THE IMPACT OF SPURIOUS MARKER ASSOCIATIONS ON GENOMIC SELECTION

by

ASHLEY LING

(Under the Direction of Romdhane Rekaya)

ABSTRACT

High-density and sequence genotype data was expected to improve the accuracy of genomic predictions in commercial animal genomic improvement programs by increasing the capture of genetic variation, however realized results have found only modest gains in accuracy relative to low- to mid-density SNP chip panels. This has raised a question of how much of this phenomenon can be attributed to lower-density SNP panels already sufficiently capturing segregation patterns of causal variants due to the strong linkage disequilibrium structure in these populations or if highly informative variants in high-density and sequence data are obfuscated by the dimensionality of the data.

This work employs a simulated data design to evaluate the impact of biallelic SNP markers with low information content regarding segregation of causal variants on genomic predictions and investigate strategies for distinguishing spurious, inconsistent trait associations among these markers from true associations driven by linkage between SNPs and causative variants. It is shown that the dimensionality of high-density and sequence data will inevitably result in false SNP-trait associations that lead to overfitting

of genomic models and lower quality genomic predictions. Furthermore, efforts to identify the most informative and biologically-relevant SNP markers by genome-wide association studies are impeded by the presence of these spurious associations. It may be possible to identify such spurious associations by comparing multiple criteria that measure trait-specific SNP relevance and eliminating those markers with inconsistent signals of relevance. A fuzzy inference system approach is evaluated for its potential to aggregate two SNP preselection criteria, F_{ST} scores and p-values, into a single composite score for preselection. Though gains in genomic prediction accuracy using this strategy were moderate compared to results for preselection based on the individual input criteria, it is demonstrated that there is potential for such an approach to consolidate information on SNP markers from multiple sources to better distinguish true from spurious trait associations.

INDEX WORDS: SNP prioritization, SNP preselection, high-density genotypes, sequence genotypes, genomic prediction, fuzzy inference systems

THE IMPACT OF SPURIOUS MARKER ASSOCIATIONS ON GENOMIC SELECTION

by

ASHLEY LING

BSA, University of Georgia, 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

2022

© 2022

Ashley Ling

All Rights Reserved

THE IMPACT OF SPURIOUS MARKER ASSOCIATIONS ON GENOMIC SELECTION

by

ASHLEY LING

Major Professor:	Romdhane Rekaya
Committee:	Samuel E. Aggrey
	El Hamidi Hay
	Ignacy Misztal
	Daniela Lourenco

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2022

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Dr. Romdhane Rekaya, for all the knowledge, guidance, and opportunities over these past several years; for always providing an environment where I was encouraged and supported in pursuing my own ideas; for all the understanding when roadblocks were met; and for the many pieces of wisdom shared to prepare for the post-PhD journey. I couldn't have asked for a better mentor.

I would also like to thank the other members of my committee, Dr. Samuel Aggrey, Dr. El Hamidi Hay, Dr. Ignacy Misztal, and Dr. Daniela Lourenco. Each of you have contributed to this dissertation and my professional development in unique and invaluable ways and I am incredibly fortunate for that.

There is a long list of professors, researchers, postdocs, visitors, and students that I've met and learned from during my time at UGA and I appreciate the expertise and conversations shared with all of them. I would especially like to thank Sungbong Jang, Fernando Bussiman, Ling-Yun Chang, Taylor McWhorter, Andre Garcia, Natalia Leite, Fiona Guinan, Amanda Maiorano, and Jiayi Qu for being my dear friends over the years, getting this far would have been a lot less interesting without you all.

Lastly, to my family, who has supported me through all my academic endeavors, but especially to my mom, Michelle Ling, