National Beef Cattle Evaluation Models Used by the American Hereford Association and International Genetic Solutions, Computational Aspects, and Implications of Marker Selection . In Proc. 11th Gen. Pred. Worksh., Kansas City, Dec 5-6.

Habier, D., R. L. Fernando, and D. J. Garrick. 2013. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics 194: 597-607 doi:10.1534/genetics.113.152207

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41: 51 doi:10.1186/1297-9686-41-51

Hayes, B. J., and M. Goddard. 2008. Prediction of breeding values using marker-derived relationship matrices. J. Anim. Sci 86: 2089-2092 doi:10.2527/jas.2007-0733

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, and P. Croiseau. 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J. Dairy Sci. 97: 3918-3929 doi:10.3168/jds.2013-7761

Hsu, W.-L., D. J. Garrick, and R. L. Fernando. 2017. The Accuracy and Bias of Single-Step Genomic Prediction for Populations Under Selection. G37: 2685-2694 doi:10.1534/g3.117.043596

Jónás, D., V. Ducrocq, S. Fritz, A. Baur, M.-P. Sanchez, and P. Croiseau. 2017. Genomic evaluation of regional dairy cattle breeds in single-breed and multibreed contexts. J. Anim. Breed. Genet. 134: 3-13 doi:10.1111/jbg.12249

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

Khansefid, M., J.E. Pryce, S. Bolormaa, S.P. Miller, Z. Wang, C. Li, and M. E. Goddard. 2014. Estimation of genomic breeding values for residual feed intake in a multibreed cattle population. J. Anim. Sci. 92:3270-3283. doi: 10.2527/jas2014-7375

Karoui, S., M. J. Carabaño, C. Díaz, and A. Legarra. 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genet. Sel. Evol. 44: 39 doi:10.1186/1297-9686-44-39

Kizilkaya, K., R. Fernando, and D. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J. Anim. Sci 88: 544-551 doi:10.2527/jas.2009-2064

Legarra, A., O. F. Christensen, Z. G. Vitezica, I. Aguilar, and I. Misztal. 2015. Ancestral Relationships Using Metafounders: Finite Ancestral Populations and Across Population Relationships. Genetics 200: 455-468 doi:10.1534/genetics.115.177014

Lourenco, D.A.L., S. Tsuruta, B.O. Fragomeni, C.Y. Chen, W.O. Herring and I. Misztal. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. J.Anim.Sci. 94: 909-919. doi:10.2527/jas.2015-9748

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen, and T. H. E. Meuwissen. 2009. The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. Genetics 183: 1119-1126 doi:10.1534/genetics.109.107391

Makgahlela, M., I. Strandén, U. Nielsen, M. Sillanpää, and E. Mäntysaari. 2013a. The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. J. Dairy Sci. 96: 5364-5375 doi:10.3168/jds.2012-6523

Makgahlela, M., I. Strandén, U. Nielsen, M. Sillanpää, and E. Mäntysaari. 2014. Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population. J. Dairy Sci. 97: 1117-1127 doi:10.3168/jds.2013-7167

Makgahlela, M. L., E. A. Mäntysaari, I. Strandén, M. Koivula, U. S. Nielsen, M. J. Sillanpää, and J. Juga. 2013b. Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. J. Anim. Breed. Genet. 130: 10-19 doi:10.1111/j.1439-0388.2012.01017.x

Mäntysaari, E. A., R. D. Evans, and I. Strandén. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. J. Anim. Sci. 95(11): 4728–4737 doi: 10.2527/jas2017.1912

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92: 4648-4655 doi:10.3168/jds.2009-2064

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2014. Manual for BLUPF90 family of programs Athens: University of Georgia

Olson, K., P. VanRaden, and M. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 95: 5378-5383 doi:10.3168/jds.2011-5006

Pocrnic, I., D. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2018. Limited dimensionality of genomic information and effective population size. In: Proceedings of the World Congress on Genetics Applied to Livestock Production. p 32.

Pocrnic, I., D. A. Lourenco, C.-Y. Chen, W. O. Herring, and I. Misztal. 2019. Crossbred evaluations using single-step genomic BLUP and algorithm for proven and young with different sources of data. J. Anim. Sci doi:10.1093/jas/skz042

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016. The dimensionality of genomic information and its effect on genomic prediction. Genetics 203: 573-581 doi:10.1534/genetics.116.187013

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. J. Dairy Sci. 94: 2625-2630 doi:10.3168/jds.2010-3719

Raymond, B., A. C. Bouwman, C. Schrooten, J. Houwing-Duistermaat, and R. F. Veerkamp. 2018. Utility of whole-genome sequence data for across-breed genomic prediction. Genet. Sel. Evol. 50: 27 doi:10.1186/s12711-018-0396-8

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680-681 doi:10.1093/bioinformatics/btp045

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123: 218-223 doi:10.1111/j.1439-0388.2006.00595.x

Spelman, R., C. Ford, P. McElhinney, G. Gregory, and R. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. J. Dairy Sci. 85: 3514-3517 doi:10.3168/jds.S0022-0302(02)74440-8

Stam, P. 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet. Res. 35: 131-155 doi:
10.1017/S0016672300014002

Strandén, I., and O. F. Christensen. 2011. Allele coding in genomic evaluation. Genet. Sel. Evol. 43: 25 doi:10.1186/1297-9686-43-25

Su, G., R. F. Brøndum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (~ 54,000) and high-density (~ 777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J. Dairy Sci. 95: 4657-4665 doi:10.3168/jds.2012-5379

Thaller, G., A. Winter, R. Fries, W. Krämer, B. Kaupe, and G. Erhardt. 2003. Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci 81: 1911-1918 doi:10.2527/2003.8181911x

Van den Berg, I., D. Boichard, B. Guldbrandtsen, and M. S. Lund. 2016. Using Sequence Variants in Linkage Disequilibrium with Causative Mutations to Improve Across Breed Prediction in Dairy Cattle: A Simulation Study. G3 doi:10.1534/g3.116.027730

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414-4423 doi:10.3168/jds.2007-0980

Vitezica, Z., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. Genet. Res. 93: 357-366 doi:10.1017/S001667231100022X

Wientjes, Y. C., M. P. Calus, M. E. Goddard, and B. J. Hayes. 2015. Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genet. Sel. Evol. 47: 42 doi:10.1186/s12711-015-0124-6

Wientjes, Y. C. J., P. Bijma, J. Vandenplas, and M. P. L. Calus. 2017. Multi-population Genomic Relationships for Estimating Current Genetic Variances Within and Genetic Correlations Between Populations. Genetics 207: 503-515 doi:10.1534/genetics.117.300152

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16: 97

Xiang, T., G. Su, O. F. Christensen, A. Legarra, and B. Nielsen. 2016. Application of single-step genomic evaluation for crossbred performance in pig. J. Anim. Sci 94: 936-948 doi:10.2527/jas.2015-9930

Zhou, L., M. S. Lund, Y. Wang and G. Su. 2014. Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. J. Anim. Breed. Genet. doi:10.1111/jbg.12089

TABLES

Table 3.1: Summary of parameters used to simulate the five different breeds for the evaluation

Simulation	Breed A	Breed B	Breed C	Breed D	Breed E
Sire replacement	0.50	0.50	0.60	0.60	0.50
Dam replacement	0.20	0.30	0.30	0.20	0.20
Mating design	Random	Random	+ assortative	Random	-assortative
Smaller Ne					
Ne	98	118	117	98	117
Initial males	25	30	30	25	30
Initial females	1,200	1,500	1,200	1,500	1,200
Final data	13,225	16,530	13,230	16,525	13,230
Genotyped	6,600	6,900	6,600	6,900	6,600
Larger Ne					
Ne	196	236	234	197	234
Initial males	50	60	60	50	60
Initial females	2,400	3,000	2,400	1,500	2,400
Final data	26,450	33,060	26,460	33,050	26,460
Genotyped	7,800	8,400	7,800	8,400	7,800

Table 3.2: The number of eigenvalues explaining 98% of the variation of each breed and in a full multi-breed scenario using 45k SNP markers when effective population size (Ne) is smaller and larger.

	Breed A	Breed B	Breed C	Breed D	Breed E	**ABCDE**
Smaller Ne	3,780	3,627	3,280	3,764	3,737	**13,024**
Larger Ne	5,072	5,189	4,661	5,172	5,073	**18,059**

Table 3.3: The correlation between true breeding value (TBV) and direct genomic value (DGV) of the validation populations when 45k SNP effects based on one breed is used to predict within breed (diagonal) or to predict across-breed (off-diagonal). Results are for the smaller effective population scenario.

Breed Effects	**Breed Predicted**				
Used	**Breed A**	**Breed B**	**Breed C**	**Breed D**	**Breed E**
Breed A	**0.67 ± 0.01**	0.15 ± 0.04	0.11 ± 0.04	0.10 ± 0.03	0.15 ± 0.03
Breed B	0.12 ± 0.02	**0.71 ± 0.01**	0.12 ± 0.04	0.11 ± 0.03	0.13 ± 0.03
Breed C	0.09 ± 0.01	0.10 ± 0.02	**0.72 ± 0.03**	0.07 ± 0.04	0.13 ± 0.03
Breed D	0.10 ± 0.02	0.16 ± 0.03	0.12 ± 0.04	**0.73 ± 0.01**	0.07 ± 0.03
Breed E	0.13 ± 0.02	0.10 ± 0.02	0.11 ± 0.02	0.13 ± 0.04	**0.69 ± 0.02**

Table 3.4: Accuracies obtained for breeds A-E with smaller Ne using 9k and 45k SNP markers. Single-breed evaluations were performed as well as multi-breed. For multi-breed, SNP effects were first assumed to be the same in a shared scenario, and then SNP effects were treated as different in a non-shared scenario.

SNP Density	**Breed**	**Single-breed**		**Multi-breed shared**		**Multi-breed non-shared**	
		Acc	**SE**	**Acc**	**SE**	**Acc**	**SE**
9k	Breed A	0.63	0.01	0.54	0.01	0.63	0.01
	Breed B	0.63	0.01	0.54	0.03	0.60	0.00
	Breed C	0.70	0.03	0.63	0.06	0.72	0.04
	Breed D	0.74	0.02	0.67	0.04	0.73	0.01
	Breed E	0.57	0.01	0.48	0.03	0.58	0.01
	Average	**0.65**	**0.01**	**0.57**	**0.04**	**0.65**	**0.02**
45k	Breed A	0.67	0.01	0.68	0.01	0.67	0.01
	Breed B	0.71	0.01	0.71	0.01	0.69	0.01
	Breed C	0.72	0.03	0.72	0.02	0.71	0.03
	Breed D	0.73	0.01	0.74	0.01	0.73	0.01
	Breed E	0.70	0.02	0.71	0.02	0.69	0.02
	Average	**0.70**	**0.02**	**0.71**	**0.01**	**0.70**	**0.02**

Table 3.5: Accuracies obtained for breeds A-E with larger Ne using 9k and 45k SNP markers. Single-breed evaluations were performed as well as multi-breed. For multi-breed, SNP effects were first assumed to be the same in a shared scenario, and then SNP effects were treated as different in a non-shared scenario.

SNP Density	**Breed**	**Single-breed**		**Multi-breed shared**		**Multi-breed non-shared**	
		Acc	**SE**	**Acc**	**SE**	**Acc**	**SE**
9k	Breed A	0.61	0.01	0.51	0.01	0.62	0.01
	Breed B	0.63	0.01	0.52	0.03	0.61	0.00
	Breed C	0.63	0.03	0.49	0.06	0.63	0.04
	Breed D	0.70	0.02	0.60	0.04	0.69	0.01
	Breed E	0.60	0.01	0.49	0.03	0.59	0.01
	Average	**0.64**	**0.01**	**0.52**	**0.04**	**0.63**	**0.02**
45k	Breed A	0.65	0.02	0.66	0.02	0.65	0.02
	Breed B	0.66	0.00	0.66	0.01	0.65	0.00
	Breed C	0.73	0.03	0.71	0.03	0.73	0.04
	Breed D	0.76	0.01	0.75	0.01	0.74	0.01
	Breed E	0.60	0.01	0.60	0.01	0.60	0.01
	Average	**0.68**	**0.01**	**0.68**	**0.02**	**0.67**	**0.02**

Table 3.6: Bias measured as the regression coefficient when true breeding value (TBV) is regressed over genomic estimated breeding value (GEBV) for single breed evaluations and multi-breed evaluations using shared or non-shared SNP effects, 9k and 45k SNP markers in breeds with a smaller or larger effective population size (Ne)

		Single	Multi-breed (Shared)	Multi-breed (Non-shared)
9k	Smaller Ne	0.96 ± 0.03	0.87 ± 0.03	0.94 ± 0.03
	Larger Ne	0.92 ± 0.01	0.76 ± 0.03	0.89 ± 0.02
45k	Smaller Ne	0.99 ± 0.03	0.98 ± 0.03	0.97 ± 0.03
	Larger Ne	0.98 ± 0.02	0.94 ± 0.02	0.95 ± 0.02

FIGURES

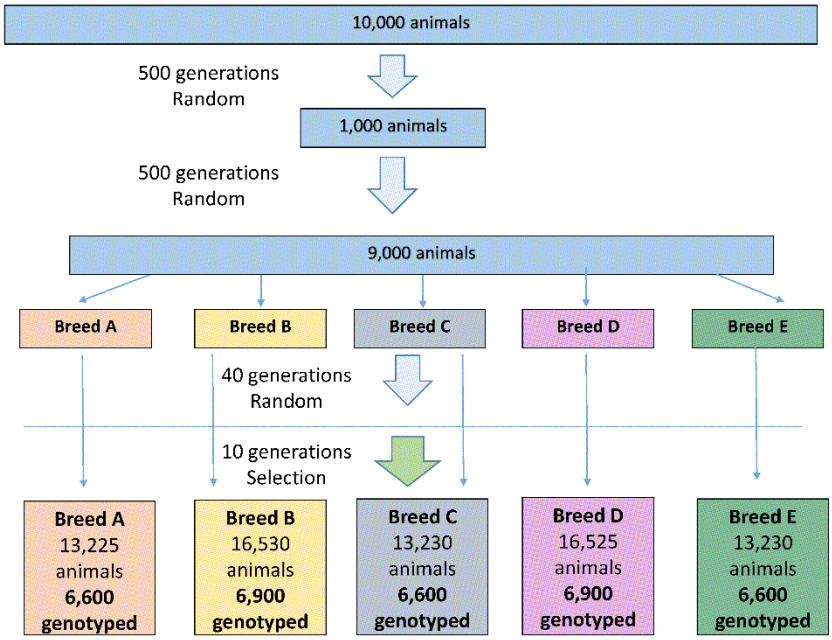


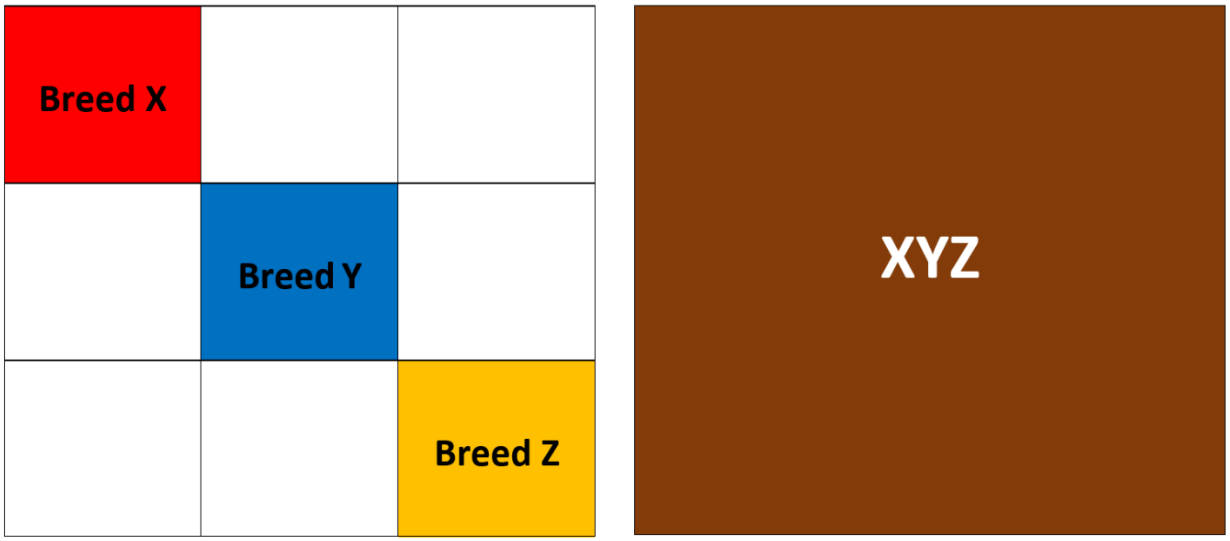
Figure 3.1: Visual presentation of the simulated data. The historic population of 10,000 animals was mated randomly for 1,000 generations, undergoing a bottleneck in generation 500. Founder animals for five breeds were selected and mated randomly for 40 generations followed by 10 generations of selection, resulting in different breed sizes and selected number of genotyped animals.

Shared	SNP ID		
Breed	1	2	3
Breed X			
Breed Y			
Breed Z			

Non-shared	SNP ID								
Breed	1	2	3	1	2	3	1	2	3
Breed X									
Breed Y									
Breed Z									

Figure 3.2: A graphic presentation of the SNP file for the shared and non-shared scenarios using three hypothetical breeds (X,Y and Z) corresponding to the three primary colors (red, blue and yellow), and only 3 SNP markers for all animals. When SNP effects were shared, the number of SNPs in the file is 3 and all animals have non-missing markers that overlap completely. When SNPs were treated as non-shared, the total number of SNPs in the file is 9 (3 SNPs x 3 breeds) and animals from each breed have 6 missing SNPs. Although physically all animals have SNPs in the same position on the chromosome, the file treats them as if they are in different, non-overlapping positions.

Figure 3.3: A visual presentation of the genomic relationship matrix (G) in a shared and non-shared scenario using 3 hypothetical breeds (X, Y and Z) corresponding to 3 primary colors (red, blue and yellow). In the shared scenario, the genotypes of all breeds are scaled to a single allele-frequency base (assuming a correlation of 1 between breeds) and all values are based on the combined information from all breeds. In the non-shared scenarios, there are no SNPs in common and therefore each breed is based and centered according to its own allele frequencies and animals from different breeds are not genetically correlated to each other. The G matrix has mostly zero elements.



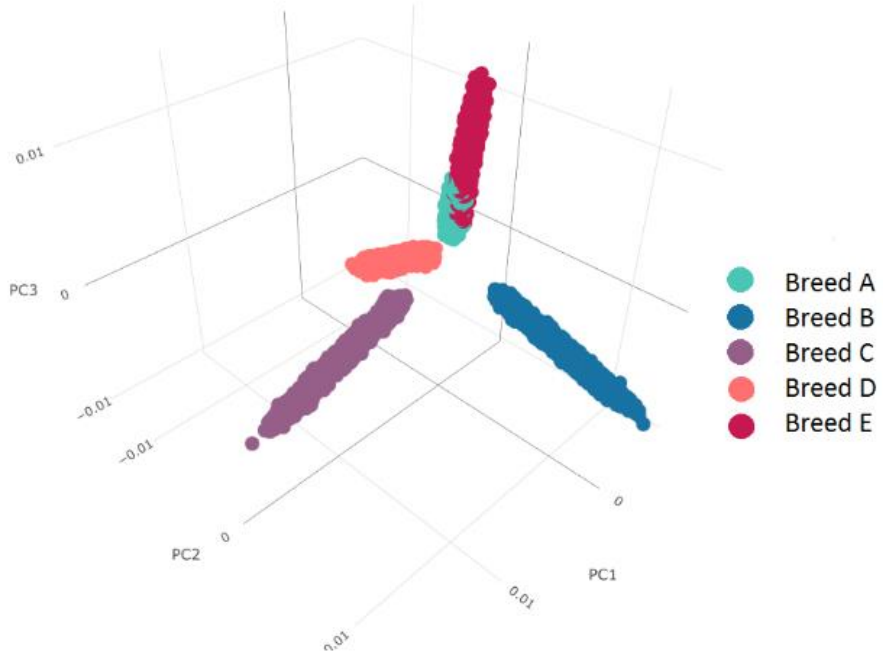


Figure 3.4: A principal component analysis with 1st, 2nd and 3rd principal components (PC) with the 5 different simulated breeds in one replicate using 45k SNP markers.

CHAPTER 4

OPTIMIZING THE REFERENCE POPULATION FOR VARIANT IMPUTATION IN CROSSBRED DAIRY CATTLE POPULATIONS

---

[1]Y.Steyn, D. Gonzalez-Pena, N. Vukasinovic, S.K. DeNise, D.A.L. Lourenco, I. Misztal. 2021. To be submitted to the Journal of Dairy Science

ABSTRACT

The objective of this study was to evaluate the accuracy of imputation of genotypes of crossbred animals using different reference populations and different SNP densities. Medium density genotypes (approximately 41,000 SNP markers) were available on 795 Holstein x Jersey crossbred animals as well as about 21,000 and 1,000 purebred Holstein and Jersey animals, respectively. Imputation accuracies were investigated from four different SNP panels – 3k (2 901 markers), LD (6 910 markers), ZL4 (18 820 markers), ZL5 (35 335 markers). This study evaluated three different reference populations – 500 crossbreds, 500 Holstein, 500 Jersey, and all three of these combined. The target population consisted of randomly selected 295 crossbred animals. Using a reference of crossbred animals resulted in the same average (0.91, 0.97, 0.98, and 0.98 for 3k, LD, ZL4 and ZL5 SNP panels, respectively) and maximum imputation accuracies (1.00 for all panels) as using a combined group. However, minimum accuracies were higher using only crossbreds (0.55, 0.60. 0.67 and 0.68 for the SNP panels in the same aforementioned order) compared to a combination (0.52, 0.58, 0.65 and 0.65). Jersey animals were considerably better as a reference group (averages 0.91, 0.97, 0.98 and 0.98) compared to Holstein (averages 0.81, 0.91, 0.94 and 0.94). This reflects the greater similarity to the Jersey breed based on available breed proportions, because the target crossbred animals had an average Jersey proportion of 0.63. The results suggest that using crossbred information improves the imputation accuracy rather than considering only information on each pure breed.

Key Words: imputation accuracy, Jersey, Holstein, genomic selection

INTRODUCTION

Genomic selection greatly improves accuracy of prediction in dairy cattle (Hayes et al., 2009, VanRaden et al., 2009). Although genotyping costs are becoming affordable, the majority

of animals in genomic evaluation are still being genotyped with lower density chips and imputed to higher densities (Hayes et al., 2012). Accuracy of imputation may depend on the statistical method used, minor allele frequency (MAF) of the markers to be imputed, linkage disequilibrium between markers, discrepancy in marker densities between high and low density panels, size and composition of the reference population, and the degree of relatedness to the target population (Schrooten et al., 2014; Lashmar et al., 2019).

The pig industry commonly uses crossbreeding strategies to obtain a commercial population that exploits heterosis (Christensen et al., 2015). The beef industry also makes use of composite breeds and crosses and therefore most imputation studies have been done in these fields (pigs – Duerte et al., 2013; Xiang et al., 2015; beef – Ventura et al., 2014; Chud et al., 2015; Wang et al., 2016). Studies focusing on imputation accuracy on dairy cattle have been less common (Oliveira Junior et al., 2017, Aliloo et al., 2018). Generally, crossbreeding is less common in the dairy industry, however, more dairy producers are choosing crossbreeding in order to improve efficiency of dairy animals (Olson et al., 2012; Shonka-Martin et al., 2019) and reduce potentially undesirable effects of inbreeding (Hazel et al., 2017). Inclusion of crossbred animals in the genomic evaluation is essential to make the correct selection decisions.

Imputation is generally more accurate when the animals to be imputed are of the same breed as the reference population (Berry et al., 2014, Moghaddar et al., 2014, Xiang et al., 2015). This creates a challenge for crossbred animals - these animals will have haplotypes originating from different purebred ancestors while imputation accuracy greatly depends on the proportion of shared haplotypes between the reference and target populations (Xiang et al., 2015). The objective of this study was to evaluate the accuracy of imputation of genotypes of crossbred animals using different reference populations and different SNP densities.

Data for US Holstein and Jersey dairy cattle and their crosses were provided by Zoetis. Genotypes were available for over 22k crossbred animals, but only 795 had 40k or more SNP markers, with a maximum of 41 008. Raw genotypes were previously edited using criteria described in Wiggans et al. (2011). Imputation was therefore applied to reach ~41k markers. None of the 795 crossbred animals had genotyped parents. The target population consisted of 295 crossbred animals randomly selected from the 795 crossbred animals. The remaining 500 crossbred animals were used as a crossbred reference population (CROSS). Breed proportions were obtained from the Council on Dairy Cattle Breeding (CDCB). The average proportion of Jersey present in the 795 crossbred animals was 63.11%. The crossbred training population had an average Jersey proportion of 62.72%, and the validation population had 63.76%.

To represent each group equally, the Holstein reference population consisted of 500 animals (HOL), and the Jersey reference population of 500 Jersey animals (JER), randomly selected from a set of about 21,000 and 1,000 Holstein and Jersey animals. These three populations were combined (COMB) to have a reference population that included all breeds and crosses (1,500 animals).

Marker names and positions for four different Illumina SNP chips were used to select and impute markers. The first was the Bovine 3k Beadchip with 2,901 SNP markers, BovineLD with 6,910 markers, a custom Zoetis SNP chip (ZL4) with 18,820 markers, and a custom Zoetis SNP chip (ZL5) with 35,335 markers. The original genotypes of 295 target animals were reduced to mimic the panels described above. Pedigree information was not used because previous research by Zoetis showed that it does not provide benefits to imputation accuracy (results not published). FIMPUTE software (Sargolzaei et al., 2014) was used for imputation. Accuracy of imputation was measured as concordance, i.e., the proportion of markers that were correctly imputed.

The minimum, maximum and mean concordance for all scenarios are presented in Table 4.1. The number of animals within a certain range of concordance are presented in Table 4.2. Imputation accuracy was lowest when imputing from 3k and LD. The imputation accuracy essentially remains unchanged from ZL4 and ZL5.

On average, using COMB as the reference was as successful as using CROSS and both scenarios could impute the genotypes of some animals perfectly regardless of the density imputed from. However, using CROSS gave slightly higher minimum concordance – 0.55 vs 0.52, 0.60 vs 0.58, 0.67 vs 0.65, and 0.68 vs 0.65 for 3k, LD, ZL4, and ZL5, respectively. In all different SNP panel scenarios, three more animals reached concordance over 0.95 when using CROSS compared to COMB (285 vs 282).

Accuracy of imputation was considerably higher when using JER as a reference instead of HOL, which reflects the level of relatedness based on breed proportion. Previous studies also found that imputation becomes more accurate as the relatedness between the reference and target population increases (Ventura et al., 2014, Xiang et al. 2015). Moghaddar et al. (2015) found an accuracy of 0.88 when Merino-cross animals were imputed from 3,000 crossbred animals but increased to 0.96 when the reference population was strategically selected instead of being random. The highest average obtained when using HOL as reference was 0.76, whereas the highest average using JER was 0.94. No animals imputed from 3k or LD had 0.95 or higher concordance when using HOL, and only two when imputing from ZL4 and ZL5. When using JER as reference, 117 animals had a concordance of 0.95 or more when imputing from ZL4, and 122 when imputing from ZL5. In crossbred dairy in Brazil, Oliveira Junior et al. (2017) also found that including crossbred animals in the reference population, whether by themselves or in combination with one or all component pure breeds deliver the best imputation accuracy. However, in their study, only

including component breeds delivered almost identical results. Work on pigs by Xiang et al. (2015), found that only including component breeds without any cross animals deliver great results, but it was not compared to a reference population that also included crossbred animals.

The accuracy of imputation is essential to obtain reliable genomic predictions for crossbred animals. Moghaddar et al. (2014) found that moderately well (0.62 to 0.86 accuracy) imputed genotypes (12k to 50k) still gave higher prediction accuracy than using 12k observed markers. When imputation was poor, it was more accurate to use 12k observed markers than 50k imputed ones. Bolormaa et al. (2015) found that, provided the imputation accuracy is greater than 0.90, there is no apparent difference in prediction accuracy of genomic selection when using 50k observed or imputed markers. In this study, most animals had a concordance of 0.90 or higher when using only crossbred animals, or a combination of all pure and crossbred animals. This is true even when imputing from low density panel. It should be mentioned that all crossbred animals in this study originated from one single farm and therefore were more related to each other than the average population, which may have an influence on the results. This could be reflective of recently formed composite breeds where animals are closely, but not directly, related to the pure component breeds, and more closely related to each other.

In conclusion, we found that, in order to accurately impute genotypes of crossbred animals, the reference population should contain crossbred animals. Similar accuracies can be obtained by combining the purebreds and crossbreds together in a single reference population. However, using only a pure breed to impute crossbred animals did not result in an adequate imputation accuracy. Imputation accuracies when imputing from 3k or LD were lower compared to ZL4 and ZL5.

ACKNOWLEDGMENTS

The authors declare that they have no competing interests. This study was partially supported by grants from Zoetis, and by Agriculture and Food Research Initiative Competitive Grants no. 2015-67015-22936 from the US Department of Agriculture's National Institute of Food and Agriculture.

REFERENCES

Aliloo, H., R. Mrode, A.M. Okeyo, G. Ni, M.E. Goddard, and J.P. Gibson. 2018. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. J. Dairy Sci.

Berry, D. P., M. C. McClure, and M. P. Mullen. 2014. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. J Anim Breed Genet 131(3):165-172.

Bolormaa, S., K. Gore, J. Van Der Werf, B. Hayes, and H. Daetwyler. 2015. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Anim Genet 46(5):544-556.

Christensen, O.F., P. Madsen, B. Nielsen, and G. Su. 2015. Genomic evaluation of both purebred and crossbred performances. Genet Sel Evol 46:23

Chud, T.C.S., R.V. Ventura, F.S. Schenkel, R. Carvalheiro, M.E. Buzanskas, J.O. Rosa, M.A. Maduda, M.V.G.B da Silva, F.B. Mokry, C.R. Marcondes, L.C.A. Regitano, and D.P. Munari. 2015. Strategies for genotype imputation in composite beef cattle. BMC Genet 16:99

Duarte, J. L., R. O. Bates, C. W. Ernst, N. E. Raney, R. J. C. Cantet, and J. P. Steibel. 2013. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genet 14(1):38.

Hazel, A. R., B. J. Heins, and L. B. Hansen. 2017. Fertility, survival, and conformation of Montbéliarde × Holstein and Viking Red × Holstein crossbred cows compared with pure Holstein cows during first lactation in 8 commercial dairy herds. J. Dairy Sci. 100:9447–9458.

Hayes, B., P. Bowman, H. Daetwyler, J. Kijas, and J. Van der Werf. 2012. Accuracy of genotype imputation in sheep breeds. Anim Genet 43(1):72-80.

Hayes, B. J., P. J. Bowman, A. Chamberlain, and M. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci 92(2):433-443.

Lashmar, S., F. Muchadeyi, and C. Visser. 2019. Genotype imputation as a cost-saving genomic strategy for South African Sanga cattle: A review. S Afr J Anim Sci 49(2):262-280.

Moghaddar, N., K. P. Gore, H. D. Daetwyler, B. J. Hayes, and J. H. J. van der Werf. 2015. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. Genet Sel Evol 47(1):97.

Moghaddar, N., A. A. Swan, and J. H. Van Der Werf. 2014. Comparing genomic prediction accuracy from purebred, crossbred and combined purebred and crossbred reference populations in sheep. Genet Sel Evol 46(1):58.

Oliveira Júnior, G. A., T. C. S. Chud, R. V. Ventura, D. J. Garrick, J. B. Cole, D. P. Munari, J. B. S. Ferraz, E. Mullart, S. DeNise, S. Smith, and M. V. G. B. da Silva. 2017. Genotype imputation in a tropical crossbred dairy cattle population. J. Dairy Sci., 100(12):9623-9634.

Olson,K. M., B. G. Cassell, M. D. Hanigan, and R. E. Pearson. 2012. Short communication: Interaction of energy balance, feed efficiency, early lactation health events, and fertility in first-lactation Holstein, Jersey, and reciprocal F1 crossbred cows. J. Dairy Sci. 94:507-511.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15(1):478.

Schrooten, C., R. Dassonneville, V. Ducrocq, R. F. Brøndum, M. S. Lund, J. Chen, Z. Liu, O. González-Recio, J. Pena, and T. Druet. 2014. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip. Genet Sel Evol 46(1):10.

Shonka-Martin, B.N., A.R. Hazel, B.J. Heins, L.B. Hansen. 2019. Three-breed rotational crossbreds of Montbéliarde, Viking Red, and Holstein compared with Holstein cows for dry matter intake, body traits, and production J. Dairy Sci, 102:871-882.

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

VanRaden, P. M., M. E. Tooker, T. C. S. Chud, H. D. Norman, J. H. Megonigal, I. W. Haagen, and G. R. Wiggans. 2020. Genomic predictions for crossbred dairy cattle. Journal of Dairy Science 103(2):1620-1631.

Ventura, R. V., D. Lu, F. S. Schenkel, Z. Wang, C. Li, and S. P. Miller. 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J Anim Sc 92(4):1433-1444.

Wang, Y., G. Lin, C. Li, and P. Stothard. 2016. Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. Springer Sci Rev 4(2):79-98.

Wiggans, G. R., P. M. VanRaden, and T. A. Cooper. 2011. The genomic evaluation system in the United States: Past, present, future. J. Dairy Sci. 94:3202–3211.

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

TABLES

Table 4.1: Summary statistics for accuracy of imputation, measured as the proportion of SNP markers correctly imputed when using four different reference groups described in the footnotes. The target population was 295 Jersey-Holstein crossbred animals and four different SNP panels were used to impute to 41k markers from – 3k, low density (LD), and two customized Zoetis chips (ZL4 and ZL5).

Reference[1]	SNP chip[2]	Minimum	Mean	Maximum
CROSS	3k	0.55	0.91	1.00
	LD	0.60	0.97	1.00
	ZL4	0.67	0.98	1.00
	ZL5	0.68	0.98	1.00
HOL	3k	0.51	0.61	0.85
	LD	0.57	0.70	0.94
	ZL4	0.65	0.75	0.96
	ZL5	0.66	0.76	0.97
JER	3k	0.54	0.81	0.94
	LD	0.60	0.91	0.98
	ZL4	0.67	0.94	0.99
	ZL5	0.68	0.94	0.99
COMB	3k	0.52	0.91	1.00
	LD	0.58	0.97	1.00
	ZL4	0.65	0.98	1.00
	ZL5	0.65	0.98	1.00

[1] CROSS consisted of 500 Jersey-Holstein crossbred animals, HOL of 500 purebred Holstein animals, JER of 500 purebred Jersey animals, and COMB of all three of these reference populations

[2] The 3k panel had 2 901 SNP markers, LD had 6 910, ZL4 had 18 820, and ZL5 had 38 335

Table 4.2: The number of animals with an imputation accuracy less than or equal to 0.90, greater than 0.90 and less than 0.95, and more than 0.95. The target population was 295 crossbred animals and four different reference populations were used, as described in the footnote. Four different SNP panels were used to impute to 41k markers from – 3k, low density (LD), and two customized Zoetis chips (ZL4 and ZL5).

		Number of animals within concordance range		
Reference[1]	SNP chip[2]	≤ 0.90	> 0.90 to 0.95	> 0.95
CROSS	3k	63	195	37
	LD	8	12	275
	ZL4	8	2	285
	ZL5	8	2	285
HOL	3k	295	0	0
	LD	295	0	0
	ZL4	293	2	2
	ZL5	293	2	2
JER	3k	275	20	0
	LD	101	150	44
	ZL4	14	164	117
	ZL5	15	158	122
COMB	3k	94	168	33
	LD	12	12	271
	ZL4	11	2	282
	ZL5	11	2	282

[1]CROSS consisted of 500 Jersey-Holstein crossbred animals, HOL of 500 purebred Holstein animals, JER of 500 purebred Jersey animals, and COMB of all three of these reference populations[2]

[2] The 3k panel had 2 901 SNP markers, LD had 6 910, ZL4 had 18 820, and ZL5 had 38, 335

CHAPTER 5

INDIRECT GENOMIC PREDICTIONS FOR MILK YIELD IN CROSSBRED HOLSTEIN-JERSEY DAIRY CATTLE[1]

---

[1]Y.Steyn, D. Gonzalez-Pena, Y.L. Bernal Rubio, N. Vukasinovic, S.K. DeNise, D.A.L. Lourenco, I. Misztal. 2021. *Journal of Dairy Science.* 104(5):5728-5737. Reprinted here with permission of publisher.

ABSTRACT

The objective of this study was to predict genomic breeding values for milk yield of crossbred dairy cattle under different scenarios using single-step genomic BLUP (ssGBLUP). There were 13,880,217 milk yield measurements on 6,830,415 cows. Genotypes of 89,558 Holstein, 40,769 Jersey, and 22,373 Holstein-Jersey crossbred animals were used, of which all Holstein, 9,313 Jersey, and 1,667 crossbred animals had phenotypic records. Genotypes were imputed to 45k SNP markers. The SNP effects were estimated from single-breed evaluations for Jersey (JE), Holstein (HO) and crossbreds (CROSS), and multi-breed evaluations including all Jersey and Holstein (JE_HO) or approximately equal proportions of Jersey, Holstein and crossbred animals (MIX). Indirect predictions (IP) of the validation animals (358 crossbred animals with phenotypes excluded from evaluations) were calculated using the resulting SNP effects. Additionally, breed proportions (BP) of crossbred animals were applied as a weight when IP were estimated based on each pure breed. The predictive ability of IP was calculated as the Pearson correlation between IP and phenotypes of the validation animals adjusted for fixed effects in the model. Regression of adjusted phenotypes on IP was used to assess the inflation of IP. The predictive ability of IP for CROSS, JE, HO, JE_HO and MIX scenario was 0.50, 0.50, 0.47, 0.50, and 0.46, respectively. Using BP was least successful, with a predictive ability of 0.32. The inflation of the IP for crossbred animals using CROSS, JE, HO, JE_HO, MIX, and BP scenarios were 1.17, 0.65, 0.55, 0.78, 1.00, and 0.85, respectively. The IP of crossbred animals can be predicted using ssGBLUP under a scenario that includes pure breed genotypes.

Keywords*:* single-step GBLUP, breed proportions, direct genomic value, independent chromosome segments

INTRODUCTION

Genomic breeding values are typically estimated within pure breeds, especially in dairy cattle. Interest in combined purebred and crossbred evaluation is limited to countries with a large number of crossbred animals that are potentially used for breeding., e.g. in New Zealand (Harris and Johnson, 2010). As of May 2019, the total number of genotyped dairy cattle in the U.S. exceeded 3 million, of which only about 2% were of crossbred cattle (VanRaden et al., 2020). By September 2020, the number of genotyped US dairy animals increased to over 4.5 million, of which 86% are Holstein and 12% are Jersey animals (CDCB, 2020a). Although the proportion of crossbred animals is small, it amounts to a substantial financial cost that requires a return on investment. Crossbreds are becoming increasingly popular. An analysis of trends in the breed composition of U.S. Dairy Herd Improvement (DHI) herds showed that the percentage of dairy cattle reported as crossbred increased from 0.1% to 5.3% from 1990 to 2018 (Guinan et al., 2019). In April 2019, the Council on Dairy Cattle Breeding (CDCB) extended genomic evaluation services to provide estimates for crossbred animals (Wiggans et al., 2019, CDCB, 2020b).

Joint modeling of purebreds and crossbreds may require adjustments to account for non-additive effects and heterogeneous variance of the breeds (Wei and van der Werf, 1994; Christensen et al., 2014). For purebred parents and F_1 crossbred animals, an optimum strategy may be based on separating the genomic information for F_1 due to each parent. However, simpler methods based on combining all genotypes in a single relationship matrix may work as well (Lourenco et al., 2016). Another possibility is providing evaluations based on purebreds and estimated breed proportions (VanRaden et al., 2020).

Genomic evaluation works primarily by estimating the value of chromosome segments (Daetwyler et al., 2010; Habier et al., 2013; Pocrnic et al., 2016a; Pocrnic et al., 2019a). Such segments are different for each pure breed and probably only partially overlap with those of purebreds for crossbreds. In such cases, it would be useful to include the crossbred data for crossbred prediction.

Single-step genomic BLUP (ssGBLUP) is widely used for different species in numerous countries. This method includes both genotyped and non-genotyped animals independently of the phenotyping status (Legarra et al., 2009). Therefore, it allows the joint evaluation of purebred- and crossbred animals, as well as separate evaluations. Additionally, there is an interest to apply genomic selection in commercial crossbred animals that may not be part of the official evaluation (i.e., not registered). These animals could have indirect genomic predictions (IP) computed based on SNP effects back-solved from the official evaluation. The objective of this study was to evaluate the predictive ability and inflation of indirect genomic predictions for crossbred animals using SNP effects estimated with ssGBLUP methods. Genotypes of Holsteins, Jerseys, and Holstein-Jersey crossbreds were used in ssGBLUP for the estimation of SNP effects. Additional comparisons involved a method that incorporates breed proportions based on genomic information.

MATERIALS AND METHODS

The trait of interest was milk yield and 13,880,217 records were available on 6,830,415 animals, consisting of Holstein, Jersey, and animals classified as Holstein-Jersey crossbreds. Phenotypic and pedigree data were directly obtained from producers in the USA through on-farm software using proprietary scripts. Genotypes were obtained from the Zoetis Genotyping Lab (Zoetis Genetics, Kalamazoo, MI) and a variety of low density chips with a number of SNP

(ranging from about 3,000 to over 35,000, and medium density chips with 50-80K markers) were used. Animals genotyped on chips with less that 40K SNPs were imputed to 45,245 markers using the program FImpute (Sargolzaei et al., 2011). Details on the data sources is provided in Vukasinovic et al. (2017). The crossbred animals were imputed using a reference population of 795 Holstein-Jersey crossbred animals. The number of animals genotyped (and phenotyped) were 375,487 (89,558) Holstein, 40,769 (9,313) Jersey, and 22,373 (1,667) crossbred animals. Among the parents of crossbred animals, 1,331 Holstein sires and 202 Holstein dams, 147 Jersey sires and 122 Jersey dams were genotyped. In total, 9% of the purebred parents of crossbred animals were genotyped. Breed proportions (BP) were obtained for crossbreds based on genotypes using an ADMIXTURE (Alexander et al., 2009) analysis supervised with two clusters. The available crossbreds were of varying breed proportions (1%-99% Holstein or Jersey). Figure 5.1(A) shows the distribution for the Holstein proportion in all crossbred animals. The average Holstein BP for all genotyped crossbred animals was 59% and Jersey BP was 41%. Figure 5.2. shows the first two principal components (PC) for all crossbred animals. All genotyped Jersey and crossbred animals were used and 89,558 Holstein animals with phenotypes and genotypes were selected to lower computational requirements.

Validation animals for milk yield consisted of animals that had measurements for the first lactation only, and were born within 2015 to 2017, such that the validation populations were 20% of all the genotyped animals. There were three different validation populations – for Holstein (15,695 animals), Jersey (2,186 animals), and crossbreds (358 animals). Among the parents of the 358 validation crossbred animals, 61 Holstein sires and 10 Holstein dams, and 26 Jersey sires and 19 Jersey dams were genotyped. This amounted to 32% of the known purebred parents of crossbred validation animals. Breed proportions of validation animals ranged from 1%

to 99% of either Holstein or Jersey. Figure 5.1(B) shows the distribution of the proportion of Holstein in validation crossbred animals. The average Holstein BP for crossbred validation animals was 48%, and Jersey BP was 52%. Phenotypes of the validation populations were removed to estimate SNP effects based on that breed/group. All evaluations included the same pedigree and data file of all available animals, but the selection of genotypes differed for each scenario.

The repeatability animal model for milk yield was:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z_1u} + \mathbf{Z_2pe} + \mathbf{Z_3hs} + \mathbf{e}$$

where **y** is the vector of milk yield, **X** is the incidence matrix assigning the measurement to fixed effects that included age group, management group, pedigree-based inbreeding, and pedigree-based heterosis (obtained using the R package OptiSel (Wellman, 2019,2020)); **b** is a vector of solutions for fixed effects, $\mathbf{Z_1}$ is the incidence matrix assigning the measurement to the random animal effect; **u** is a vector of solutions for animal, $\mathbf{Z_2}$ is the incidence matrix for random permanent environment effect; **pe** is a vector of solutions for permanent environment effect; $\mathbf{Z_3}$ is an incidence matrix for the random effect of sire nested within herd; **hs** is a vector of solutions for herd × sire interaction, and **e** is the residual. It was assumed that $\mathbf{u} \sim N(0, \mathbf{H}\sigma_u^2)$ where σ_u^2 = additive genetic variance and **H** is the relationship matrix combining genotyped and non-genotyped animals in single-step GBLUP (ssGBLUP) (Legarra et al., 2009); $\mathbf{pe} \sim N(0, \mathbf{I}\sigma_{pe}^2)$ where σ_{pe}^2= permanent environment variance; $\mathbf{hs} \sim N(0, \mathbf{I}\sigma_{hs}^2)$ where σ_{hs}^2= herd × sire variance; $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ and **I** is the identity matrix. A heritability of 0.30 was assumed for milk yield (Wiggans, 1997).

The inverse of **H**, which is required for the ssGBLUP evaluations was constructed as in Aguilar et al. (2010):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix}$$

In this study, **G** was obtained using the formula $\mathbf{G} = \frac{\mathbf{MM'}}{2\sum p_i(1-p_i)}$ where **M** is a matrix of SNP content centered by twice the current allele frequencies, and p_i is the allele frequency for SNP i (VanRaden, 2008). The pedigree-based relationship matrix between genotyped animals is referred to as $\mathbf{A}_{22}$. To reduce bias due to the different genetic level of genotyped and non-genotyped animals, **G** was tuned to be compatible with $\mathbf{A}_{22}$ using the method described by Chen et al. (2011). To avoid singularity problems, 5% of $\mathbf{A}_{22}$ was combined with 95% of **G**.

The **G** was constructed differently for each scenario to account for the difference in allele frequencies. Three evaluations were within-breed – Jersey (JER), Holstein (HOL), and crossbreds (CROSS). The respective **G** matrices were constructed using within-breed allele frequencies. Two evaluations were multi-breed and used the relevant genotypes to construct different **G** matrices – 1) all Jersey and Holstein genotypes used as reference (JER_HOL) with **G** constructed using allele frequencies of all Jersey and Holstein combined, and 2) equal proportions (~20k each) of Jersey, Holstein and crossbreds (MIX) with **G** constructed using allele frequencies of only these animals. Because the number of genotyped animals was large in all scenarios, except CROSS, the inverse of **G** was computed using the APY algorithm with core animals selected randomly (Fragomeni et al., 2015; Misztal, 2016). The number of core animals corresponded to approximately the number of eigenvalues explaining 99% of the variation in **G**. The number of eigenvalues explaining 90%, 95%, 98%, and 99% are presented in Table 5.1. The direct inverse of **G** was only used for the CROSS evaluation.

Single-step genomic BLUP implemented in the BLUPF90 software suite (Misztal et al., 2014), was used for analyses of these five scenarios. After removing phenotypes of all validation animals, genomic breeding values (GEBV) for all animals were estimated in each scenario. The

BLUP90IOD2 v3.102 software was used to compute these GEBV. The combination of breeds/groups together assumes that SNP effects and variances are the same across all populations. The SNP effects were estimated for each scenario separately based on these GEBV using the POSTGSF90 v1.63 software package with the formula (VanRaden, 2008; Wang et al., 2012):

$$\hat{\mathbf{a}} = \lambda \mathbf{D}\mathbf{M}'\mathbf{G}^{-1}(\text{GEBV})$$

where $\hat{\mathbf{a}}$ is a vector of estimated SNP effects, λ is the ratio of SNP to additive genetic variance, **D** is a diagonal matrix of weights for SNP (in this case an identity matrix), and **M** was defined before. The $\mathbf{G}^{-1}$ for JER, HOL, JER_HOL and MIX, were obtained using APY. The IP from ssGBLUP have been shown to be stable when using different core animals as long as the size of the core is at least equal to the number of eigenvalues required to explain 98-99% of the genomic variation (Garcia et al., 2020), which was the case in this study.

Based on SNP effects, IP for validation animals were calculated as the sum of SNP effects weighted by the genotype content using the PREDF90 v1.04 software. Across-breed predictions were obtained using SNP effects estimated in one scenario to predict the IP of validation animals not included in that evaluation. The IP obtained using different methods were compared to the GEBV obtained with full data in the CROSS ($GEBV_{Full\_CROSS}$) and MIX ($GEBV_{Full\_MIX}$) scenarios, as well as GEBV obtained using CROSS without data of the validation populations ($GEBV_{Partial}$).

Indirect predictions for crossbred animals were also estimated using breed proportions and IP obtained using SNP effects of both the component breeds with the following formula:

$IP_{BP} = BP_H(IP_H) + BP_J(IP_J)$

where IP_{BP} is the IP of crossbred animals using breed proportions, BP_H is the proportion of Holstein in the crossbred animal, and BP_J is the proportions of Jersey in the crossbred animal, IP_H is the IP of the crossbred animal estimated using Holstein SNP effects, and IP_J is the IP of the same crossbred animal estimated using Jersey SNP effects. This weighting of the IP with the breed proportion follows the same concept as VanRaden et al. (2020) and Strandén and Mäntysaari (2013). A requirement for this method is for IP to be on an all-breed scale, which was achieved by including phenotypes of all breeds and a full pedigree in every evaluation.

The predictive ability was determined using a Pearson correlation between the phenotype adjusted for all other effects, and the IP of the relevant validation population. The adjusted phenotype was obtained using the PREDICTf90 v1.3 software package, with the same model as described before, and data of all animals, regardless of breed. No genotypes were included for this purpose. Inflation was measured as the regression coefficient when regressing adjusted phenotype on IP. A coefficient of 1 indicates no inflation, above 1 indicates an under-estimation (deflation) and below 1 an over-estimation (inflation).

RESULTS

The number of eigenvalues required to explain 90%, 95%, 98% and 99% of the variation in the **G** matrix for the different scenarios are presented in Table 5.1. To apply APY, it is recommended to have a core size equal to the number of eigenvalues that explains between 98% and 99% of the variation of **G** (Pocrnic et al., 2016b). Since available resources were able to handle larger core sizes, the core was selected to explain at least 99%.

Results of the predictive abilities of indirect predictions are presented in Table 5.2. The predictive ability within-breed was 0.48, 0.45, and 0.50 for JER, HOL, and CROSS, respectively. The JER could not predict HOL well (0.13) and HOL had even lower ability to predict JER

(0.09), which is expected as across-breed predictions have been shown to have low accuracies in dairy cattle (Olson et al., 2012; Pryce et al., 2011). Both pure breeds could predict CROSS approximately the same as when using only crossbred animals (0.50 with JER effects, 0.47 with HOL, and 0.50 with CROSS). The CROSS was better at predicting the pure breeds (0.24 for JER and 0.26 for HOL) than the pure breeds were able to predict each other (across-breed), however the predictive ability was still low. The JER_HOL scenario could predict all three groups with relatively similar predictive ability as the single-breed analyses (0.45, 0.44, and 0.50 for Jersey, Holstein, and crossbreds, respectively). Making use of breed proportions was the least successful of all scenarios, with a predictive ability equal to 0.32.

The resulting IP for crossbred validation animals in the different scenarios were compared to the $GEBV_{Partial}$, $GEBV_{Full\_CROSS}$ and $GEBV_{Full\_MIX}$. Table 5.3 summarizes these correlations. The estimated GEBV were adjusted for the genetic base consisting of animals born in 2015. The means were -70.13 for IP based on CROSS, -929.48 for $GEBV_{Partial}$, -1003.76 for $GEBV_{Full\_CROSS}$ and -1083.12 for $GEBV_{Full\_MIX.}$ Figure 5.3 compare the distribution of adjusted phenotype to the IP of crossbred animals obtained from different approaches and $GEBV_{Partial}$. The means and standard deviations differ between scenarios, especially when crossbred animals were not part of the reference population. The distribution of the adjusted phenotypes is wide, while distributions are more centered around zero for the $GEBV_{Partial}$ and IP when SNP effects were based on CROSS, BP, and MIX. Their ranges are narrower compared to the IP based on JER, HOL or JER_HOL. The IP for crossbred validation animals are generally above zero when based on JER and below zero when based on HOL.

The inflation of the IP of crossbred animals when using SNP effects based on CROSS, JER, HOL, JER_HOL, and MIX was 1.17, 0.65, 0.55, 0.78, and 1.00. Inflation when IP was based on pure breeds or with breed proportions was 0.85.

DISCUSSION

The number of eigenvalues associated with the genomic relationship matrix is an indicator of the number of independent chromosome segments (M_e) (Pocrnic et al., 2016a). A smaller M_e indicates longer chromosome segments and less genetic diversity within the population. The Jersey breed required considerably fewer eigenvalues compared to the Holstein. However, this may be an artifact of the number of genotypes available for Jersey. In the case of JER_HOL, more eigenvalues were required compared to either JER or HOL, but less than the sum of eigenvalues obtained for those pure breeds separately. This deviation from the sum suggests that the Holstein and Jersey breeds share similarities (Pocrnic et al., 2019b) albeit not enough to do across-breed predictions. It is not unexpected for breeds, especially the ones with similar breeding goals (such as dairy breeds) to share similarities, such as the *DGAT1* gene (Spelman et al., 2002; Thaller et al., 2003). In fact, genetic similarities can occur across species (Raymond et al., 2020). The number of eigenvalues required in the MIX scenario was considerably higher than in the other scenarios. Since crossbred animals receive their genes from the component pure breeds, adding the crossbred genotypes is not expected to provide any additional genomic content information not already captured by purebred parents. However, crossbred animals in this study have a wide range of breed proportions instead of only F_1. This may produce new haplotypes, recombination, and LD not present in the pure breeds, thereby resulting in the higher number of eigenvalues needed to explain 98-99% of variation. More importantly, only a small proportion of known purebred parents were genotyped in our study.

Thus, our crossbred population provided additional information from purebred animals that do not have genotypes. The study by Pocrnic et al. (2019b) on pigs found that adding F_1 crossbred animals did not increase the number of eigenvalues beyond that of the purebred animals. However, the size of their data was larger than the M_e in pigs.

The low ability of one pure breed to predict the other has been observed in previous work, both on real and simulated data (Pryce et al., 2011; Olson et al., 2012; Raymond et al., 2018; Steyn et al., 2019). Results of our study show that when both purebreds are combined in the JER_HOL multi-breed evaluation, the predictive ability of the purebred validation animals was slightly lower than single-breed evaluations (0.01 lower for Holstein and to 0.03 for Jersey), which corresponds to other studies that did not find much of a difference (Olson et al., 2012; Pryce et al., 2011). The change in predictive ability for Holstein is small enough to be negligible. In multi-breed evaluations, it is important for component breeds to be present in the reference population (Toosi et al., 2010; Pryce et al., 2011; Olson et al., 2012; Steyn et al., 2019). The small, but slightly larger, decrease in predictive ability for Jersey could be because the number of Holstein genotypes far outweighs the number of Jersey genotypes in the JER_HOL reference population. When the number of genotypes per group is more balanced in the MIX scenario, the predictive ability of IP for Holstein decreased by 0.05 but only 0.02 for Jersey IP compared to single-breed evaluations.

In the JER_HOL, the number of Holstein animals was considerably higher than the number of Jersey animals, yet the crossbred IP were predicted better than when all groups were represented equally (MIX). The ability to predict IP for Holstein or crossbreds noticeably decreased in the MIX scenario compared to single-breed evaluations while predictive ability of Jersey IP changed very little. The change in Holstein but not Jersey, could be because the

number of Holstein genotypes removed for the MIX scenario was proportionally larger for Holstein (from ~90K to ~20K) compared to the Jersey genotypes (~40K to ~20K). More influential Holstein animals may have been removed, decreasing the average relatedness between the MIX reference population and the Holstein validation populations. Additionally, the Jersey population may be genomically less diverse and more related to the MIX reference population.

In pigs, it was shown that the accuracy for crossbred animals was higher when using both parent breeds or equal number of animals from the parent breeds, compared to using either pure breed by itself or only crossbreds for most traits (Hidalgo et al., 2015). Pocrnic et al. (2019b) also found an increase in predictive ability for crossbred pigs when both pure lines were combined compared to only one pure breed. However, their study was on F_1 crossbred animals and had more animals with both genotypes and phenotypes compared to our study. No increase in predictive ability of crossbred IP was observed in our study using JER_HOL compared to only using JER or CROSS, but there was a slight increase compared to using only HOL. The crossbreed structure of this study was complex with a range of breed proportions instead of F_1 crossbred animals. The purebred dairy breeds may also be genetically more diverse than the pig breeds, as reflected in the different effective population sizes in Pocrnic et al. (2016b). The MIX scenario included all the crossbred animals and equal proportions of each pure breed. This inclusion of purebred animals for crossbred evaluations is expected to yield higher predictive abilities but this was not the case in our study. The random sampling of purebred animals may have excluded influential parents that contributed to previous predictive abilities.

In the study by VanRaden et al. (2020), the quality of prediction for crossbreds was evaluated by squared correlations of later milk yield deviation on earlier prediction by breed base representation. For all crossbred cows, those correlations for several traits was 0.01 to 0.05

higher for genomic predictions using breed proportions than for parent average, indicating limited improvement. While the correlation was 0.05 higher for milk, it was much higher (up to 0.34) when separated by breed base representation. This shows the variation within crossbred evaluations, as the same approach to all can lead to different responses. Their study used much larger data for Holsteins and Jerseys, but crossbred animals were not used in the prediction process, only in validation.

In our study, accounting for breed proportions for the estimation of IP of crossbred animals led to a decrease in predictive ability compared to not making any adjustments at all (0.32 using proportions versus 0.50 using CROSS). This does not correspond to VanRaden et al. (2020), where using breed proportions was slightly more accurate than using the nearest pure breed. They also found that the accuracy was highest when the major breed proportion was from 75% to 90% (0.52), and lowest from 50% to 74% (0.35). The validation population in our study was, on average, 52% Jersey. The SNP effects in the study by VanRaden et al. (2020) were estimated from much larger data, while our study used approximately 2k crossbred animals with both genotypes and phenotypes. The relatively high predictive ability of crossbreds based on the crossbred SNP effects in this study despite the small reference population could be due to the limited number of purebred parents.

Predictive abilities assume that the model was adequate to adjust for all effects other than the additive genetic effect. Breed effect was not specifically included in the model. Animals are generally compared within breed and therefore breed adjustment will not compromise the ranking within the pure breed. Crossbred animals are most likely compared only with crossbreds to select replacement animals, but the range of breed proportions of animals adds a layer of complexity regarding breed effect. Accounting for non-additive factors and accounting for breed

origin of markers may increase the predictive ability for crossbred animals. However, Lopes et al. (2017) used a genomic matrix as described by Christensen et al. (2014) that takes breed-effects into account for the estimation of GEBV before back-solving for SNP effects. They found that accounting for breed-specific effects did not change the accuracy of prediction compared to only using effects of crossbreds. Simulations showed that accounting for breed origin of alleles is only beneficial when lines are distantly related (Esfandyari et al., 2015a; Esfandyari et al., 2015b), which might be the case with Jersey and Holstein. Obtaining breed-specific effects in those studies required knowledge of the breed origin of alleles. Assigning breed of origin to alleles achieved accuracies greater than 90% in both simulated and real data (Sevillano et al., 2016; Vandenplas et al., 2016).

The IP of crossbred animals were inflated when using HOL (0.55), slightly less using JER (0.65) and even less when using JER_HOL (0.78). Using CROSS was the only scenario with a deflated IP of crossbred animals (1.17). Although using MIX had a lower predictive ability than using JER_HOL or CROSS, the IP were neither inflated nor deflated. Inflation when using breed proportions was 0.85, which was better than using either pure breed or a combination of both, but not as successful as using MIX. This trade-off regarding predictive ability and inflation may influence the decision made by breed associations regarding the chosen reference population. Inflation in the study by VanRaden et al. (2020) was smaller when using breed proportions compared to using the nearest pure breed, which corresponds to this study.

Crossbred animals can be selected to be backcrossed with purebred animals with the objective to transform the current herd to a specific component pure breed (Holstein or Jersey in this study). In this case, the resulting progeny will be expected to perform within the genetic background of the pure breed. Therefore, SNP effects based on that pure breed could be useful

(JER or HOL). The performance of crossbreds is expected to be an intermediate of the parental breeds but often deviates from their average due to heterosis (Buckley et al., 2014). On average, Holstein milk production is considerably higher than Jersey milk production (CDCB, 2020c). This is reflected by the mean of the different IP, as presented in Figure 5.3. When IP are based on Jersey SNP effects for milk production, the IP for crossbreds are generally positive, while they are generally negative when using Holstein SNP effects. When IP were based on both Jersey and Holstein, the values were more intermediate.

The correlation between IP using CROSS and $GEBV_{Partial}$ is almost 1, which is expected since these GEBV were used to estimate the SNP effects. These GEBV were not used to estimate SNP effects for all other scenarios. Therefore, correlations between the other IP and $GEBV_{Partial}$ are considerably lower, but they are generally still high (over 0.65). The correlations between the IP from CROSS and $GEBV_{Full\_CROSS}$ and $GEBV_{Full\_MIX}$ are high (0.86 and 0.73), but lower than those with $CROSS_{Partial}$. These correlations are expected to be lower because more information became available to improve the predictions. It is important for these correlations to still be high since high correlations show that the IP was a strong indicator of a future breeding value. All correlations with IP based on breed proportions were much lower, confirming that this approach is not appropriate for this data. However, the results are affected by sample size and the number of crossbred animals in this data is relatively small. There were also differences in means, as shown in Figure 5.3. To compare or rank animals using predictions obtained from different sources of SNP information, they can be transformed to the same scale (Legarra et al., 2018; Lourenco et al., 2018).

In this study, less than 10% (24%) of the genotyped crossbred (purebred) animals had phenotypes. This is a challenge since the success of the estimation of SNP effects depends on the

number of animals with both phenotypes and genotypes (Gonzalez-Recio et al., 2014), however, it is important to research methods to accurately predict breeding values of young animals based on information that is present before phenotypes become available.

CONCLUSION

Indirect predictions for crossbred animals can be computed via single-step GBLUP using various combinations of purebred and crossbred data. The best reference population considering predictive ability and inflation of prediction is the mix of purebred and crossbred animals. The IP themselves provide a useful tool to select crossbred animals.

ACKNOWLEDGEMENTS

The authors want to acknowledge Zoetis Genetics providing the data for this study. The research described herein was a part of Ms. Steyn's internship program sponsored through the research agreement between UGA and Zoetis.

REFERENCES

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93(2):743-752. https://doi.org/10.3168/jds.2009-2730

Alexander, D., Novembre, J. and Lange, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19:1655-1664.

Boichard, D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, K. J. Gietzen, B. J. Hayes, C. T. Lawley, T. S. Sonstegard, C. P. Van Tassell, P. M. VanRaden, K. A. Viaud-Martinez, G. R. Wiggans, and L. D. C. for the Bovine. 2012. Design of a Bovine Low-Density SNP

Array Optimized for Imputation. PLOS ONE 7(3):e34130. doi: 10.1371/journal.pone.0034130

Buckley, F., N. Lopez-Villalobos, B.J. Heins. 2014. Crossbreeding: Implications for dairy cow fertility and survival. Animal 8(s1):122-133. doi: 10.1017/S1751731114000901

CDCB. 2020a. Council on Dairy Cattle Breeding Activity Report Oct19/Sep2020. https://www.uscdcb.com/wp-content/uploads/2020/10/2020-CDCB-Activity-Report_103020_lowres.pdf . Accessed Dec. 2, 2020

CDCB. 2020b. https://www.uscdcb.com/wp-content/uploads/2020/04/USApro.pdf . Accessed Aug.1,2020

CDCB. 2020c. https://queries.uscdcb.com/publish/dhi/current/laall.shtml . Accessed Oct.8,2020

Chen, C.-Y., I. Misztal, I. Aguilar, A. Legarra, and W. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci 89(9):2673-2679. doi: https://doi.org/10.2527/jas.2010-3555

Christensen, O. F., P. Madsen, B. Nielsen, and G. Su. 2014. Genomic evaluation of both purebred and crossbred performances. Genet Sel Evol 46(1):23. https://doi.org/10.1186/1297-9686-46-23

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185(3):1021-1031. https://doi.org/10.1534/genetics.110.116855

Esfandyari, H., A. C. Sørensen, and P. Bijma. 2015a. A crossbred reference population can improve the response to genomic selection for crossbred performance. Genet Sel Evol 47(1):76. https://doi.org/10.1186/s12711-015-0155-z

Esfandyari, H., A. C. Sørensen, and P. Bijma. 2015b. Maximizing crossbred performance through purebred genomic selection. Genet Sel Evol 47(1):16. https://doi.org/10.1186/s12711-015-0099-3

Fragomeni, B., D. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T. Lawlor, and I.

Illumina Inc. 2011a. GoldenGate Bovine3K Genotyping BeadChip. Accessed Sept. 29 2020. http://support.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_bovine3k.pdf

Illumina Inc. 2011b. BovineSNP50 Genotyping BeadChip. Accessed Sept. 29 2020. http://www.illumina.com/Documents/products/datasheets/datasheet_bovine_snp5O.pdf

Misztal. 2015. Hot topic: use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. J Dairy Sci 98(6):4090-4094. https://doi.org/10.3168/jds.2014-9125

Garcia, A. L. S., Y. Masuda, S. Tsuruta, S. Miller, I. Misztal, and D. Lourenco. 2020. Indirect predictions with a large number of genotyped animals using the algorithm for proven and young. J. Anim. Sci. 98(6). doi: 10.1093/jas/skaa154

Gonzalez-Recio, O., M. P. Coffey, and J. E. Pryce. 2014. On the value of the phenotypes in the genomic era. J Dairy Sci 97(12):7905-7915. https://doi.org/10.3168/jds.2019-16903

Guinan, F.L., Norman, H.D., and Dürr, J.W. 2019. Interbull Bull. 55:11-16

Habier, D., R. L. Fernando, and D. J. Garrick. 2013. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics 194(3):597-607. https://doi.org/10.1534/genetics.113.152207

Harris, B. L. and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J Dairy Sci 93(3):1243-1252. https://doi.org/10.3168/jds.2009-2619

Hidalgo, A. M., J. W. M. Bastiaansen, M. S. Lopes, B. Harlizius, M. A. M. Groenen, and D.-J. de Koning. 2015. Accuracy of Predicted Genomic Breeding Values in Purebred and Crossbred Pigs. G3.5(8):1575-1583. https://doi.org/10.1534/g3.115.018119

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci 92(9):4656-4663. https://doi.org/10.3168/jds.2009-2061

Legarra A., Lourenco D.A.L., Vitezica Z., Bases for genomic predictions [on line] (2018) http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=gsip.pdf

Lopes, M. S., H. Bovenhuis, A. M. Hidalgo, J. A. M. van Arendonk, E. F. Knol, and J. W. M. Bastiaansen. 2017. Genomic selection for crossbred performance accounting for breed-specific effects. Genet Sel Evol 49(1):51. https://doi.org/10.1186/s12711-017-0328-z

Lourenco, D., S. Tsuruta, B. Fragomeni, C. Chen, W. Herring, and I. Misztal. 2016. Crossbreed evaluations in single-step genomic best linear unbiased predictor using adjusted realized relationship matrices. J Anim Sci 94(3):909-919. https://doi.org/10.2527/jas.2015-9748

Lourenco D.A.L., Legarra A., Tsuruta S., Moser D., Miller S., Misztal I. Tuning indirect predictions based on SNP effects from single-step GBLUP. Interbull Bulletin. 2018;53:48-53

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics. 115.182089. https://doi.org/10.1534/genetics.115.182089

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2014. Manual for BLUPF90 family of programs. in Athens: University of Georgia.

Olson, K., P. VanRaden, and M. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci 95(9):5378-5383. https://doi.org/10.3168/jds.2011-5006

Pocrnic, I., D. A. Lourenco, Y. Masuda, A. Legarra, and I. Misztal. 2016a. The dimensionality of genomic information and its effect on genomic prediction. Genetics 203(1):573-581.

Pocrnic, I., D. A. L. Lourenco, Y. Masuda, and I. Misztal. 2016b. Dimensionality of genomic information and performance of the Algorithm for Proven and Young for different livestock species. Genet Sel Evol 48(1):82. https://doi.org/10.1186/s12711-016-0261-6

Pocrnic, I., D. A. L. Lourenco, Y. Masuda, and I. Misztal. 2019a. Accuracy of genomic BLUP when considering a genomic relationship matrix based on the number of the largest eigenvalues: a simulation study. Genet Sel Evol 51(1):75. https://doi.org/10.1186/s12711-019-0516-0

Pocrnic, I., D. A. Lourenco, C.-Y. Chen, W. O. Herring, and I. Misztal. 2019b. Crossbred evaluations using single-step genomic BLUP and algorithm for proven and young with different sources of data. J Anim Sci. 97(4):1513-1522. https://doi.org/10.1093/jas/skz042

Pryce, J. E., B. Gredler, S. Bolormaa, P. J. Bowman, C. Egger-Danner, C. Fuerst, R. Emmerling, J. Sölkner, M. E. Goddard, and B. J. Hayes. 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. J Dairy Sci 94(5):2625-2630. https://doi.org/10.3168/jds.2010-3719

Raymond, B., A. C. Bouwman, C. Schrooten, J. Houwing-Duistermaat, and R. F. Veerkamp. 2018. Utility of whole-genome sequence data for across-breed genomic prediction. Genet Sel Evol 50(1):27. https://doi.org/10.1186/s12711-018-0396-8

Raymond, B., L. Yengo, R. Costilla, C. Schrooten, A. C. Bouwman, B. J. Hayes, R. F. Veerkamp, and P. M. Visscher. 2020. Using prior information from humans to prioritize genes and gene-associated variants for complex traits in livestock. PLOS Genetics 16(9):e1008780. 10.1371/journal.pgen.1008780

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2011. FImpute: An efficient imputation algorithm for dairy cattle populations. J. Dairy Sci. 94(E-Suppl. 1):421. (Abstr.)

Sevillano, C. A., J. Vandenplas, J. W. M. Bastiaansen, and M. P. L. Calus. 2016. Empirical determination of breed-of-origin of alleles in three-breed cross pigs. Genet Sel Evol 48(1):55. doi: 10.1186/s12711-016-0234-9

Spelman, R., C. Ford, P. McElhinney, G. Gregory, and R. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. Journal of Dairy Science 85(12):3514-3517. https://doi.org/10.3168/jds.S0022-0302(02)74440-8

Steyn, Y., D. A. L. Lourenco, and I. Misztal. 2019. Genomic predictions in purebreds with a multibreed genomic relationship matrix. J Anim Sci 97(11):4418–4427. https://doi.org/10.1093/jas/skz296

Strandén, I. and E. A. Mäntysaari. 2013. Use of random regression model as an alternative for multibreed relationship matrix. Journal of Animal Breeding and Genetics 130(1):4-9. https://doi.org/10.1111/jbg.12014

Thaller, G., A. Winter, R. Fries, W. Krämer, B. Kaupe, and G. Erhardt. 2003. Effects of DGAT1 variants on milk production traits in German cattle breeds. Journal of Animal Science 81(8):1911-1918. https://doi.org/10.2527/2003.8181911x

Toosi, A., R. Fernando, J. Dekkers, and R. Quaas. 2010. Genomic selection in admixed and crossbred populations. Journal of Animal Science 88(1):32. https://doi.org/10.2527/jas.2009-1975

Vandenplas, J., M. P. L. Calus, C. A. Sevillano, J. J. Windig, and J. W. M. Bastiaansen. 2016. Assigning breed origin to alleles in crossbred animals. Genet Sel Evol 48(1):61. https://doi.org/10.1186/s12711-016-0240-y

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414-4423. https://doi.org/10.3168/jds.2007-0980

VanRaden, P. M., M. E. Tooker, T. C. S. Chud, H. D. Norman, J. H. Megonigal, I. W. Haagen, and G. R. Wiggans. 2020. Genomic predictions for crossbred dairy cattle. J Dairy Sci 103(2):1620-1631. https://doi.org/10.3168/jds.2019-16634

Vukasinovic, N., N. Bacciu, C. A. Przybyla, P. Boddhireddy, and S. K. DeNise. 2017. Development of genetic and genomic evaluation for wellness traits in US Holstein cows. J Dairy Sci 100(1):428-438. https://doi.org/10.3168/jds.2016-11520

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics Research 94(2):73-83. https://doi.org/10.1017/S0016672312000274

Wei, M. and J. H. J. van der Werf. 1994. Maximizing genetic response in crossbreds using both purebred and crossbred information. Animal Science 59(3):401-413. https://doi.org/10.1017/S0003356100007923

Wellmann, R. 2019. Optimum contribution selection for animal breeding and conservation: the R package optiSel. BMC Bioinformatics 20: 25. https://doi.org/10.1186/s12859-018-2450-5

Wellmann, R. 2020. Package 'optiSel' manual. https://cran.r-project.org/web/packages/optiSel/optiSel.pdf

Wiggans, G. R. 1997: Genetic evaluation systems in the United States. Accessed October 26,2020. https://aipl.arsusda.gov/publish/other/1997/conf_isap97_19.html.

Wiggans, G.R., VanRaden, P.M., Nicolazzi, E.L., Tooker, M.E., Megonigal, Jr., and Walton, L.M. 2019. Interbull Bull. 55:46-49

TABLES

Table 5.1: Eigenvalues explaining 90%, 95%, 98% and 99% of variation in the genomic relationship matrix (**G**) when all genotyped animals are considered.

	Animals	90%	95%	98%	99%
Holstein	89 558	4 637	8 101	14 068	19 046
Jersey	40 769	3 325	5 775	9 841	13 216
Cross	22 373	4 939	6 698	8 676	9 887
Jersey & Holstein	130 327	5 661	9 666	16 250	21 564
Mix[1]	61 275	7 396	11 857	18 017	22 483

[1] The mixed population contains around equal numbers of Holstein, Jersey, and crossbred animals

Table 5.2: Predictive ability (Pearson correlation between IP and adjusted phenotype) when using marker effects based on a breed, or group, to predict the indirect genomic value of itself, and that of others. Inflation for predictions on crossbreed animals are included.

Breed used	Breed predicted			
	Jersey	Holstein	Cross	
		Predictive ability		Inflation
Jersey	0.48[1]	0.13	0.50	0.65
Holstein	0.09	0.45[1]	0.47	0.55
Cross	0.24	0.26	0.50[1]	1.17
Jersey & Holstein	0.45[1]	0.44[1]	0.50	0.78
Mix	0.46[1]	0.40[1]	0.46[1]	1.00
Proportions	-	-	0.32[1]	0.85

[1]Indicates whether the breed predicted was also represented in the training population

Table 5.3: The correlations between indirect predictions of crossbred validation animals (IP_{Cross}) obtained using different SNP effects and GEBV of crossbred animals estimated with different datasets.

IP_{Cross}	$GEBV_{Partial}$[1]	$GEBV_{Full\_CROSS}$[2]	$GEBV_{Full\_MIX}$[3]
Population used for SNP effects			
Crossbreds	0.97	0.86	0.73
Jersey	0.75	0.72	0.65
Holstein	0.68	0.66	0.62
Jersey and Holstein	0.65	0.67	0.75
MIX[4]	0.76	0.72	0.84
Other method			
Breed proportions	0.29	0.37	0.56

[1] The $GEBV_{Partial}$ were obtained from an evaluation with crossbred genotypes and only phenotypes of the training population.

[2] The $GEBV_{Full\_CROSS}$ were obtained from an evaluation with crossbred genotypes and all available phenotypes.

[3] The $GEBV_{Full\_MIX}$ were obtained from an evaluation with all crossbred genotypes, ~20K Holstein genotypes, ~20K Jersey genotypes and all available phenotypes.

[4] The mixed population contains approximately equal numbers of Holstein, Jersey and crossbred animals

FIGURES

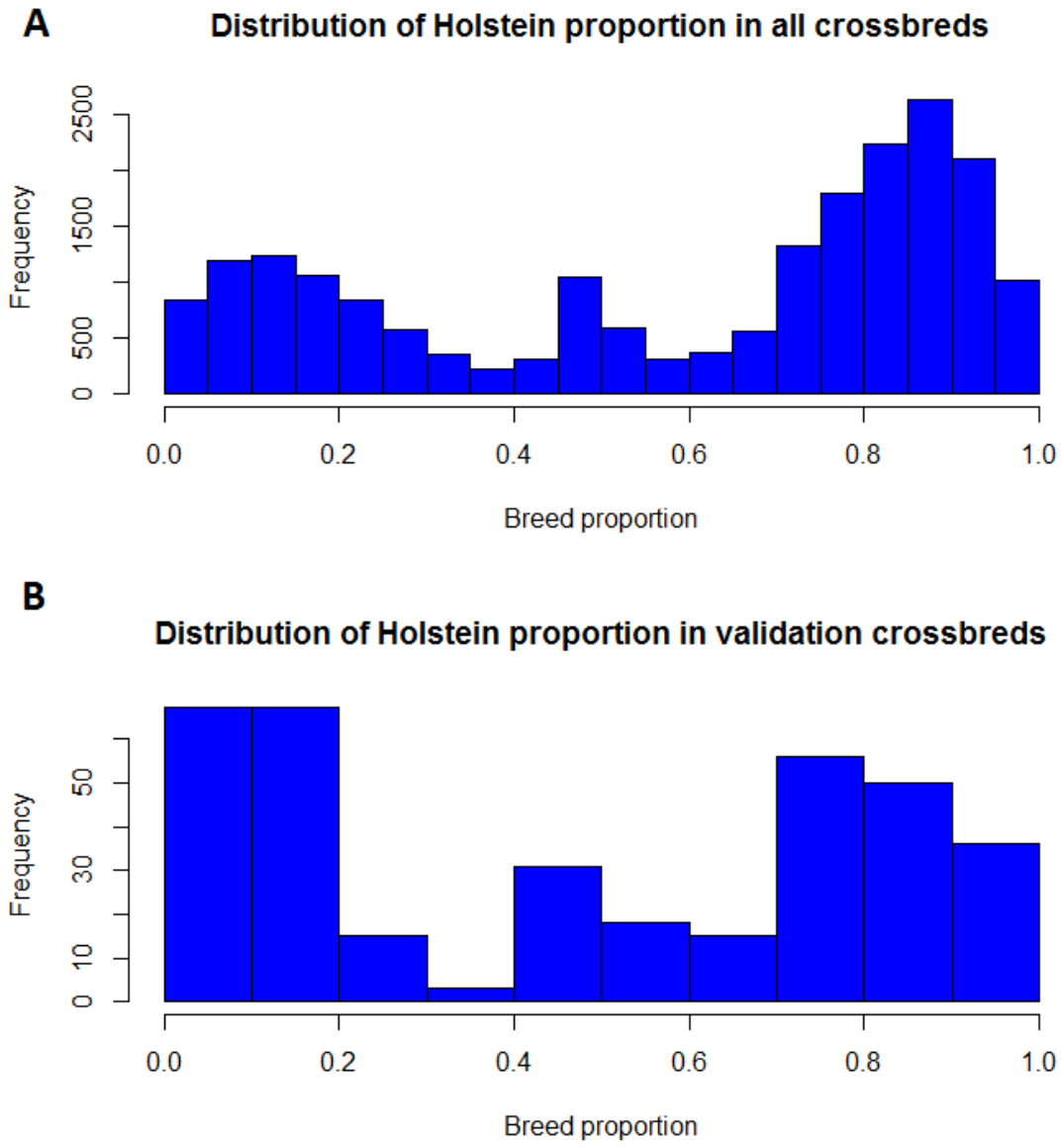


Figure 5.1**.** Distribution of the proportion of the crossbred genotypes that are assigned as Holstein. The first plot applies to all crossbred animals, while the second applies to only validation crossbred animals

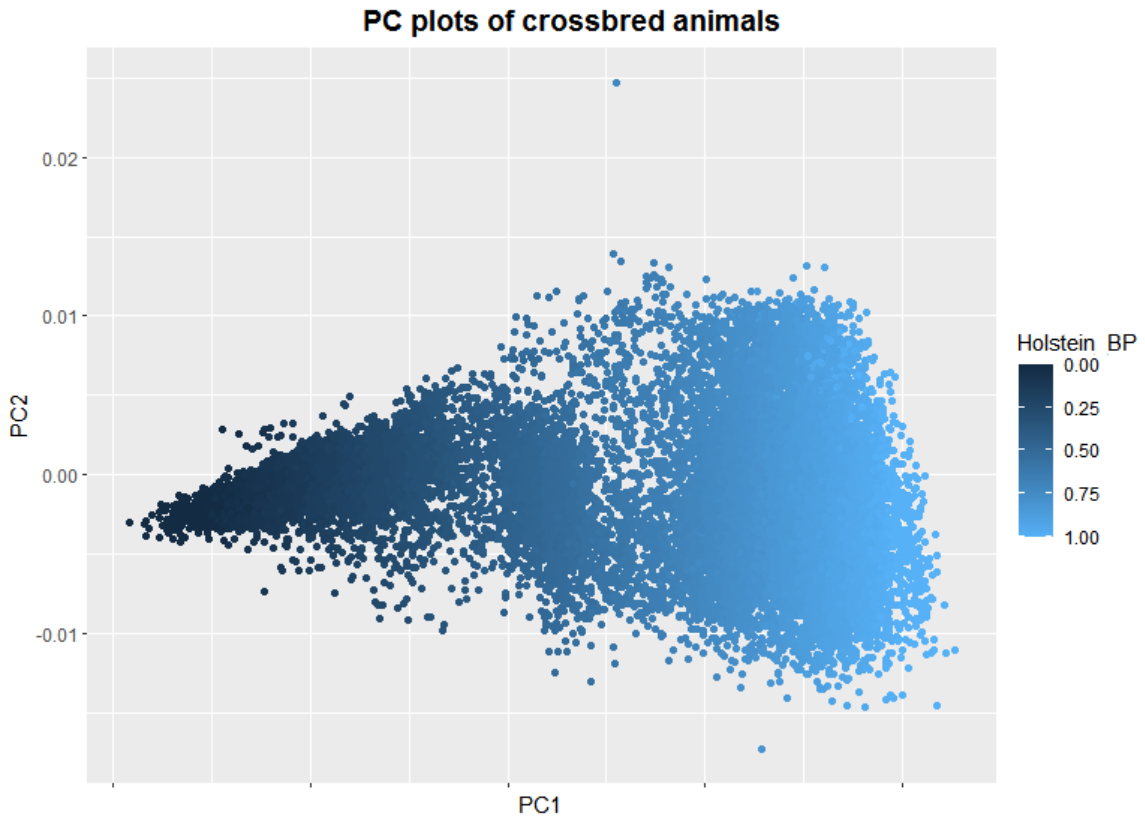


Figure 5.2: Principal component (PC) plot for all crossbred animals. Color intensity indicates the Holstein breed proportion (BP) of each crossbred animal. Animals with a Holstein BP lower than 0.50 have higher Jersey BP

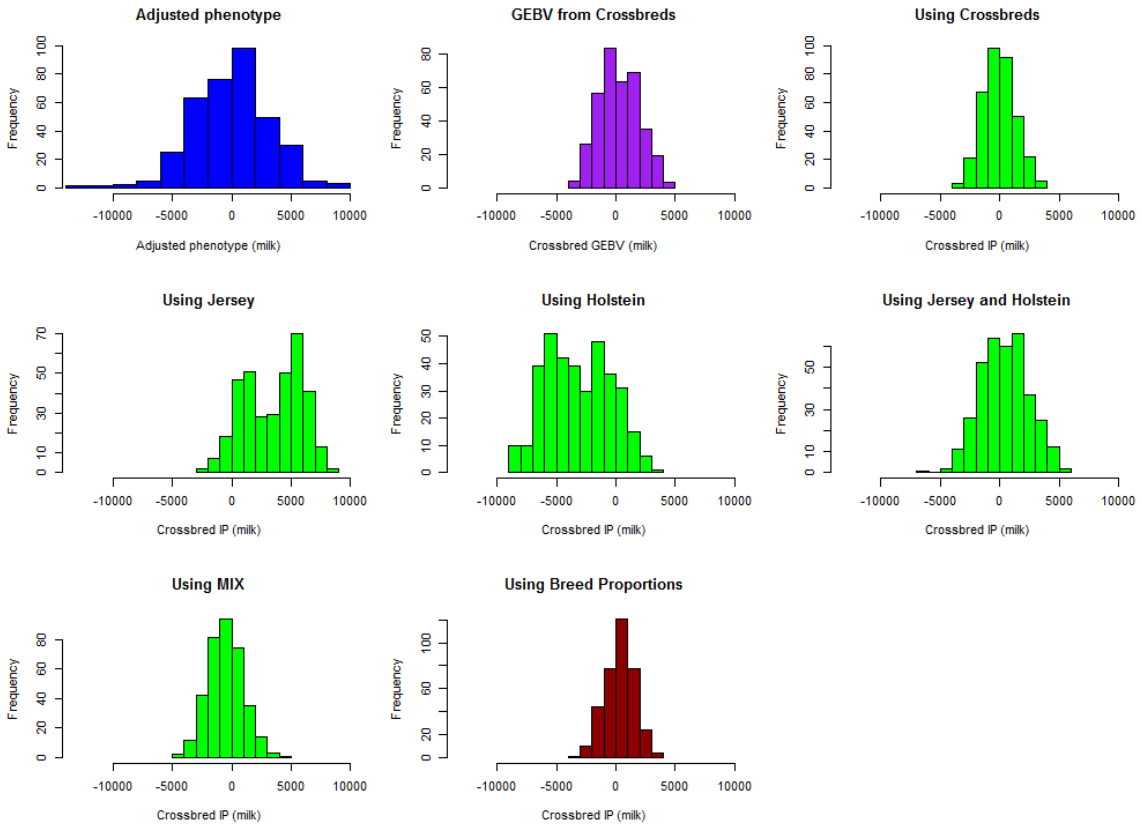


Figure 5.3: The distributions for the resulting estimations for the crossbred validation animals. Distributions include adjusted phenotype, Genomic Breeding Value (GEBV) obtained when genotypes of only crossbred animals are included and phenotypes of the validation populations are excluded, and Indirect Predictions (IP) obtained from SNP effects estimated when excluding phenotypes of validation animals and including genotypes of specific groups. The different reference groups to estimate SNP effects for the calculation of IP were composed of 1) only crossbred animals, 2) MIX as described before, 3) only Jersey genotypes, 4) only Holstein genotypes, or 5) both Jersey and Holstein genotypes. Another IP was obtained by using genomic breed proportions as weights to sum the IP obtained when using Jersey SNP effects, and when using Holstein SNP effects.

CHAPTER 6

EFFECT OF POPULATION STRATIFICATION, GENOME CHANGES, AND GENETIC REDUNDANCY UPON DIVERSITY WITHIN THE U.S. HOLSTEIN BREED

Y. Steyn, T. Lawlor, Y. Masuda, S. Tsurutu, D.A.L. Lourenco, I. Misztal. To be submitted to the Journal of Dairy Science

ABSTRACT

Maintaining genetic variation in a population is important for long-term genetic gain. The existence of sub-populations within a breed helps maintain genetic variation and diversity. The 20,990 selected candidates, representing the breeding animals in the year 2014, were identified as the sires of animals born after 2010 with at least 25 progeny, and females measured for type traits within the last 2 years of data. K-means with 5 clusters (C1,C2,C3,C4, and C5) was applied to the genomic relationship matrix based on 58,990 SNP markers to stratify the selected candidates into subpopulations. The general higher inbreeding resulting from within-cluster mating compared to across-cluster mating, suggests the successful stratification into genetically different groups. The largest cluster (C4) contained animals that were diverse enough to allow low inbreeding both within and across cluster. The average F_{st} was 0.03, indicating that allele differences across the sub-populations are not due to drift alone. Starting with the selected candidates within each cluster, a family unit was identified by tracing back through the pedigree, identifying the genotyped ancestors and assigning them to a generation. Each of the five families (F1, F2, F3, F4 and F5) were traced back for 10 generations, allowing for changes in frequency of individual SNPs over time to be observed, which we call allele frequency (AF) change. Alternative procedures were used to identify SNPs changing in a parallel or non-parallel way across families. For example, literature search of previously identified genes known to be changing in frequency, markers with the greatest association with time, markers that have changed most in the whole populations, and markers that have changed differently across families. The genomic trajectory taken, by each family, involves selective sweeps, polygenic changes, hitchhiking and epistasis. The Replicate Frequency Spectrum (RFS) was used to measure the similarity of change across families, and showed that populations have changed

differently, supporting the presence of genetic redundancy. The proportion of markers that reversed direction in AF change varied from 0.00 to 0.02 if the magnitude of change was greater than 0.02 per generation, or from 0.14 to 0.24 if the rate of change was greater than 0.005 within each family. Genome-wide association studies (GWAS) for stature was done for all clusters combined (ALL) and within each cluster. Utilizing the different SNP effects for indirect genomic predictions (IGP) resulted in reranking within family, indicating epistasis. Correlations between IGP for stature based on SNP effects from a specific cluster, or all combined, show that the exclusion of a cluster from the training population may reduce the accuracy of genomic prediction. Further research is required to determine how this knowledge can be applied to maintain diversity and optimize selection decisions in the future.

Keywords: k-means, clustering, polygenic adaptation, hitchhiking, selection sweeps, epistasis

INTRODUCTION

Understanding the population structure of a breed is critical in revealing its genetic diversity and the changes occurring within its genome over time. Stratification allows for the identification of SNP that change in frequency in a uniform way across all subpopulations, as well as those that change in a unique way within one, or more, subpopulations. Without stratification, the pooling of all animals together masks these family specific changes. In recent years, an abundance of genomic information on different species undergoing adaptive response to environmental change or selection for different agricultural goals has become available. This has led to new ideas on how to evaluate adaptation and understand the genetic architecture of traits.

The additive genetic model does an excellent job of allowing breeders to change the phenoypic average of a population towards a desired goal. However, it does not expose the

genetic complexity and diversity that help maintain the genetic variation that allows for both current and future genetic change. The additive breeding value itself is the sum of all markers affecting the trait, while the total breeding value also include interactions and other non-additive effects. The different combinations may be infinite, thus, populations may show non-parallel changes in gene frequencies. Heterogenous change in allele frequencies among sub-populations is an indication of genetic redundancy (Barghi et al., 2019).

Genetic redundancy allows multiple genetic pathways to achieve the same phenotype, essentially providing more beneficial variants than needed (Goldstein and Holsinger, 1992, Nowak et al., 1997). Therefore, populations that have been separated and selected for the same trait, may have undergone different changes in allele frequencies (AF) even when phenotypes are similar. This is due to different sets of loci responding differently to the same selection pressure (Barghi et al., 2019). Genetic redundancy arise through multiple factors. One or more genes may serve the same function and therefore, the absence of expression in one may not affect the phenotype (Pickett and Meeks-Wagner, 1995). Additionally, highly polygenic traits are influenced by many genes that each have a relatively small contribution to the phenotype, hence the infintisimal model (Bulmer, 1971, Turelli, 2017). Each allele would then be expected to slowly change by subtle shifts, instead of selection sweeps of few genes (Höllinger et al., 2019). Additionally, many genes not directly involved in obvious biological pathways of trait expression may collectively explain more variation in traits than core genes, reflecting the omnigenic nature of traits (Boyle et al., 2017). It has been shown that up to 70% of trait variance can be attributed to trans-chromosomal effects through peripheral genes that impact the expression of core genes (Liu et al., 2019). These trans-chromosomal effects are partly due to pleiotropy (where genes are involved in the expression of more than one trait), and epistasis

(where the expression of one gene influences the expression of another). In the US Holstein population, the percentage of epistatic effects that were inter-chromosomal varied from 1.9% to 84.2%, depending on trait (Prakapenka et al., 2021).

The global dairy industry is dominated by a few breeds, in particular the Holstein. Concern has been expressed that artificial insemination has resulted in the widespread use of semen from a handful of bulls (Yue et al., 2015). Which can lead to higher inbreeding and genetic similarities worldwide. Although this genetic connectedness can be advantageous for genetic evaluations and the similarity of animals provides a more uniform and predictable product, it may be problematic for long-term genetic improvement and adaptability. While inbreeding can increase the frequency of favorable genes for traits under selection, it leads to the decrease in performance of other traits, in particular fertility and overall health (Pryce et al., 2014), as well as the loss of rare alleles that could be of importantance in the future. Maintaing genetic diversity is considered crucial for a population to adapt to changing environments, such as climate change and consumer preferences.

The objectives of this study were to investigate the amount of stratification occurring within the U.S. Holstein population, and the role that genetic redundancy contributes to the differences in these sub-populations.

MATERIALS AND METHODS

Genotypes were available for the US Holstein population up to 2014. The number of animals in the pedigree was 9,817,252, which contained 330,837 sires and 5,471,039 dams. The most progeny for a sire was 58,266 (USAM000001626813 - Mars). The average number of progeny per sire was 29. The data file contained only type traits, and totaled 10,067,745 records.

After removal of unmapped and sex chromosomes, 58,990 SNP markers remained. Genotypes were available for 569,404 animals.

The genomic relationship matrix (G) was obtained using the formula $\mathbf{G} = \frac{\mathbf{MM'}}{2\sum p_i(1-p_i)}$ where M is a matrix of SNP content centered by twice the current allele frequencies, and p_i is the current allele frequency for SNP i (VanRaden, 2008).

Clustering the selected candidates

Potential selection candidates in 2014 were identified as sires of animals born after 2010 with at least 25 progeny (3,902 animals), and cows that were recorded for type traits in 2013 or 2014 (16,197 animals). The animals represented 14 countries, including Australia, Austria, Canada, Czech Republic, Germany, Denmark, Finland, France, Great Britain, Hungary, Italy, Netherlands, Sweden, and the US. K-means clustering with a built-in R package (Hartigan and Wong, 1979) and 50 iterations was based on the genomic relationship matrix (**G**) to identify five clusters of animals (C1, C2, C3, C4, and C5) as giving shape to the genetic diversity of the Holstein. The number of animals in each cluster is presented in table 6.1. A principal component analysis plot is presented in Fig 6.1. Hypothetical matings were performed within and across clusters with the INBUPGf90 software package within the BLUPF90 software suite (Misztal et al., 2014b). Expected inbreeding of offspring was calculated for every possible mating between a specific group of sires and specific group of dams. Solutions were based on the complete pedigree information of the Holstein population assuming non-zero inbreeding for unknown parents (Aguilar and Misztal, 2008). The average expected inbreeding of animals when mating within-cluster, and of all animals in across-cluster scenarios are presented in table 6.2.

Families

Five families were formed, not based on a stringent pedigree structure, but by the transmission of alleles identical by descent flowing down through the generations amongst a subgroup of highly related animals. To track the gene flow over time, the selection candidates were labeled as generation 10 (G10). Within each cluster of G10, the pedigree was traced back for 10 generations, with parents, grandparents, and greatgrandparents of G10 labeled as G9, G8, and G7 (up to G0) to form five families (F1, F2, F3, F4, F5). Figure 6.2 visually explains how families were constructed. Grouping of animals into different families is highly dependent upon the most recent ancestors. Ancestors varying in their occurrence in different families, alter the gene flow, resulting in different allele frequencies across families. The number of genotyped animals in each generation of each family are presented in table 6.1. Due to the overlapping of generations and abundance of admixture amongst families, animals may appear in different families and in several different generation. Table 6.3 shows the number of times that several prominent bulls appear as a sire of a G10 animal, or genotyped G9 animal, in different families, while table 6.4 shows the number of times highly influential bulls appear in each family. Table 6.5 shows the number of animals unique to each family, or common to all, within each generation. As shown in Fig 6.3, 82% of animals appear in more than one family in G0, while 54% of the animals are common to all families. As generations pass, the proportion of ancestors shared becomes smaller. By G6, the number of animals in common for all families is greater than the number of animals unique to any family. The total number of unique animals for each family is 4,355 (F1), 3,555 (F2), 3,574 (F3), 7,107 (F4), and 4,879 (F5).

Changes in allele frequencies

Changes in allele frequencies (AF) for the whole breed were calculated from the differences in AF between each generation for all families combined. Specific within family AF changes were determined solely from within family generational data. Six different procedures were used to identify SNPs changing in a parallel or non-parallel way across families. These procedures included investing AF changes for specific genes (GENES) known to be changing in our specific population, GWAS on the whole population with birth year considered a trait (TIME), largest regression coefficient (COEF) when regressing AF on generation number, the variance and range in the absolute difference in AF between the youngest and oldest generation across the five families (VAR and RANGE), and those SNPs identified by the Lewontin and Krakauer's test (LK).

Selected genes (GENES)

The *DGAT1* gene on chromosome 14 was chosen to observe over time due to its known significant genetic effect on milk production (Thaller et al., 2003, Barbosa da Silva et al., 2010). Additionally, *AVEN* (chromosome 10), *SPATA6* (chromosome 3), *ERBB4* (Chromosome 2), *SKIV2L* (chromosome 23), and *USP13* (chromosome 1) were chosen based on results from Ma et al. (2019), which showed that these genes are among those that have changed the most in the US Holstein population. Their findings were based on a comparison between modern animals and the animals from the University of Minnesota Holstein control line. This line has been unselected since 1964. The chosen genes were not on the same chromosome, nor on the sex chromosomes.

SNP identified with GWAS for birth year (TIME)

To identify alleles with unknown functions that are most associated with time, we performed a single-step genome-wide association study (GWAS) using blupf90 software (Misztal et al., 2014b) with birth year as trait (Rowan et al., 2020). The model was:

$$\mathbf{y} = \mathbf{xb} + \mathbf{Zu} + \mathbf{e}$$

where **y** is the vector containing birth year, **x** is a vector of ones to assign the mean (**b)** to all records, **Z** is an incidence matrix assigning the measurements to the random animal effect, **u** is a vector of solutions for animal effect, and **e** is the residual. Single-step genomic BLUP (ssGBLUP) was performed with BLUPf90IOMD2 to obtain genomic breeding values (GEBV) for all animals. This required a relationship matrix that combines both genotyped and ungenotyped animals (H). The inverse of the **H** matrix is constructed as (Aguilar et al., 2010):

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix}$$

Where **A** is the traditional pedigree relationship matrix among all animals, and $\mathbf{A_{22}}$ is the pedigree-based relationship matrix between only genotyped animals. The approximated inverse of **G** was constructed using the algorithm for proven and young (APY) with 15,000 randomly selected animals as core (Misztal et al., 2014a). To reduce bias due to the different genetic level of genotyped and non-genotyped animals, **G** was tuned to be compatible with $\mathbf{A}_{22}$ using the method described by Chen et al. (2011). To avoid singularity of the relationship matrix, 5% of $\mathbf{A}_{22}$ was combined with 95% of **G**. A weighted G was constructed using non-linear methods (VanRaden, 2008) and evaluations followed 3 iterations. The SNP effects were estimated based on these GEBV using the POSTGSF90 software package with the formula (VanRaden, 2008, Wang et al., 2012):

$$\hat{\mathbf{a}} = \lambda \mathbf{D}\mathbf{M}'\mathbf{G}^{-1}(\text{GEBV})$$

where $\hat{\mathbf{a}}$ is a vector of estimated SNP effects, λ is the ratio of SNP to additive genetic variance, **D** is a diagonal matrix of weights for SNP, and **M** was defined before. The 5 SNP that showed the greatest effects were selected.

Change in AF over generations (COEF)

All families were combined to determine the AF change over time for the overall population. The AF was calculated for each generation and regressed over generations to obtain a slope for each SNP marker. The formula was $y_{ij} = \beta_{0i} + \beta_{1i}x_j + e_{ij}$, where y_{ij} is the vector of AF for SNP i in generation j, β_{0i} is the mean AF for SNP i, β_{1i} is the regression coefficient over generations for SNP i, x_j is a vector of generation j (11 levels), and e_{ij} is the error associated with SNP i in generation j. The 20 SNP with the greatest change over time were identified. Of these SNP, 5 that were further apart from each other were identified. The ranking based on β_1 for these SNP were 1, 5, 6, 16, and 19.

Greatest variance of change over generations within families (VAR)

The absolute difference between G10 and G0 was calculated within each family, providing an estimate for change in AF over time. To identify SNP markers that have changed differently across families, the variance of these differences in the five families was calculated. The 20 SNP with the highest variance were identified and 5 were selected to avoid having markers close to each other. The ranking based on the variance were 1, 2, 3, 5, and 19.

Greatest range of change over generations within families (RANGE)

The range between the AF change of the family with the least change, and that of the family with the greatest change was calculated. This is an additional measure to identify SNP markers that have changed differently across families. The top 20 SNP were identified and five

were chosen to avoid markers close to each other and eliminate those that were also among the top 20 based on variance. The ranking based on range were 2, 5, 7, 9, and 13.

Lewontin and Krakauer test (LK)

Genetic drift and migration are also contributors to changes in gene frequencies over time (Falconer and Mackay, 1996). The Lewontin and Krakauer test (T_{LK}) aims to identify markers that changed due to selection, not drift or migration (Lewontin and Krakauer, 1973). The test makes use of a measure of genetic differences among sub-populations, namely the F_{st} test as defined by Wright (1943). The F_{st} test is effectively the fraction of total genetic variance due to differences among sub-populations. Only SNP markers with a MAF > 0.05 based on all families combined were used for this test. The AF of these SNP within G10 of each family were calculated. Let $p = (p_1,\ldots,p_5)$ be a vector of the AF of an individual SNP in each of the 5 families. The F_{st} for each SNP was calculated as:

$$F_{st} = \frac{s_p^2}{\bar{p}(1-\bar{p})}$$

where $\bar{p}$ is the mean of vector p and s_p^2 is the sampling variance of each SNP across families. This is used in T_{LK} formula (Lewontin and Krakauer, 1973):

$$T_{LK} = \frac{n-1}{\bar{F}_{st}} F_{st}$$

where n is the number of families and $\bar{F}_{st}$ is the mean F_{st} of all markers. The T_{LK} follows a χ^2 distribution with n-1degrees of freedom. P-values were obtained from this distribution based on the T_{LK} of each marker. The Bonferroni adjustment, and the false discovery rate (FDR) were used as measures of significance. The five SNP markers with the highest -log(p-value) were selected.

Epistasis

The interaction between loci (epistasis) is a non-additive genetic contributor to the genetic architecture of quantitative traits (Mackay, 2014). If epistatic effects are present, it is expected that SNP effects may be significant in sub-populations, but not in the whole population. Therefore, a modified mechanical heuristic approach was taken to compare effects across populations and with the whole population (Reverter et al., 2018). Six GWAS were performed with stature as trait – one with all female animals of G10 combined (ALL), and one with all females within each cluster of G10. The model was as described by Tsuruta et al. (2021). Unlike the GWAS with birth year as trait, GBLUP was performed instead of ssGBLUP. Thus, only genotyped animals were included in the analysis and the genomic relationship matrix (**G**) was not blended with the traditional pedigree relationship matrix. To make **G** invertible, 0.05 was added to the diagonal of **G**. Significance was determined based on the p-value with Bonferroni correction, and the false discovery rate (FDR). Since stature is only measured on females, males were removed and more daughters with measurements were added to increase the representation of genotyped males. These daughters were born after 2005, had phenotypes, and were unique to each cluster. The number of females in C1, C2, C3, C4, and C5 before adding these daughters were 2 866, 2 627, 2 572, 4 908, and 3224, respectively. The number of unique daughters added to each cluster were 678, 1 222, 1 287, 5 462, and 658, respectively. Only genotypes of female animals were used as training population. The genotypes of male animals were used as validation populations to compare indirect genetic predictions (IGP) obtained using SNP effects based on each cluster or ALL. The number of training (and validation) animals were 3 544 (711), 3 849 (446), 3 859 (730), 10 370 (1023), 3 882 (992), and 25 504 (3902) for C1, C2, C3, C4, C5, and ALL, respectively. The IGP for validation animals in each cluster when using different SNP

effects were compared using a Pearson correlation, with IGP based on within-cluster SNP effects as the benchmark for comparison (Table 6.6). This is a measure of correlations between two popolations, as proposed by Duenk et al. (2020). The method provides two different correlations since one refers to the IGP of population 1 based on the SNP effects of population 2, and the other to the IGP of population 2 based on the SNP effects of population 1. Animals from each cluster were ranked according to their IGP based on different SNP effects. Table 6.7 shows the ranking of 10 bulls (two bulls from each cluster) when using different SNP effects.

Genetic redundancy

The replicate frequency spectrum (RFS) can be used as a measure for heterogeneity and genetic redundancy across populations (Barghi et al., 2019). This identifies the percentage of the top 100 SNP markers (based on different criteria) that change similarly in different clusters. Criteria included the 100 markers with the highest F_{st}, VAR, and RANGE. Table 6.8 shows how many of these markers had AF changes greater than 0.10 or 0.30 (from the earliest generation to most recent) in each family. Additionally, the regression coefficient when regressing AF over generations within each family was used to identify the 100 markers that have changed the most in each family. Table 6.9 shows how many of these markers also changed in other families.

The proportion of markers that have changed direction within each family (initially increased but later decreased, or initially decreased but later increased) was calculated. Change was measured by comparing the regression coefficient when regressing AF over G0 to G5, and the regression coefficient when regressing AF over G5 to G10. Instead of using zero as cut-off point for directional change, we identified those that changed in one direction (increased or decreased) at a rate of at least 0.02 per generation in the first phase, and in the opposite direction at a similar rate in the second phase. Less strict cut-offs of 0.01 or 0.005 per generation were also

used to detect more subtle directional changes. The number and proportion of markers that have changed direction are presented in Table 6.10.

RESULTS AND DISCUSSION

Studies have shown that cross-validation accuracy to determine the usefulness of genomic prediction was lowest when using k-means as clustering (Saatchi et al., 2012, Boddhireddy et al., 2014, Baller et al., 2019). Since the accuracy of genomic predictions depends on the relationships between the training and target populations (Habier et al., 2010, Clark et al., 2012, Pszczola et al., 2012), this suggests that k-means clustering is successful at separating groups that are more related to each other, but less related to other clusters. K-means clustering identified five clusters as giving shape to the genetic diversity of the Holstein breed. The PCA plot in Fig 6.1 reveal different, but overlapping subgroups within the population.

Cluster differences

High within-cluster inbreeding and low across-cluster inbreeding indicates that animals have indeed been separated into clusters with animals that are genetically more similar to each other, and more different compared to animals in other clusters. A noticeably small expected inbreeding occurs for all mating scenarios with animals in C4, whether within- or across-cluster. This shows that although C4 is more different than other clusters, it still contains enough variation within itself to allow low inbreeding (table 6.2).

An additional measure of genetic differences, is the fixation index (F_{st}), which is an indication of genetic changes not due to drift alone. The average F_{st} value for markers across the five clusters was 0.03, which is lower than the 0.07 found in a French study that compared three different dairy breeds – Holstein, Montbéliarde, and Normande (Flori et al., 2009) – but higher than the expected value (0.00) if populations were uniform. The selected candidates do not

represent one large panmictic population. They come from a complex mixture of family subgroups with differences in AF.

Genome changes

There are two main fields of thought on how alleles change under selection. Selection sweeps, in which favorable alleles quickly move towards fixation, favors traits that are influenced by few genes with large effects and less pleiotropy. Subtle frequency shifts occur more often in highly polygenic traits following the infintisimal model (Höllinger et al., 2019). Three phases for AF change of a trait under selection have been identified using simulated data for a trait with an intermediate optimum. The first is the directional selection phase where AF of multiple loci move towards the trait optimum. The second is a plateauing phase, where the frequency changes slow down and remain more or less at the same level. The third phase is where the alleles either move to fixation (or are lost) and the trait mean stabilizes. These separate phases occur over hundreds of generations. Factors that affect the duration of these phases include the effective population size, number of loci, selection intensity, distance to the new fitness optimum, the distribution of effect sizes, and the starting frequencies of the alleles (Franssen et al., 2017).

"Evolve and re-sequence" studies observe changes in AF over generations. An example is the study by Barghi et al. (2019), where they observed a natural outcross population as it adapted to a higher temperature. This led to different subpopulations with their own genetic solutions (redundancy) to converge to the same phenotypic goal. Genetic redundancy in dairy cows is expected as breeders define an overall breeding objective or fitness measure and then select breeding animals from the available candidates representing several different subpopulations.

To fairly compare the genomic changes across families from a similar starting point, it is important for the earliest generation to contain similar animals, while the more recent generations are different. The large proportion of shared ancestors in the earlier generations allows for greater similarity. This can be seen in tables 6.3 and 6.4. Prominent bulls, such as Planet, Goldwyn, Shottle, BW Marshall, and Oman are not exclusive to a specific family as they appear in the different families and generations. However, differences in their proportional influence upon a family allows them to shift the frequency of different alleles over time. Heterogenous changes in AF across families can come from similar ancestors contributing a varying percentage of descendants. Tracing back to some of the historically most prominent ancestors, such as Elevation, Chief, and Blackstar, it can be seen that they are more evenly represented in the genotyped animals of each family. This provides a homogenous early genetic base with similar initial AF across all families.

Our analysis concentrated on temporal changes in AF over several generations amongst five different families. Heterogeneity in AF changes across families indicates that different sets of SNPs are changing over time. Having several distinct families with diversity in their genome helps to maintain genetic variation over time. Comparing AF changes in our study allowed us to identify signatures of selection that are family specific. Fixation of alleles was infrequent across the whole population (3 alleles), however, it was greater and varied within families. The number of alleles (from the 58,990 SNP markers) that became fixed within each family were 38, 22, 22, 59, and 40 for F1, F2, F3, F4, and F5, respectively. None could be described as a selective sweep as all had initial frequencies near fixation. The genome trajectory taken by each family, involves selective sweeps, polygenic shifts, hitchhiking, epistasis, and genetic redundancy. These can be observed in plots of specific alleles selected based on these different criteria:

GENES

The AF change for *DGAT1* for all families are presented in Fig 6.4. It starts at a high frequency and remains relatively unchanged, or show small increases, for four of the families. This gene is known to have a significant effect on fat yield (Spelman et al., 2002, Thaller et al., 2003, Jiang et al., 2019). The high starting point of AF in G0 is not surprising given that milk production traits have undergone selection decades before the birth of our G0. If this gene was acting alone, the expectation would have been that it would be fixed in the population. However, due to its antagonistic effects with milk yield, AF have remained similar over time (Jiang et al., 2018). The change for F2 is distinctly different, clearly decreasing. Surrounding SNP markers also show different behavior compared to other families. This may reflect a change in breeding objectives in the dairy industry. While fat yield was widely considered as unfavorable for human health, the attitude towards fat in diets changed within the last few decades. This sub-group may capture animals that were selected more for solids than milk yield.

The *AVEN* gene is associated with male fertility (Laurentino et al., 2011) and shows similar increasing trends in AF changes, with surrounding SNP markers showing different behavior in F2 (not shown). The *ERBB4* gene is involved in embryonic lethality (Tidcombe et al., 2003). The F1 and F5 show sharper increases in AF compared to other families, but the overall trend may also be considered as similar. The surrounding SNP markers in F1 show signals of hitchhiking (Fig 6.5). It is possible that this is only observed in F1 because linkage disequilibrium (LD) was lost due to recombination in other families.

The *SKIV2L* gene is also involved in fertility (Ma et al., 2019). Here, more heterogeneity is observed as the AF in F3 start to decrease from G9 while others continue to increase. Differences can also be observed in the surrounding SNP markers (not shown). The *SPATA6*

gene is associated with sperm quality (Yuan et al., 2015) and also shows heterogenous changes. The AF increased up to G7 for all families, after which they plateau or slowly decrease in F2, F3 and F4, decrease more sharply in F5, and increase sharply in F1 (Fig 6.6). The surrounding SNP also behave differently in F1. The *USP13* gene is involved in the immune system (Zhang et al., 2013) and shows change in different directions across the families (Fig 6.7). This difference in direction of change may indicate epistatic effects that are different in the families.

TIME

Genes that are most associated with time are expected to be those that have been subjected to selection. There were no clear SNP markers that showed exceptionally strong associations with time. The AF of the SNP with the greatest, and fourth greatest effect on birth year are more consistent in direction across families (not shown). More non-parallel changes are observed for the three other SNP in the top 5. In F1 and F2, sharper decreases can be observed for the SNP second most associated with time. The AF in F2 and F4 appear to plateau after decreasing, and increase in F3 after an initial decrease. The surrounding SNP markers show similar behavior in F1, F2, and F3 compared to the other families (Fig 6.8). For the SNP third most associated with time, AF in F5 show a sharp change in direction, while AF slowly decrease in the other families. The surrounding SNP behave differently in F4 (Fig 6.9). The AF changes for the SNP fifth most associated with time, along with some surrounding SNP, show greater changes in F1 compared to the other families (Fig 6.10).

COEF, VAR, and RANGE

The regression coefficient when regressing AF over generations for the whole population (COEF) was used to identify SNP markers that have changed the most over time. Thus, this is not family specific. The change in AF is fairly consistent in direction and magnitude across all

five families for two of the top five selected SNP, such as in Fig 6.11 and 6.12. This rapid change in AF for all families may reflect a partial selection sweep for a gene that has undergone selection. Some surrounding SNP markers change at a similar rate, but in opposite directions, and are more pronounced in some cases. This can be a hitchhiking effect by nearby markers. It is possible that our selected marker is in fact the hitchhiker, instead of the gene under selection.

The variance of change was used as a measure to identify scenarios where at least one family show small changes while at least one other family shows a large change. For all five selected SNP, F1 and F5 follow similar trends, while F2 and F3 have trends similar to each other, but in a different direction and pattern compared to F1 and F5 (Fig 6.13 and Fig 6.14). This change in direction may indicate epistasis, where the fate of a given allele is highly contingent on the allelic makeup at other loci within the family. Surrounding SNP markers show stronger linkage in Fig 6.13, while those in Fig 6.14 show weak linkage.

The range between the family with the least change and family with the most change was also used to identify SNP that behave differently in the families. The top 3 selected SNP showed the greatest change in F1, as well as stronger responses in the surrounding SNP (Fig 6.15 shows the SNP with the greatest range). The fourth and fifth selected SNP markers show changes in different directions across families (such as Figures 6.16 and 6.17).

LK

Figure 6.18 includes the resulting Manhattan plot of the $-\log_{10}$(p-value) of each SNP marker. Peaks are observed and some approach significance, but none of the SNP markers met the criteria for statistically significant differences ($p<0.05$ with a Bonferroni adjustment for the number of markers, or FDR). The AF change for five SNP nearest to the significance threshold were investigated. All were markers that started with AF close to fixation. While the AF

remained stable for four of the families, one family (F1 for 4 of the SNP, and F2 for the other) showed a dramatic change in the recent generations (Fig 6.18). These may have been the result of strong divergent selection. Figure 6.19 shows the distribution of F_{st} values of all markers. While most markers are smaller than 0.05, a subset of markers reach F_{st} values higher than 0.10.

Genetic redundancy

The replicate Frequency Spectrum (RFS) is a measure of redundancy. Tables 6.8 and 6.9 show the number of SNP markers, identified as the top 100 based on different criteria, that change similarly in different clusters from G0 to G10 (AF change greater than 0.10, or greater than 0.30). Clear differences were observed. Based on F_{st}, F1 and F2 showed the greatest change, with over 60 SNP markers changing more than 0.10 in both, while 16 or less change similarly in the other families. In general, F4 shows the least change. Family 1 shows the greatest number of SNP with AF change greater than 0.30, and usually the most when AF change is 0.10. Table 6.9 compared changes of the 100 SNP with the greatest regression coefficient from G0 to G10 within each specific family, instead of the population as a whole. The AF change must be greater than 0.20, or greater than 0.30. Again, F4 has the fewest number of SNP that change more than 0.30. However, differences across-clusters are more similar when change is greater than 0.20. Based on these results, we conclude that genetic redundancy is indeed present.

Epistasis

Interactions among genes create more genetic diversity, may account for missing heritability (Mackay, 2014), and lead to different allele substitution effects across populations (Legarra et al., 2021). Epistasis and different substitution effects can be observed by comparing results from GWAS and the IGP obtained if different sets of SNP effects are used. The Manhattan plots for the GWAS using ALL or each cluster separately showed different peaks.

Only ALL and C4 had large enough numbers to identify significant markers, thus only they are presented in Fig 6.20. When ALL was used in our study, significant markers for stature were found on chromosomes 5, 11, 14, 15, 18, and 29, while C4 identified one on chromosome 14 and one on chromosome 20 as significant. Chromosome 5 was the only chromosome that appeared among the top 20 in all scenarios. According to Cole et al. (2011), chromosome 5 and 11 share significant markers for stature and body depth in Holstein, while the most significant markers on an autosomal chromosome for stature were on chromosome 11. More recently, Abo-Ismail et al. (2017) found that 30% of markers associated with body conformation in US Holstein cattle were located on chromosome 5, while a further 27% were on chromosome 18, and 5% on chromosome 14. Markers specifically associated with stature were on chromosomes 5, 11, 18, and 29, with chromosomes 5 and 18 having most significant markers. Chromosome 29 carried markers associated with stature in Chinese Holstein (Wu et al., 2013). Our smaller population sizes in C1, C2, C3, and C5 did not allow us the power to detect significant markers.

Table 6.6 presents the correlations between IGP based on different training populations as measures of genomic correlations between clusters. All are compared to the IGP within-cluster (e.g. IGP of males in cluster X when SNP effects were based on females of the same cluster). Correlations between within-cluster IGP and IGP based on ALL vary from 0.70 to 0.88. The across-cluster predictions have lower correlations with the IGP obtained when using within-cluster effects, ranging from 0.38 to 0.62. The correlations with ALL are expected to be higher than using across-cluster SNP effects since all clusters are represented in the training population. It has been shown that all breeds must be included in the training population in multi-breed evaluations to obtain high accuracies for each single breed. Across-breed prediction is poor, even when the training population include more than one breed (Pryce et al., 2011, Olson et al., 2012,

Raymond et al., 2018, Steyn et al., 2019). Our results suggest that the clusters among G10 are indeed genetically different. Non-additive gene actions, such as dominance and epistasis, can contribute to these differences. A correlation lower than 0.80 is not due to dominance alone. Epistasis may play a considerable role. In a simulation with realistic epistatic scenarios decreased the correlation between populations to as low as 0.45 (Duenk et al., 2020). The lowest correlation observed in our data, is 0.47, while the highest is 0.88. Thus, results suggest that epistasis is present.

Another indication of epistasis is a change in direction of AF change. The frequency may increase, or decrease initially, but change antagonistic relationships with other genes is present. These may be due to de novo mutations or a change in selection pressure, whether due to artificial or natural selection. Table 6.10 shows the number of SNP markers that have changed directions when comparing the regression coefficient of AF over G0 to G5, with the regression coefficient of AF over G5 to G10. While the number of SNP that reversed direction over time numbered in the thousands, the proportion of SNP markers that initially increased (or decreased) by 0.005 per generation until G5, and then changed at the same, but opposite, rate, range from 0.14 in F4 to 0.24 in F1.

Reranking of IGP

An important consideration when calculating the breeding value of an animal is the specification of the population(s) where the animal will be mated. Wade and Goodnight (1998) reviewed the assumptions used in Fisher's infinitesimal model and Wright's shifting balance theory. The US Holstein, under Fisher's model, would be described as a single, large, panmictic populations with a singular set of allele substitution effects. The average effect of an allele would be estimated from all animals combined and the targeted mates would be from a uniform

population. Under Wright's model, the Holstein breed consists of multiple demes (or families) with shifting allelic values depending upon the genetic background of these sub-populations. Genetic redundancy, across families, leads to the additive effect of an allele varying between families and resulting in a unique set of breeding values for each family. Our results are more in line with Wright's shifting balance theory where reranking is expected within different sub-populations.

Table 6.7 shows the ranking of ten specific bulls (two from each cluster) based on their IGP when using ALL, or each separate cluster. Reranking of bulls depending upon the target population is clearly evident. While Airlift ranks the highest of all the G10 males when using ALL SNP effects, he only ranks within the top 10 again when using SNP effects of C2. Alleles have different substitution effects in populations, and therefore alleles that are more favorable in one population, may not be favorable in another. Further investigation is required to determine whether results from this study can be used to improve the accuracy of genetic evaluations for the individual subgroups within the population. If subgroups are truly different, excluding one from the training population could mimic across-population genomic prediction, which consistently gives poor prediction accuracies even with high density markers and sophisticated techniques (Hayes et al., 2009, Karoui et al., 2012, Raymond et al., 2018). Genomic selection is more accurate when the training population is closely related to the target population (Clark et al., 2012, Chen et al., 2013). Current national evaluations include genotypes of many shared ancestors, which could ensure improvements in prediction accuracy across all subgroups.

Limitations

Replicate populations over time are useful for the observation of adaptation (Franssen et al., 2017). Ideally, these replicates should be from the same environment, share a founder

population, and evolve independently to the same environmental stressor (Barghi et al., 2020). In our study, the high overlap of animals in our families in older generations allowed us to have reasonable founding populations that are similar across families. Differences from this starting point over time reflect both divergent selection and genetic redundancy.

Genotypes of older Holstein animals in our data do not reflect a true baseline of animals that were part of the gene pool at the time, since they are generally animals that were predominantly bulls and considered to be best. However, these bulls were widely used in artificial insemination (AI) programs. Therefore, their genetic material is expected to be present in large proportions of the population. In 2015, it was shown that all AI bulls could be traced back to only 2 bulls born in 1880. Two highly influential bulls, Pawnee Farm Arlinda Chief and Round Oak Rag Apple shared Y-chromosomes with 48.78%, and 51.06% of the Holstein bull population in the 2010s (Yue et al., 2015). Both these bulls are included in our study. Our imbalance of sexes among older genotypes is not unlike the study by Barghi et al. (2019), where only females were genotyped in the founder population. During more recent decades, genotyping costs have decreased enough for breeders to genotype most of their animals, regardless of their genetic merit or sex. This will enable future studies to use AF that are more reflective of the whole population over time.

Other limitations of our study include the small number of replicates and generations. More generations will be better to detect the genetic change in different sub-populations. Additionally, our families are not closed families. Not all parents were genotyped and therefore, not all genetic changes over time are captured by our animals. This is also similar to the study by Barghi et al. (2020) where not all replicate members were genotyped.

The number of genotypes available among the selection candidates are too small to reach sufficiently accurate and significant GWAS results. However, even with our data, differences in results still support our hypothesis. Our work may serve as a preliminary study to encourage the investigation of adaptation and genetic redundancy in future once these obstacles are overcome. Considerably more genotypes have been collected since 2014. Their inclusion in future studies will greatly increase the ability to investigate the differences among sub-populations more thoroughly.

Implications

Non-parallel changes across families reveal genetic redundancy, however, since families may differ in genetic merit, this may be confounded with divergent selection. Whether due to selection or redundancy, non-parallel change shows underlying genetic diversity within the US Holstein breed. These results have important implications for the projection of long-term genetic response in the breed and other large panmictic populations. Genetic selection in Holstein cattle has achieved a continued increase in milk production with no apparent sign of reaching a selection limit for the trait. A question arises whether selection should continue treating the breed as one large population with AF that can converge to the best overall average AF for the population, or as separate lines to increase the genetic distance between families. The latter can potentially prevent the loss of alleles that may be beneficial in future and allow outcrossing that can take advantage of heterosis.

CONCLUSIONS

The Holstein breed is a complex mixture of family subgroups with different allele frequencies and gene combinations. The different families offer redundant solutions to the goals of modern-day breeders. Genetic redundancy allows for the value of individual alleles to shift

over time in unique ways within a specific family. The substitution value of different alleles, and consequently the breeding value, will differ for different target populations such as a specific family versus the overall combined population. Stratification of selection candidates into unique subpopulations provides genetic redundancy, maintains diversity, and lowers the risk of the fixation, or loss, of alleles.

REFERENCES

Abo-Ismail, M. K., L. F. Brito, S. P. Miller, M. Sargolzaei, D. A. Grossi, S. S. Moore, G. Plastow, P. Stothard, S. Nayeri, and F. S. Schenkel. 2017. Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle. Genetics Selection Evolution 49(1):82.

Aguilar, I. and I. Misztal. 2008. Technical Note: Recursive Algorithm for Inbreeding Coefficients Assuming Nonzero Inbreeding of Unknown Parents. Journal of Dairy Science 91(4):1669-1672.

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta, and T. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of dairy science 93(2):743-752.

Baller, J. L., J. T. Howard, S. D. Kachman, and M. L. Spangler. 2019. The impact of clustering methods for cross-validation, choice of phenotypes, and genotyping strategies on the accuracy of genomic predictions. Journal of Animal Science 97(4):1534-1549.

Barbosa da Silva, M. V. G., T. S. Sonstegard, R. M. Thallman, E. E. Connor, R. D. Schnabel, and C. P. Van Tassell. 2010. Characterization of DGAT1 Allelic Effects in a Sample of North American Holstein Cattle. Animal Biotechnology 21(2):88-99.

Barghi, N., J. Hermisson, and C. Schlötterer. 2020. Polygenic adaptation: a unifying framework to understand positive selection. Nature Reviews Genetics.

Barghi, N., R. Tobler, V. Nolte, A. M. Jakšić, F. Mallard, K. A. Otte, M. Dolezal, T. Taus, R. Kofler, and C. Schlötterer. 2019. Genetic redundancy fuels polygenic adaptation in Drosophila. PLoS biology 17(2):e3000128.

Boddhireddy, P., M. J. Kelly, S. Northcutt, K. C. Prayaga, J. Rumph, and S. DeNise. 2014. Genomic predictions in Angus cattle: Comparisons of sample size, response variables, and clustering methods for cross-validation. Journal of Animal Science 92(2):485-497.

Boyle, E. A., Y. I. Li, and J. K. Pritchard. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell 169(7):1177-1186.

Bulmer. 1971. The effect of selection on genetic variability. The American Naturalist 105(943):201-211.

Chen, C.-Y., I. Misztal, I. Aguilar, A. Legarra, and W. Muir. 2011. Effect of different genomic relationship matrices on accuracy and scale. Journal of animal science 89(9):2673-2679.

CDCB. 2020. Council on Dairy Cattle Breeding Activity Report Oct 19/Sep 2020. Accessed Dec. 2, 2020.https: / / www .uscdcb .com/ wp -content/ uploads/ 2020/ 10/ 2020 -CDCB -Activity -Report _103020 _lowres .pdf

Chen, L., C. Li, F. Schenkel, M. Vinsky, and D. H. Crews, Jr. 2013. Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. Journal of Animal Science 91(10):4669-4678.

Clark, S. A., J. M. Hickey, H. D. Daetwyler, and J. H. J. van der Werf. 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetics Selection Evolution 44(1):4.

Cole, J. B., G. R. Wiggans, L. Ma, T. S. Sonstegard, T. J. Lawlor, B. A. Crooker, C. P. Van Tassell, J. Yang, S. Wang, L. K. Matukumalli, and Y. Da. 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics 12(1):408.

Duenk, P., P. Bijma, M. P. L. Calus, Y. C. J. Wientjes, and J. H. J. van der Werf. 2020. The Impact of Non-additive Effects on the Genetic Correlation Between Populations. G3 Genes|Genomes|Genetics 10(2):783-795.

Falconer, D. and T. Mackay. 1996. Introduction to quantitative genetics. 4 ed. Pearson, Essex, UK.

Flori, L., S. Fritz, F. Jaffrézic, M. Boussaha, I. Gut, S. Heath, J.-L. Foulley, and M. Gautier. 2009. The Genome Response to Artificial Selection: A Case Study in Dairy Cattle. PLOS ONE 4(8):e6595.

Franssen, S. U., R. Kofler, and C. Schlötterer. 2017. Uncovering the genetic signature of quantitative trait evolution with replicated time series data. Heredity 118(1):42-51.

Goldstein, D. B. and K. E. Holsinger. 1992. Maintenance of polygenic variation in spatially structured populations: Roles for local mating and genetic redundancy. Evolution 46(2):412-429.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42(1):5.

Hartigan, J. A. and M. A. Wong. 1979. A K-Means Clustering Algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics) 28(1):100-108.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41(1):51.

Höllinger, I., P. S. Pennings, and J. Hermisson. 2019. Polygenic adaptation: From sweeps to subtle frequency shifts. PLOS Genetics 15(3):e1008035.

Jiang, J., L. Ma, D. Prakapenka, P. M. VanRaden, J. B. Cole, and Y. Da. 2019. A Large-Scale Genome-Wide Association Study in U.S. Holstein Cattle. Frontiers in Genetics 10(412).

Jiang, J., D. Prakapenka, L. Ma, J. Cole, P. VanRaden, and Y. Da. 2018. Extreme antagonistic pleiotropy effects of DGAT1 on fat, milk and protein yields. in Proc. Proceedings of the World Congress on Genetics Applied to Livestock Production.

Karoui, S., M. J. Carabaño, C. Díaz, and A. Legarra. 2012. Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. Genetics Selection Evolution 44(1):39.

Laurentino, S., J. Gonçalves, J. E. Cavaco, P. F. Oliveira, M. G. Alves, M. de Sousa, A. Barros, and S. Socorro. 2011. Apoptosis-inhibitor Aven is downregulated in defective spermatogenesis and a novel estrogen target gene in mammalian testis. Fertility and Sterility 96(3):745-750.

Legarra, A., C. A. Garcia-Baccino, Y. C. J. Wientjes, and Z. G. Vitezica. 2021. The Correlation of Substitution Effects Across Populations and Generations in the Presence of Non-Additive Functional Gene Action. bioRxiv:2020.2011.2003.367227.

Lewontin, R. C. and J. Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74(1):175-195.

Liu, X., Y. I. Li, and J. K. Pritchard. 2019. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. Cell 177(4):1022-1034.e1026.

Ma, L., T. S. Sonstegard, J. B. Cole, C. P. VanTassell, G. R. Wiggans, B. A. Crooker, C. Tan, D. Prakapenka, G. E. Liu, and Y. Da. 2019. Genome changes due to artificial selection in U.S. Holstein cattle. BMC Genomics 20(1):128.

Mackay, T. F. C. 2014. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nature Reviews Genetics 15(1):22-33.

Misztal, I., A. Legarra, and I. Aguilar. 2014a. Using recursion to compute the inverse of the genomic relationship matrix. Journal of Dairy Science 97(6):3943-3952.

Misztal, I., S. Tsuruta, D. Lourenco, I. Aguilar, A. Legarra, and Z. Vitezica. 2014b. Manual for BLUPF90 family of programs. in Athens: University of Georgia.

Nowak, M. A., M. C. Boerlijst, J. Cooke, and J. M. Smith. 1997. Evolution of genetic redundancy. Nature 388(6638):167-171.

Olson, K., P. VanRaden, and M. Tooker. 2012. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science 95(9):5378-5383.

Pickett, F. B. and D. R. Meeks-Wagner. 1995. Seeing double: appreciating genetic redundancy. Plant Cell 7(9):1347-1356.

Prakapenka, D., Z. Liang, J. Jiang, L. Ma, and Y. Da. 2021. A Large-Scale Genome-Wide Association Study of Epistasis Effects of Production Traits and Daughter Pregnancy Rate in U.S. Holstein Cattle. Genes 12(7):1089.

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

Pryce, J. E., M. Haile-Mariam, M. E. Goddard, and B. J. Hayes. 2014. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genetics Selection Evolution 46(1):71.

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

Raymond, B., A. C. Bouwman, C. Schrooten, J. Houwing-Duistermaat, and R. F. Veerkamp. 2018. Utility of whole-genome sequence data for across-breed genomic prediction. Genetics Selection Evolution 50(1):27.

Reverter, A., J. Henshall, L. Porto-Neto, F. Raidan, Y. Li, M. Naval-Sánchez, S. Lehnert, K. Meyer, and Z. Vitezica. 2018. A rapid method for the identification of epistatic ‘dormant’SNPs. in Proc. Proceedings of the 11th World Congress on Genetics Applied to Livestock Production Methods and Tools-GWAS.

Rowan, T. N., H. J. Durbin, C. M. Seabury, R. D. Schnabel, and J. E. Decker. 2020. Powerful detection of polygenic selection and environmental adaptation in US beef cattle. bioRxiv:2020.2003.2011.988121.

Saatchi, M., R. D. Schnabel, M. M. Rolf, J. F. Taylor, and D. J. Garrick. 2012. Accuracy of direct genomic breeding values for nationally evaluated traits in US Limousin and Simmental beef cattle. Genetics Selection Evolution 44(1):38.

Spelman, R., C. Ford, P. McElhinney, G. Gregory, and R. Snell. 2002. Characterization of the DGAT1 gene in the New Zealand dairy population. Journal of Dairy Science 85(12):3514-3517.

Steyn, Y., D. A. L. Lourenco, and I. Misztal. 2019. Genomic predictions in purebreds with a multibreed genomic relationship matrix. Journal of Animal Science 97(11):4418–4427.

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

Tidcombe, H., A. Jackson-Fisher, K. Mathers, D. F. Stern, M. Gassmann, and J. P. Golding. 2003. Neural and mammary gland defects in ErbB4 knockout mice genetically rescued from embryonic lethality. Proceedings of the National Academy of Sciences 100(14):8281-8286.

Tsuruta, S., T. J. Lawlor, D. A. L. Lourenco, and I. Misztal. 2021. Bias in genomic predictions by mating practices for linear type traits in a large-scale genomic evaluation. Journal of Dairy Science 104(1):662-677.

Turelli, M. 2017. Commentary: Fisher's infinitesimal model: A story for the ages. Theoretical Population Biology 118:46-49.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91(11):4414-4423.

Wade, M. J. and C. J. Goodnight. 1998. Perspective: The theories of fisher and wright in the context of metapopulations: When nature does many small experiments. Evolution 52(6):1537-1553.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics Research 94(2):73-83.

Wright, S. 1943. Isolation by Distance. Genetics 28(2):114-138.

Wu, X., M. Fang, L. Liu, S. Wang, J. Liu, X. Ding, S. Zhang, Q. Zhang, Y. Zhang, L. Qiao, M. S. Lund, G. Su, and D. Sun. 2013. Genome wide association studies for body conformation traits in the Chinese Holstein cattle population. BMC Genomics 14(1):897.

Yuan, S., C. J. Stratton, J. Bao, H. Zheng, B. P. Bhetwal, R. Yanagimachi, and W. Yan. 2015. Spata6 is required for normal assembly of the sperm connecting piece and tight head–tail conjunction. Proceedings of the National Academy of Sciences 112(5):E430-E439.

Yue, X.-P., C. Dechow, and W.-S. Liu. 2015. A limited number of Y chromosome lineages is present in North American Holsteins. Journal of Dairy Science 98(4):2738-2745.

Zhang, J., P. Zhang, Y. Wei, H.-l. Piao, W. Wang, S. Maddika, M. Wang, D. Chen, Y. Sun, M.-C. Hung, J. Chen, and L. Ma. 2013. Deubiquitylation and stabilization of PTEN by USP13. Nature Cell Biology 15(12):1486-1494.

TABLES

Table 6.1: The number of genotyped animals per generation within each family. The most recent generation (G10) was used for clustering. Their pedigrees were traced back 10 generations (G9 to G0).

Generation	**Family 1**	**Family 2**	**Family 3**	**Family 4**	**Family 5**
G10	3,577	3,073	3,302	5,931	4,216
G9	1,513	1,471	1,478	2,830	1,859
G8	1,337	1,242	1,302	2,364	1,591
G7	1,043	975	1,004	1,666	1,161
G6	838	792	797	1,139	914
G5	645	582	608	839	669
G4	467	432	454	603	489
G3	336	304	310	426	346
G2	243	229	221	299	245
G1	189	171	171	223	183
G0	148	135	137	171	146

Table 6.2: The average expected inbreeding of offspring resulting from hypothetical mating within-cluster, and across-cluster. The expected inbreeding if animals were mated at random is 0.12.

Cluster ID	Inbreeding within-cluster	Inbreeding across-cluster
1	0.22	0.11
2	0.20	0.11
3	0.18	0.12
4	0.10	0.10
5	0.17	0.11

Table 6.3: The number of times a prominent bull appears as a sire of animals in G10 (or G9) of each family.

Name	Family 1	Family 2	Family 3	Family 4	Family 5
Planet	1 (321)	0 (42)	0 (27)	0 (13)	0 (104)
Goldwyn	0 (99)	449 (399)	0 (92)	0 (43)	0 (158)
Shottle	0 (209)	0 (171)	584 (492)	0 (44)	0 (222)
Domain	22 (0)	42 (0)	118 (0)	276 (7)	43 (1)
BW Marshall	0 (7)	0 (16)	0 (25)	34 (65)	2 (25)
Oman	0 (77)	0 (49)	0 (49)	1 (20)	95 (223)

Table 6.4: The number of times historically prominent ancestors appear across all generations in each family

Name	Family 1	Family 2	Family 3	Family 4	Family 5
Ivanhoe Star	7	7	6	7	7
Chairman	8	8	6	9	7
Chief	10	12	12	14	11
Elevation	17	20	20	26	18
Bell	29	25	24	33	26
Starbuck	24	32	23	34	26
Blackstar	49	47	44	53	41

Table 6.5: The number of animals in common for all families per generation, and number of animals unique to each family within each generation.

Generation	Common	Unique					
		Family 1	Family 2	Family 3	Family 4	Family 5	All within generation
G10	0	4,355	3,555	3,574	7,107	4,879	20,099
G9	55	783	703	623	1,862	831	6,546
G8	253	380	290	247	952	392	4,150
G7	397	169	141	120	469	169	2,478
G6	411	92	84	72	234	98	1,635
G5	351	66	42	41	159	66	1,142
G4	288	38	34	25	98	41	793
G3	218	25	14	21	75	24	540
G2	169	17	12	6	48	20	370
G1	131	8	5	5	24	7	261
G0	108	7	3	4	17	6	200

Table6. 6: The Pearson correlations between indirect genomic predictions (IGP) obtained from SNP effects estimated with different populations. The benchmark for comparison was the IGP obtained from within-cluster analysis (thus males of cluster X based on SNP effects estimated using females of the same cluster).

SNP effects used (females)	**Cluster predicted (males)**				
	C1	**C2**	**C3**	**C4**	**C5**
C1	1.00	0.47	0.59	0.48	0.58
C2	0.41	1.00	0.49	0.47	0.38
C3	0.45	0.45	1.00	0.54	0.59
C4	0.56	0.53	0.62	1.00	0.57
C5	0.53	0.38	0.58	0.44	1.00
All	0.76	0.70	0.75	0.88	0.78

Table 6.7: The ranking of two bulls from each cluster when IGP for stature is based on different SNP effects.

Bull	Relationship	Group SNP effects were based on					
		All	C1	C2	C3	C4	C5
Bulls in G10 of cluster 1							
Doorman	Planet grandson	43	158	825	871	152	777
McCutchen	Planet grandson	56	636	583	56	729	814
Bulls in G10 of cluster 2							
Airlift	Goldwyn grandson	1	101	3	64	44	276
G.W. Atwood	Goldwyn son	22	81	8	1150	212	446
Bulls in G10 of cluster 3							
Edison	Domain son	21	94	15	1	233	50
Scorpio	Domain son	326	265	272	9	281	524
Bulls in G10 of cluster 4							
Chuckie	Domain son	12	892	173	26	1	234
Domain	Sons in C3 and C4	167	465	114	15	114	153
Bulls in G10 of cluster 5							
Monreal	Oman grandson	2	46	97	140	30	2
Broch	Oman grandson	4	52	146	124	19	57

Table 6.8: Replicate frequency spectrum (RFS): The number of SNP identified as top 100 for F_{st}, greatest variance, or greatest range that have shown an allele frequency change greater than 0.10 (or greater than 0.30) from earliest to last generation in each family.

Top 100 SNP	Family 1	Family 2	Family 3	Family 4	Family 5
Highest F_{st}	61 (5)	63 (4)	15 (0)	13 (0)	16(0)
Variance	67 (56)	59 (34)	74 (30)	48 (1)	52 (17)
Range	71 (66)	56 (28)	64 (17)	60 (1)	70 (14)

Table 6.9: Replicate frequency spectrum (RFS): The number of SNP that have allele frequency changes (AFC) greater than 0.20 or 0.30 from oldest to youngest generation in each family, when SNP markers are selected based on the greatest absolute AF change when regressing allele frequencies over generations within each family.

	Top 100 SNP in Family 1				
AFC	Family 1	Family 2	Family 3	Family 4	Family 5
>0.20	100	75	89	74	86
>0.30	100	45	61	26	59
	Top 100 SNP in Family 2				
>0.20	83	100	97	89	77
>0.30	57	100	68	25	51
	Top 100 SNP in Family 3				
>0.20	76	87	100	95	75
>0.30	60	61	100	31	46
	Top 100 SNP in Family 4				
>0.20	82	90	99	100	90
>0.30	68	67	78	55	60
	Top 100 SNP in Family 5				
>0.20	69	75	93	83	100
>0.30	50	57	57	20	100

Table 6.10: The number and proportion of SNP markers that have changed direction in each family. These are SNP that changed at a rate of 0.02, 0.01, or 0.005 allele frequencies per generation in one direction (positive or negative) from G0 to G5, and the same (but opposite) rate from G5 to G10.

	Rate of allele frequency change					
Family	>0.02		>0.01		>0.005	
	Number	Proportion	Number	Proportion	Number	Proportion
Family 1	1086	0.02	6765	0.11	14132	0.24
Family 2	780	0.01	5986	0.10	13012	0.22
Family 3	833	0.01	6238	0.11	13161	0.22
Family 4	56	0.00	2172	0.04	8121	0.14
Family 5	839	0.01	6285	0.11	13450	0.23

FIGURES

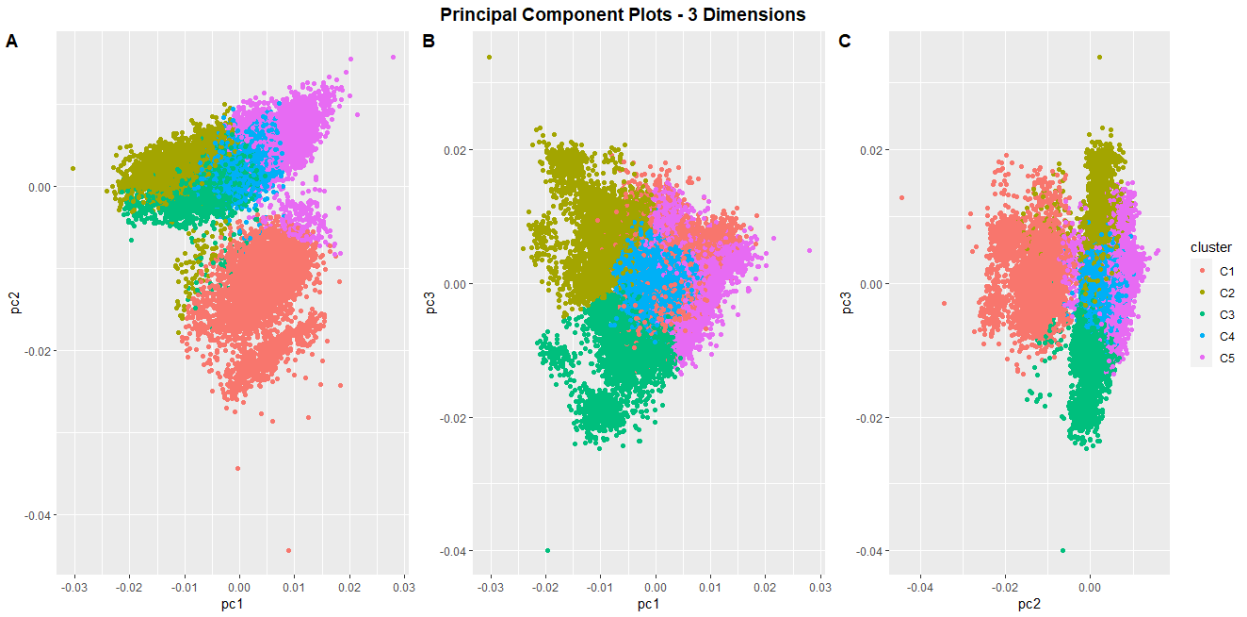


Figure 6.1: Principal component analyses plots for three dimensions showing the clustering results of selected candidates (G10).

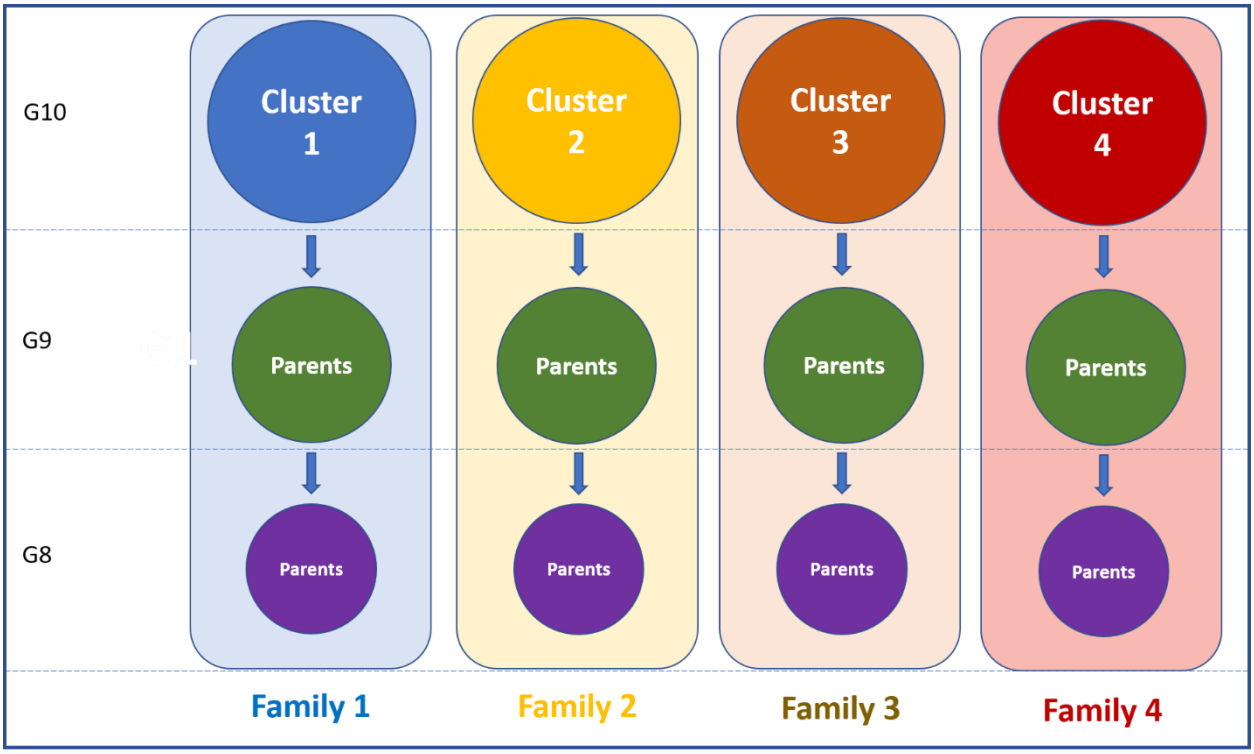
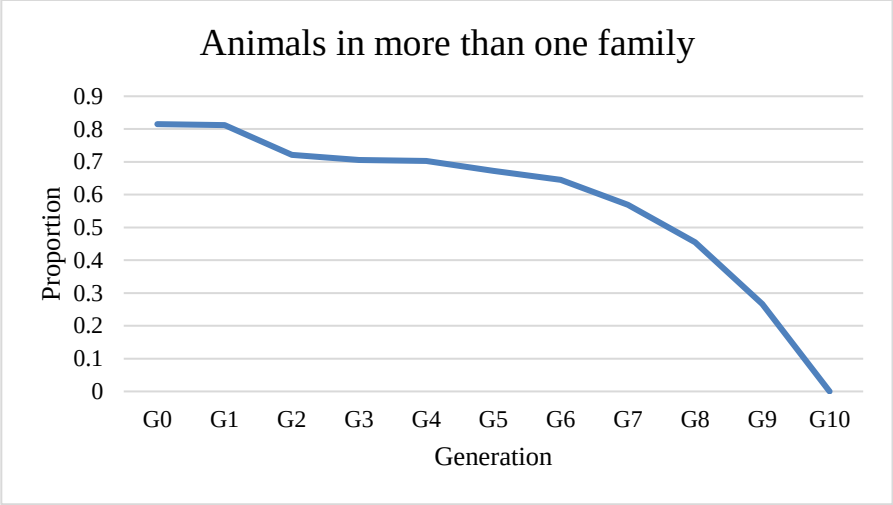


Figure 6.2: The resulting 5 clusters (here 4 are visualized) were considered as generation 10 (G10). From each cluster, pedigrees were traced back 10 generations (G9 to G0). Animals from a specific generation within a specific family were unique, but animals may appear in the same generation of other families. Due to overlapping generations, animals in one generation may also appear in other generations of the same family, or other families.



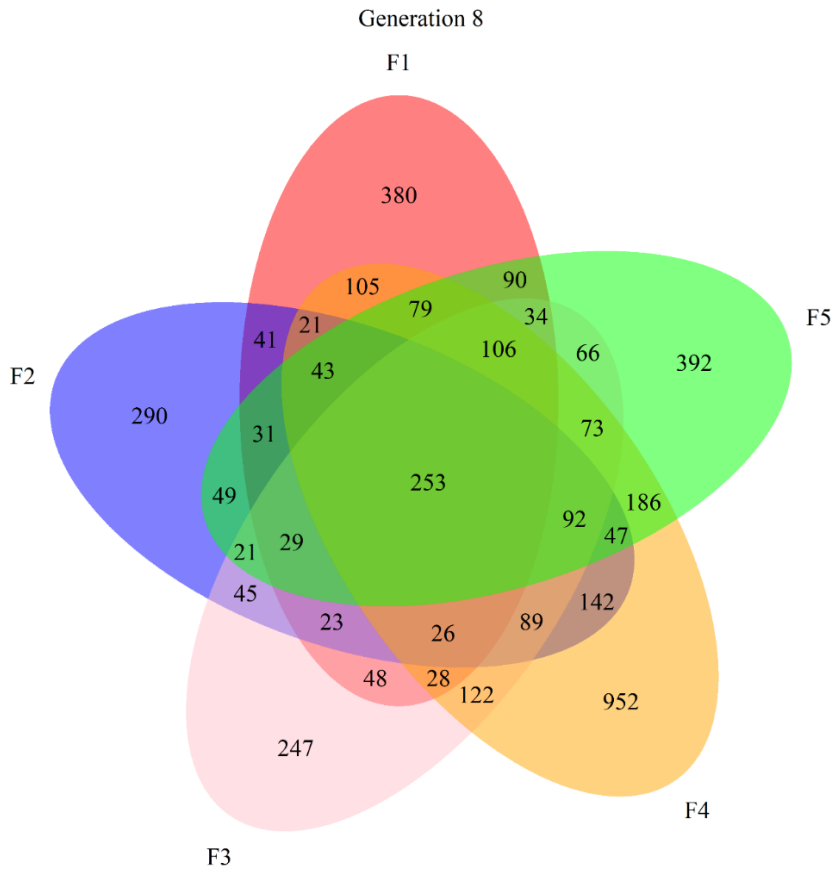


Figure 6.3: The proportion of animals that appear in more than one family in each generation, and a Venn diagram illustrating the overlapping nature of families in generation 8 (G8)

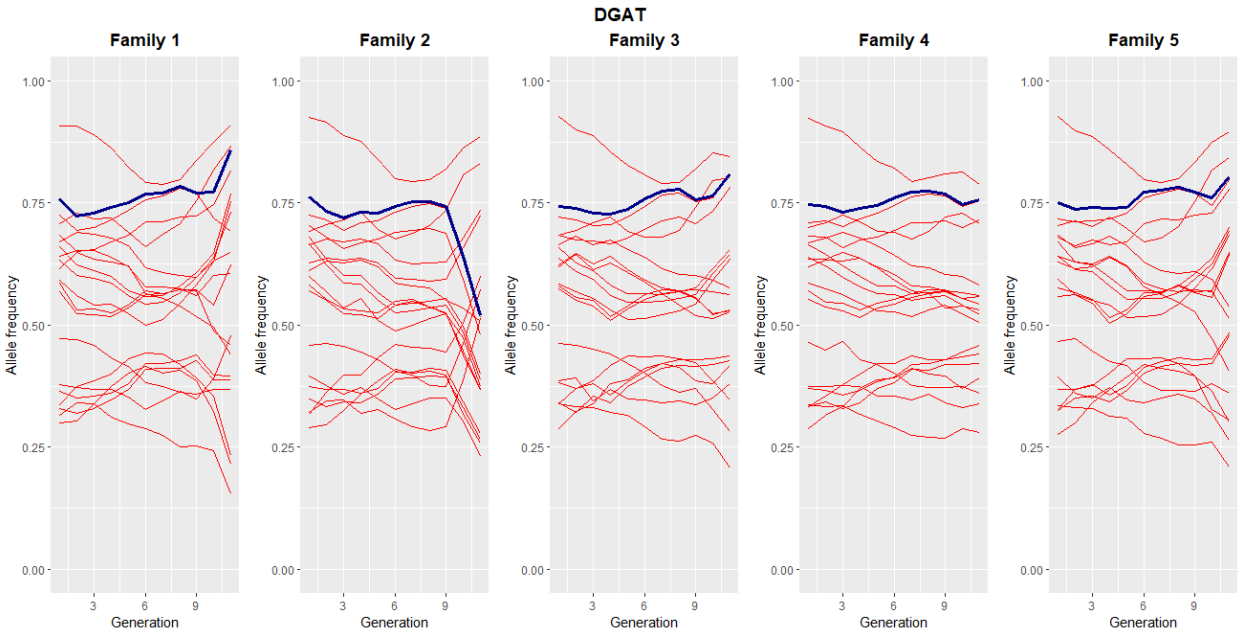


Figure 6.4: The allele frequency of the *DGAT* gene (blue) and surrounding 20 SNP markers (red) per genration within each family.

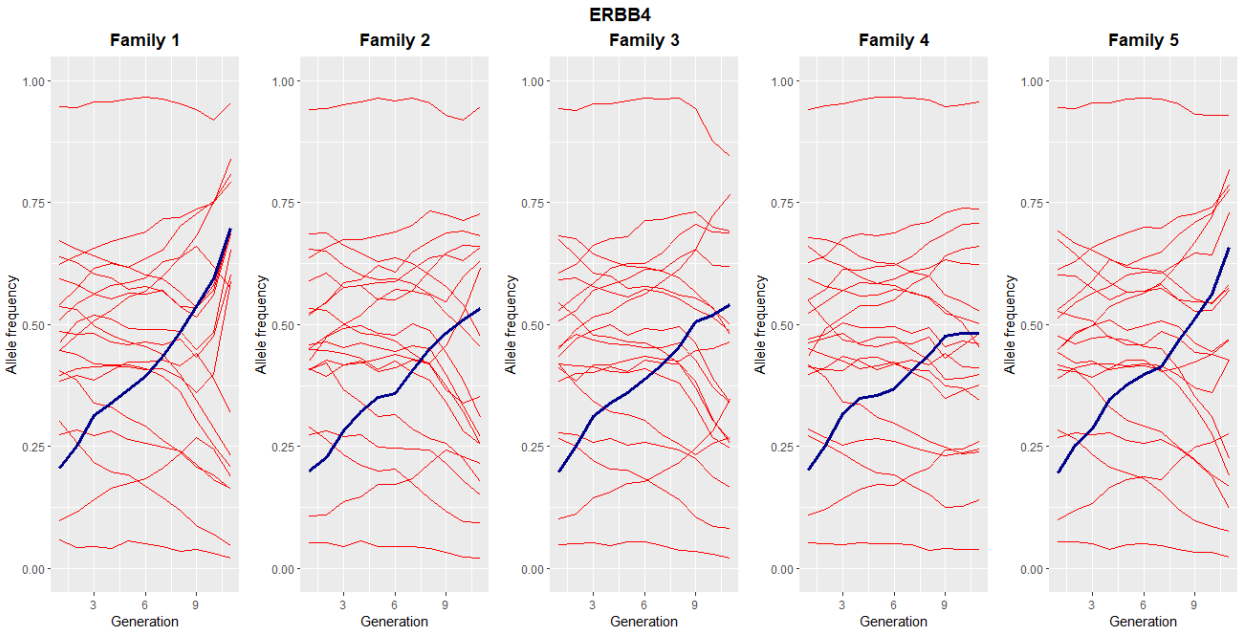


Figure 6.5: The allele frequency of the *ERBB4* gene (blue) and surrounding 20 SNP markers (red) per generation within each family.

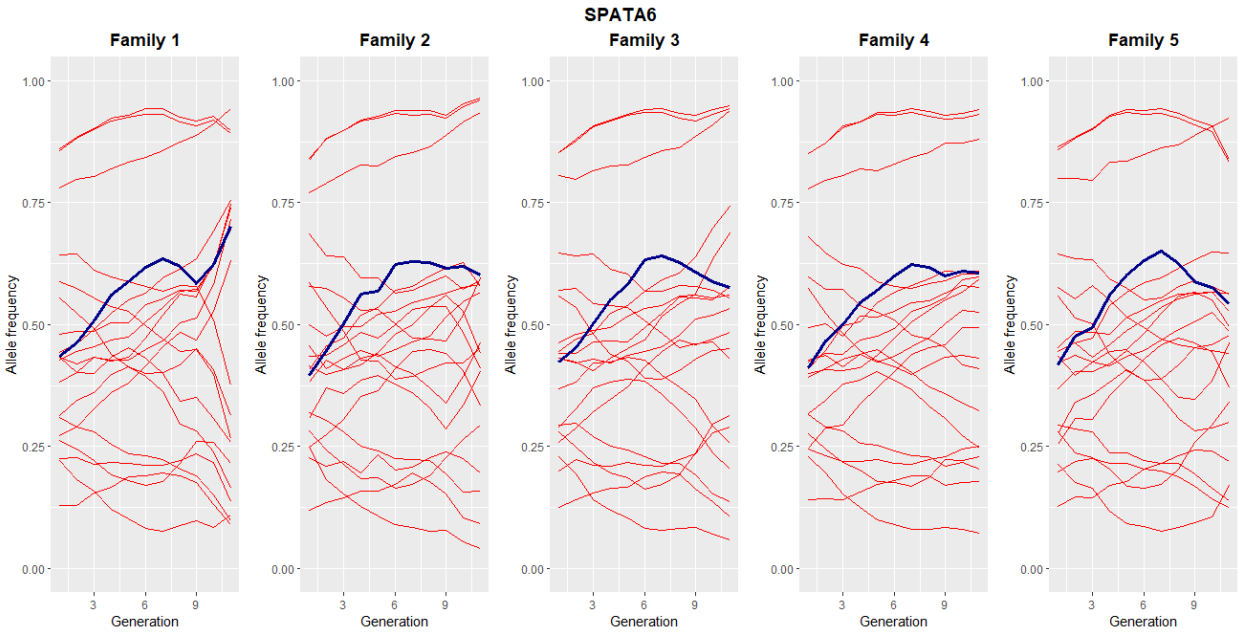


Figure 6.6: The allele frequency of the *SPATA6* gene (blue) and surrounding 20 SNP markers (red) per generation within each family.

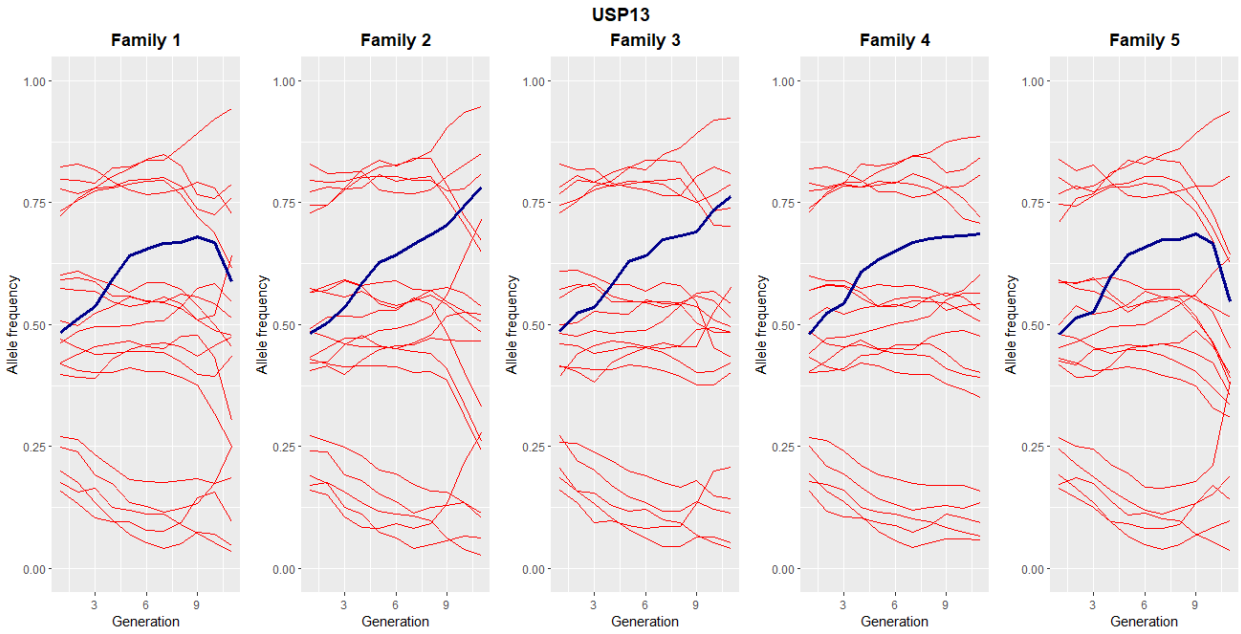


Figure 6.7: The allele frequency of the *USP13* gene (blue) and surrounding 20 SNP markers (red) per generation within each family.

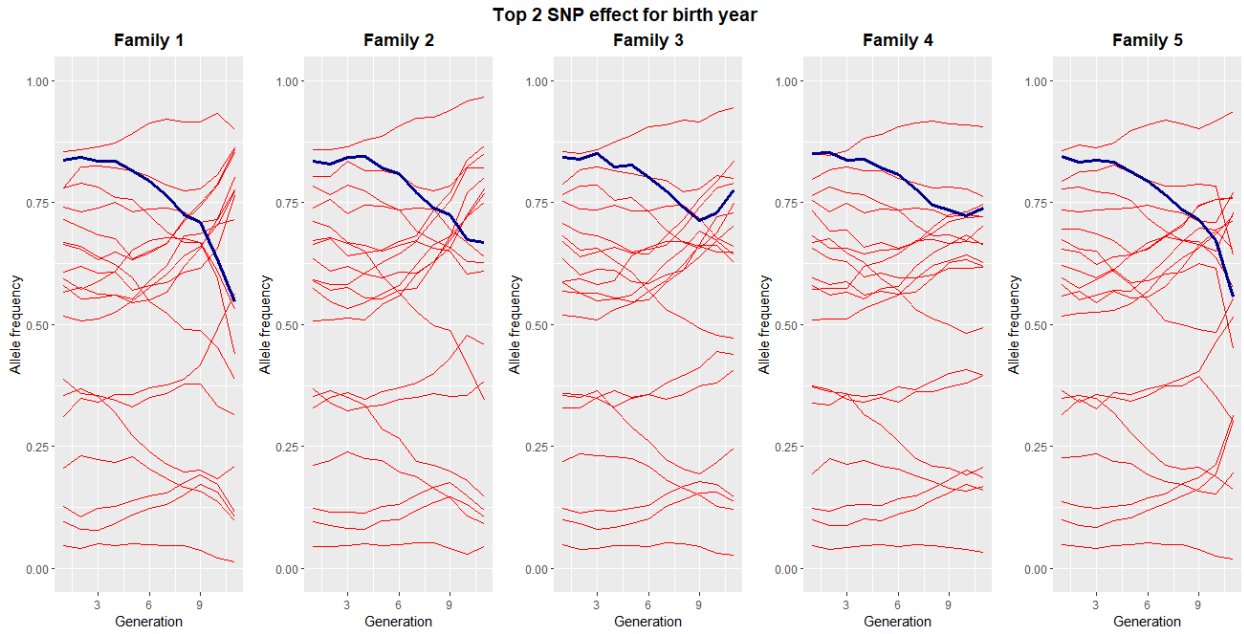


Figure 6.8: The allele frequency of the SNP second strongest associated with birth year (blue) and the surrounding 20 SNP markers (red) per generation within each family

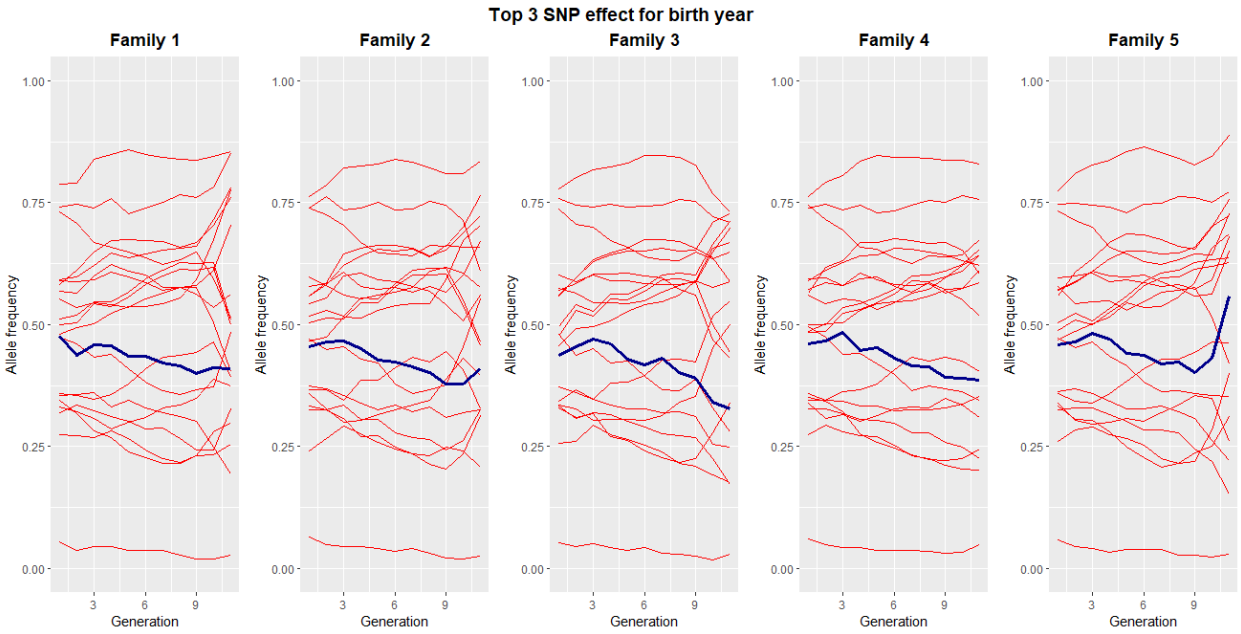


Figure 6.9: The allele frequency of the SNP third strongest associated with birth year (blue) and the surrounding 20 SNP markers (red) per generation within each family

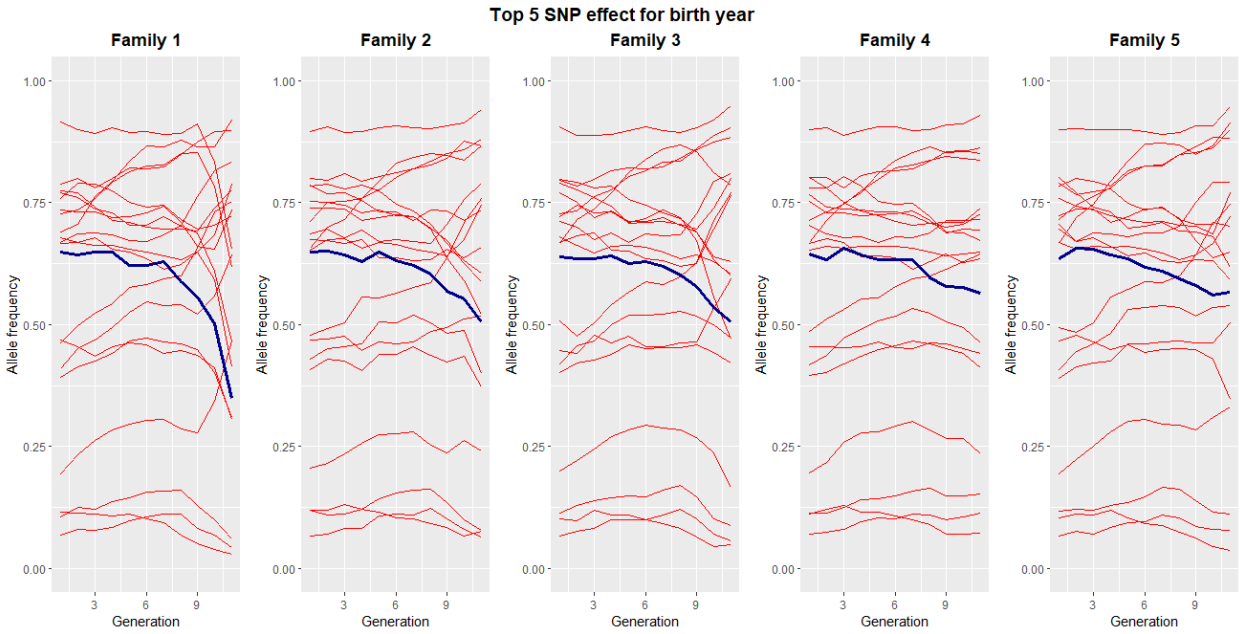


Figure 6.10: The allele frequency of the SNP fifth strongest associated with birth year (blue) and the surrounding 20 SNP markers (red) per generation within each family

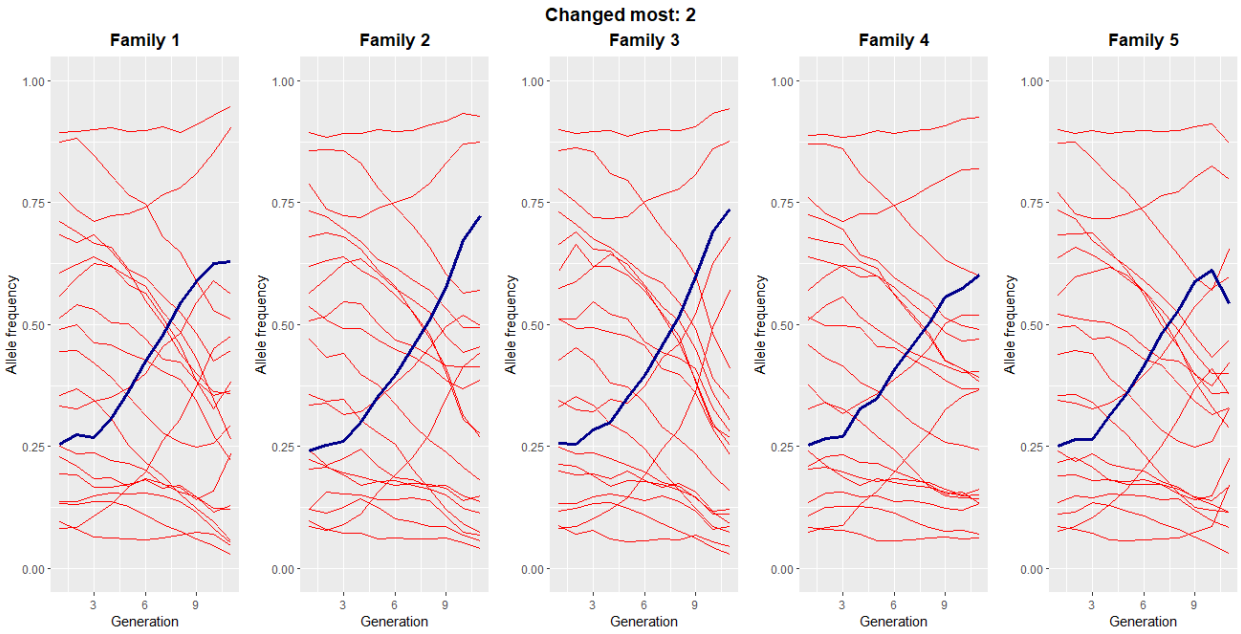


Figure 6.11: The allele frequency of the second selected SNP among the 20 SNP markers that have changed the most over time based on the regression coefficient (blue) and the surrounding 20 SNP markers (red) per generation within each family. The rate and direction of change are generally similar across families, with the exception of F5.

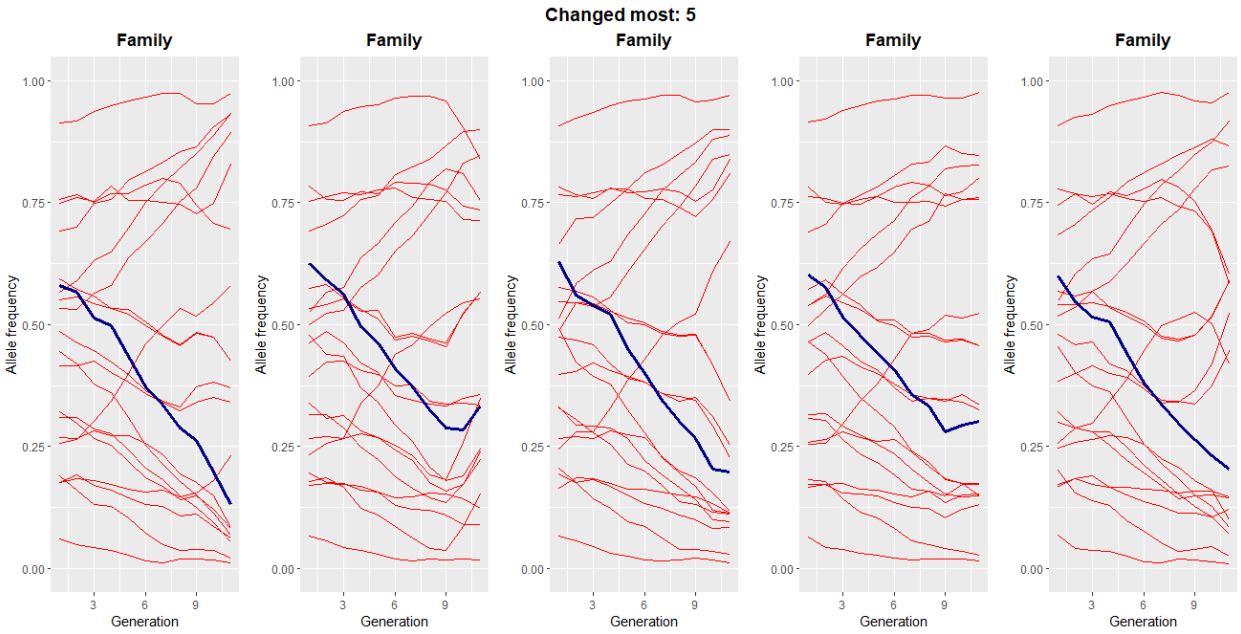


Figure 6.12: The allele frequency of the fifth selected SNP among the 20 SNP markers that have changed the most over time based on the regression coefficient (blue) and the surrounding 20 SNP markers (red) per generation within each family. The rate and direction of change are generally similar across families, with the exception of F2 and F4 that change direction from G9.

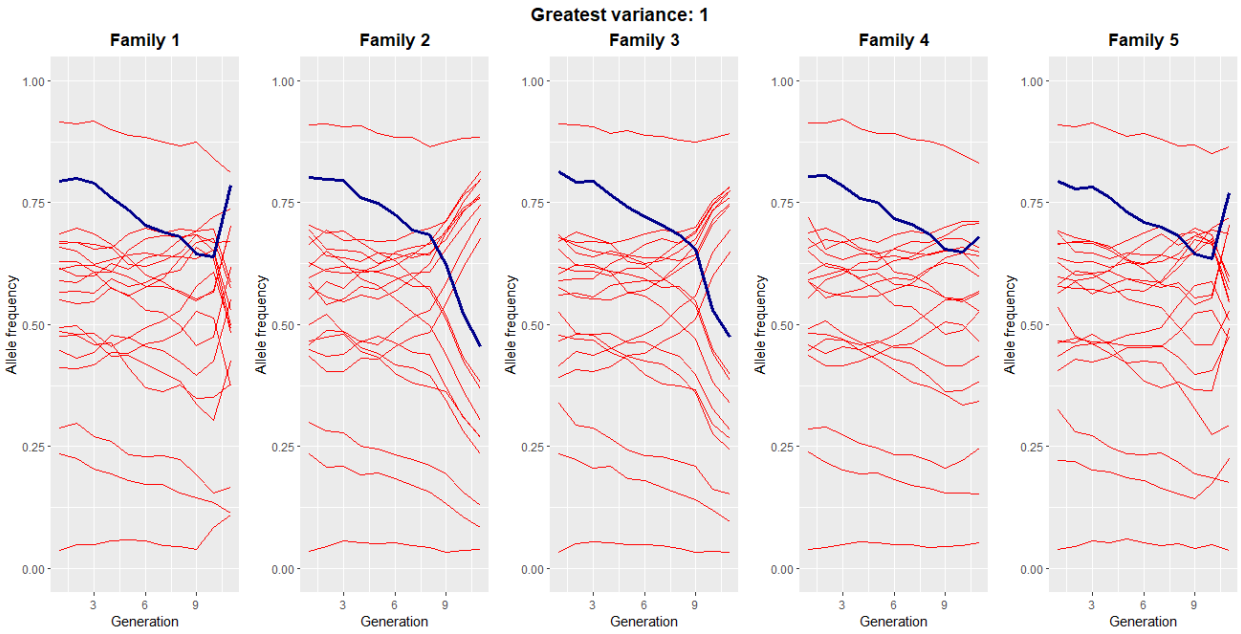


Figure 6.13: The allele frequency of the first selected SNP among the 20 SNP markers that have shown different changes across families (blue) and the surrounding 20 SNP markers (red) per generation within each family. Changes in F1 and F5 share the same pattern (decrease followed by a sharp increase), while F2 and F3 share the same pattern (continue to decrease). Many surrounding SNP markers follow the same, or opposite trend as the selected marker.

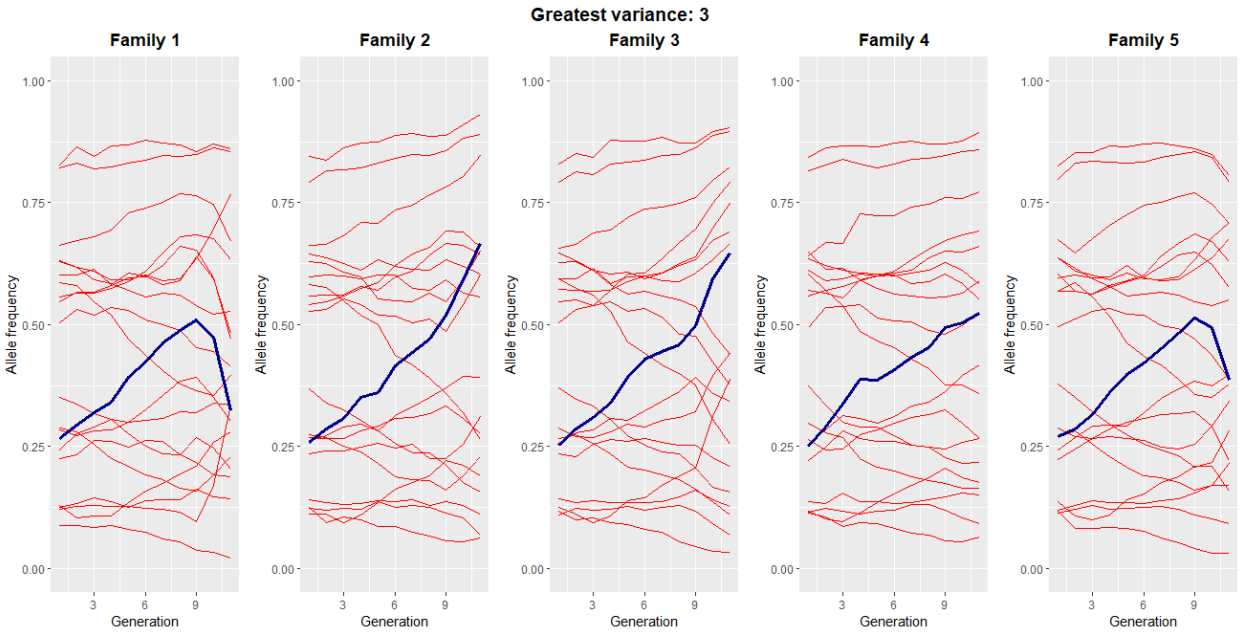


Figure 6.14: The allele frequency of the first selected SNP among the 20 SNP markers that have shown different changes across families (blue) and the surrounding 20 SNP markers (red) per generation within each family. Changes in F1 and F5 share the same pattern (decrease followed by a sharp increase), while F2 and F3 share the same pattern (continue to decrease).

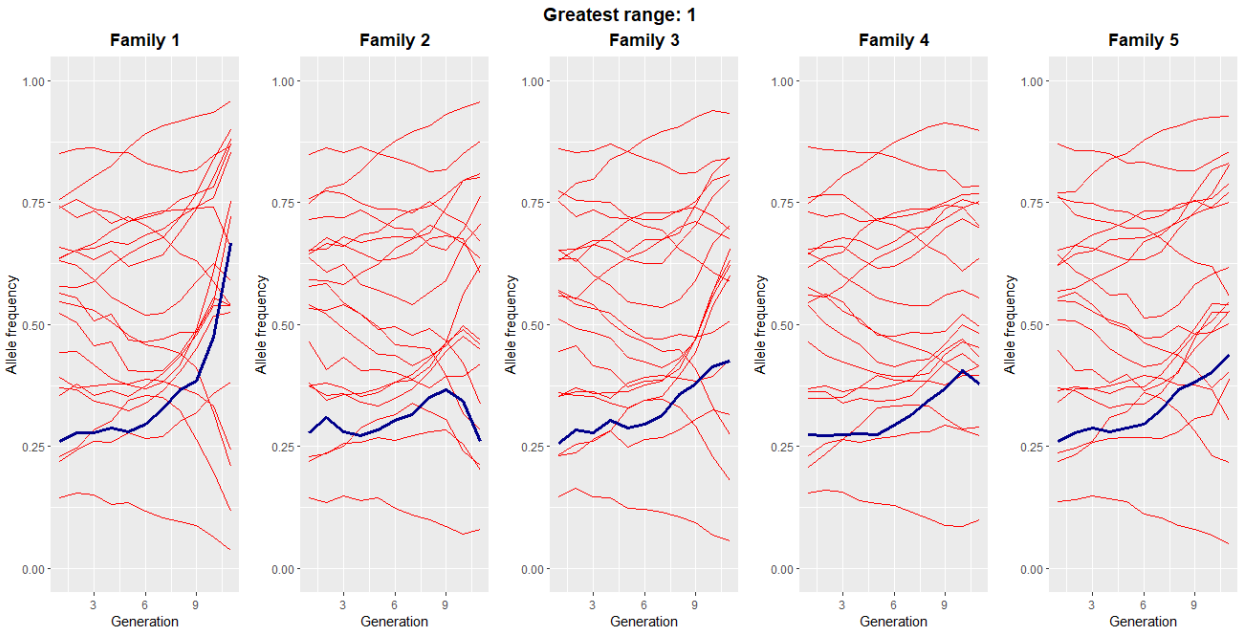


Figure 6.15: The allele frequency of the first selected SNP among the 20 SNP markers that have shown small changes in at least one family and large changes in another (blue) and the surrounding 20 SNP markers (red) per generation within each family.

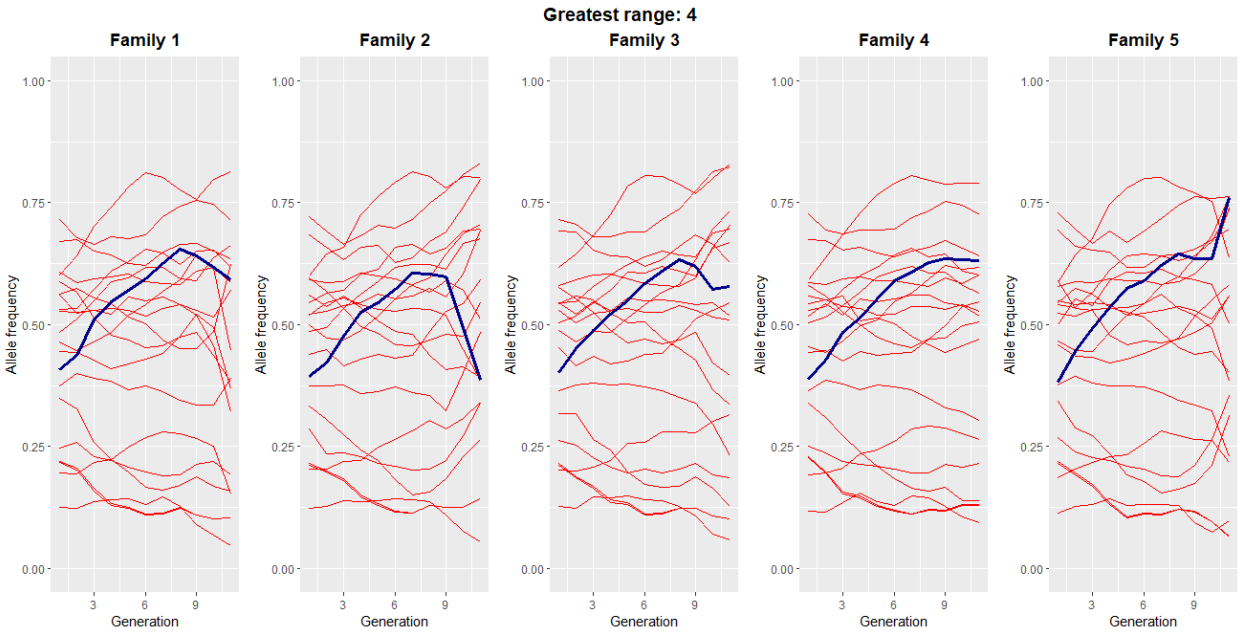


Figure 6.16: The allele frequency of the fourth selected SNP among the 20 SNP markers that have shown small changes in at least one family and large changes in another (blue) and the surrounding 20 SNP markers (red) per generation within each family

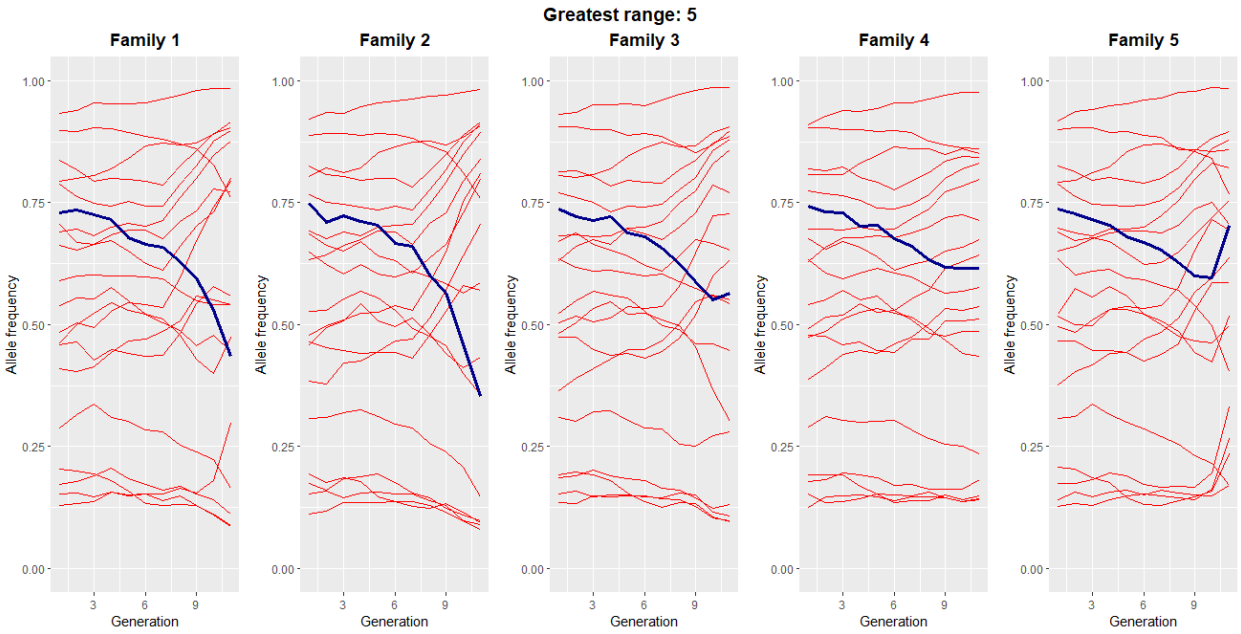


Figure 6.17: The allele frequency of the fifth selected SNP among the 20 SNP markers that have shown small changes in at least one family and large changes in another (blue) and the surrounding 20 SNP markers (red) per generation within each family.

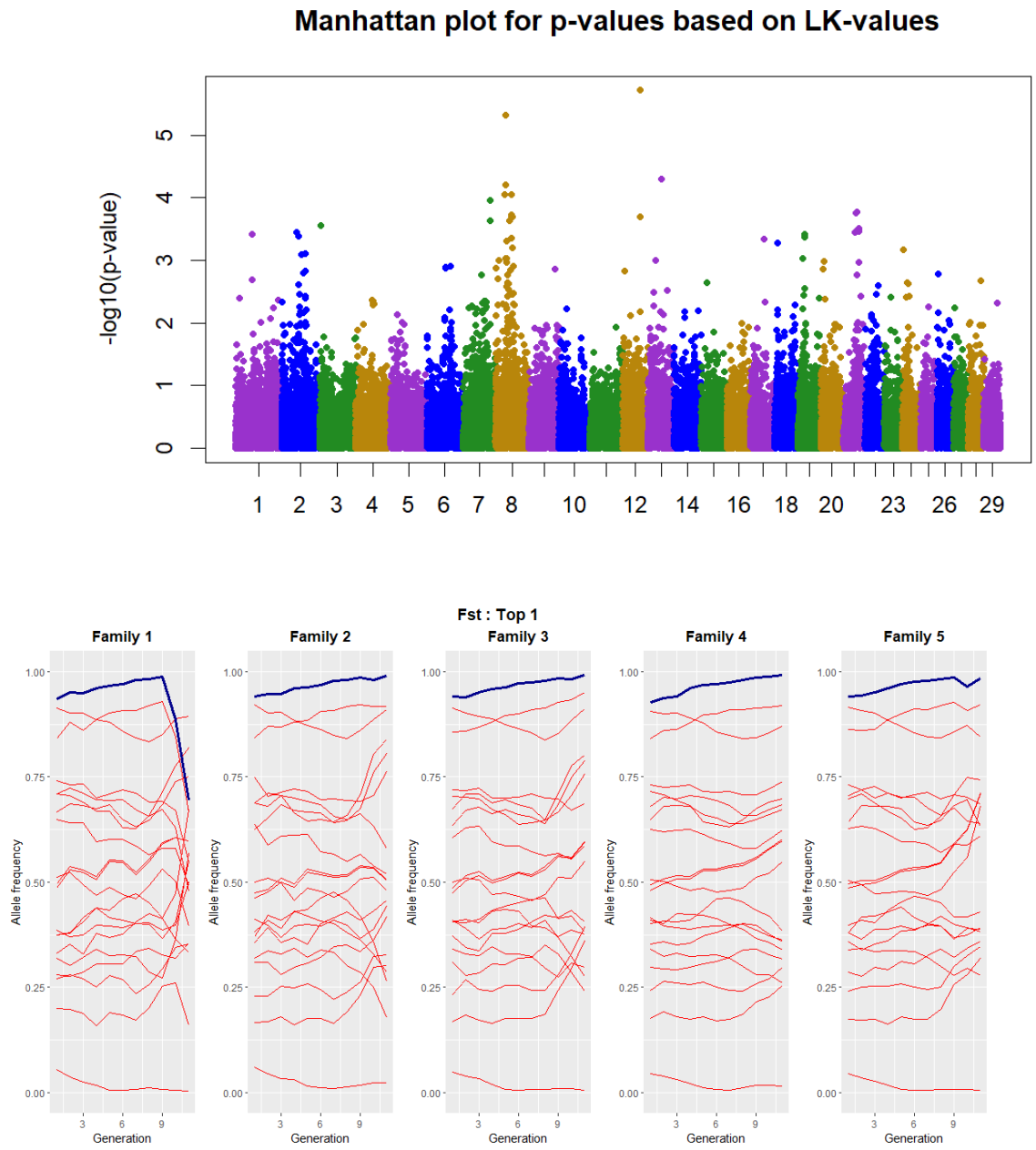


Figure 6.18: The Manhattan plot with p-values based on the Lewontin and Krakauer (LK) extension of the F_{st} test, and the allele frequency of the SNP with the highest LK-value (blue) and the surrounding 20 SNP markers (red) over generations within each family

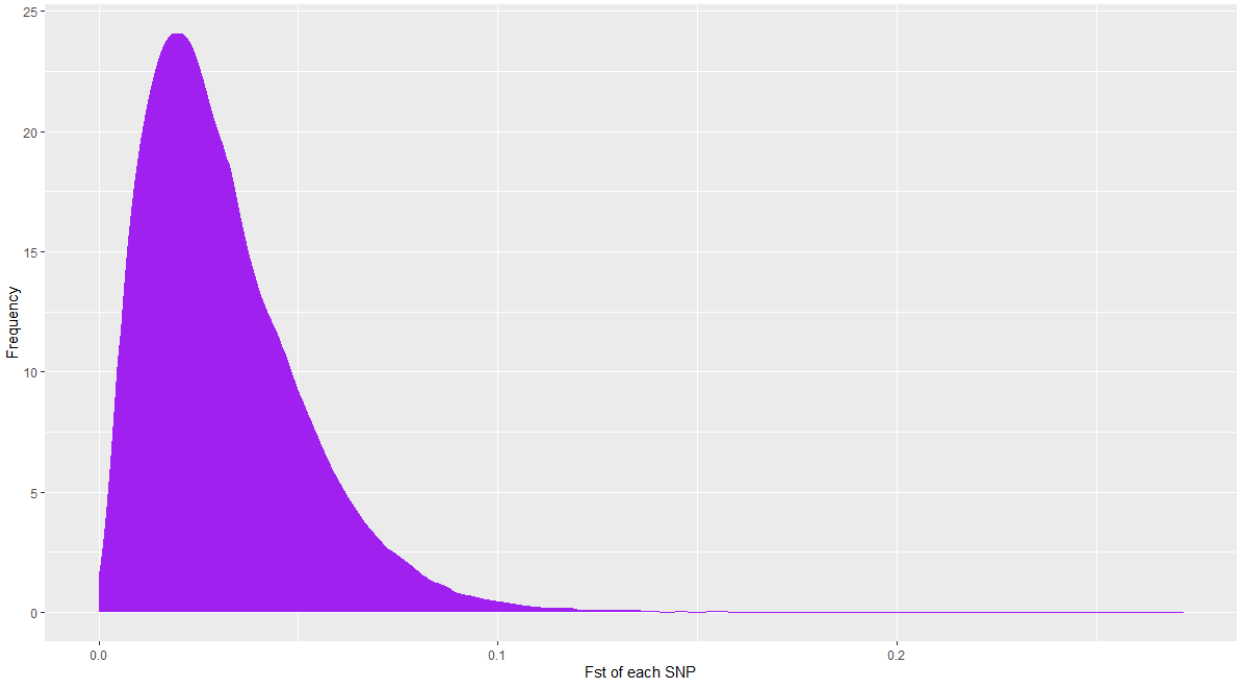


Figure 6.19: The distribution of F_{st} values of SNP markers across the five different clusters in G10.

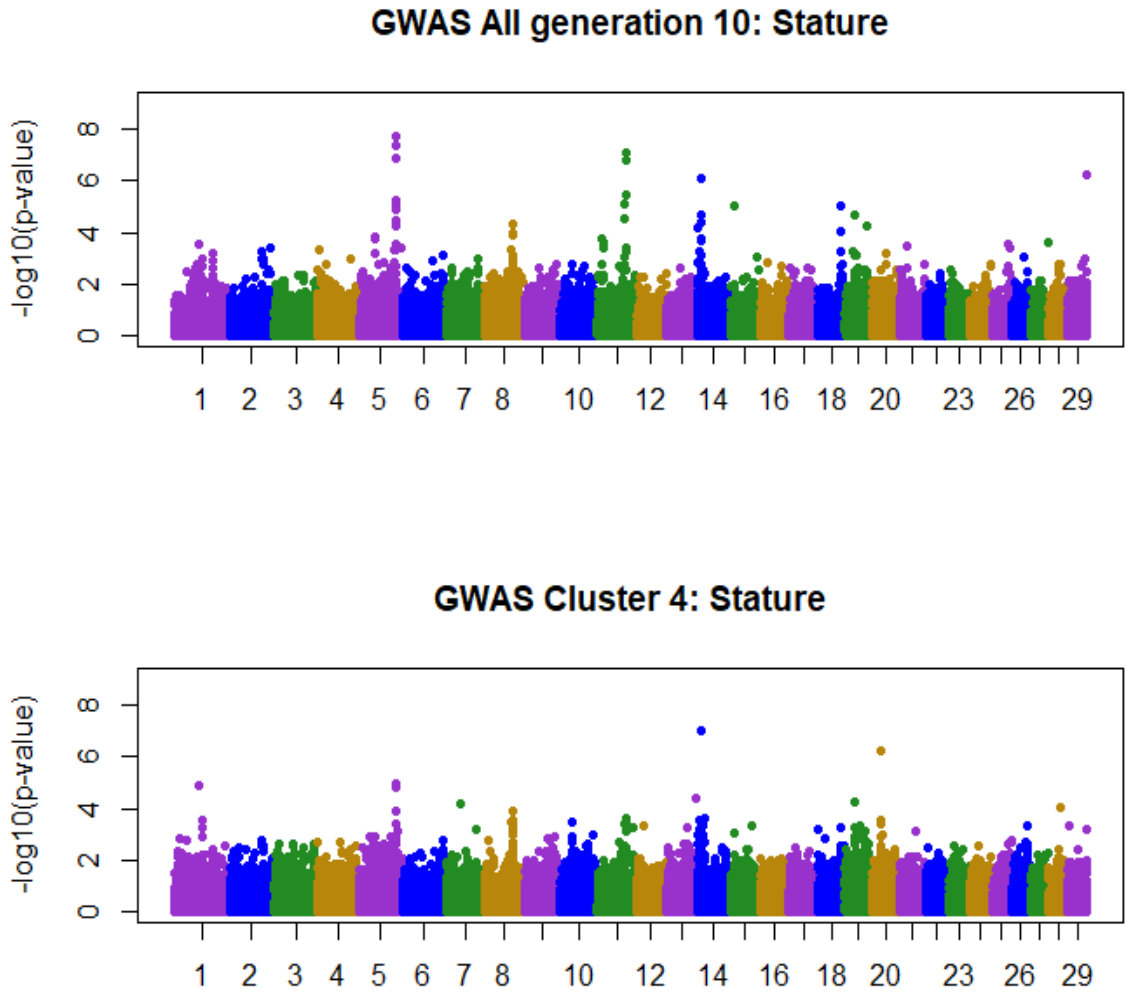


Figure 6.20: The Manhattan plot for the negative natural log of p-values for each SNP's contribution to stature when based on G10 of all clusters combined (ALL), or only cluster 4

CHAPTER 7

CONCLUSIONS

Multi-breed and across-breed evaluations with or without crossbred animals are desirable for their potential simplicity and pooling of resources to obtain accurate genomic predictions for all animals. Many approaches have failed to improve accuracy of prediction beyond what could be achieved with within-breed evaluations. In particular, across-breed predictions have been consistently poor, even with sophisticated techniques. Reasons for the lack of success or improvement include the inability to capture LD across breeds, different substitution effects, lack of allele segregation in all breeds, different allele frequencies, little representation of a specific breed, non-additive genetic factors, and inability to include all QTL in the evaluations. In general, treating breeds as separate traits, appear to be more successful, as well as approaches that take breed of origin of alleles into account. When a breed benefits, it is usually the smaller breed, although breeds that are numerically too small may have even lower accuracies when being combined with another. Marker density is crucial in multi-breed evaluations to capture LD better.

Results from our studies confirm that all breeds must be present in the reference population to at least obtain accuracies similar to within-breed evaluations. No benefits were found when combining breeds in a single, joint genomic relationship matrix. If marker density was low, the accuracy of multi-breed evaluations produced lower accuracies than within-breed evaluations, even when all breeds were in the reference population. Applying our non-shared SNP approach in compiling the genomic relationship matrix, prevented a reduction in accuracy,

but did not improve it. This is an advantage, as it can allow the use of fewer markers selected for their importance in each separate breed as non-shared SNP markers, while using those important in all breeds as shared SNP markers.

The use of crossbred animals is becoming increasingly popular in the dairy industry, which is known for their large, within-breed evaluations. Low imputation accuracy has long been a reason for their exclusion in genomic evaluations. Including component pure breeds alongside crossbred genotyped can deliver imputation accuracies high enough to include in evaluations. Approaches that have shown to be successful include the use of breed proportions along with pure-bred specific SNP effects, or taking breed origin of allele into account. Results from our studies show that the accuracy of prediction for crossbred animals were similar, regardless of whether only one pure breed, both pure breeds, equal proportions of each pure breed and crossbreds, or only crossbred animals were used as reference. Using Jersey as reference was slightly better than using only Holstein. Using breed proportions with breed-specific SNP effects resulted in the lowest accuracies. Inflation was best if an equal proportion of each pure breed and crossbreds were used as reference. Thus, crossbred animals can be included in the reference population without accounting for breed proportion. Another recent study also found that using breed-specific effects along with breed proportions was less successful than simply combining all animals, or combining all while also accounting for breed origin of allele.

Even within a single breed, distinct sub-populations can exist, either across-country or within. Animals with similar genetic merit can still be genetically different. This is mostly due to genetic redundancy, pleiotropy, and epistasis. Results from our study found that different sub-populations have evolved differently over time due to these factors. Genetic redundancy is advantageous to maintain genetic diversity over time, even within a population that has been

strongly selected for production. This could allow even greater genetic improvement without great losses of diversity. Further work is required to establish the best way to apply the identification of distinct groups in mating program.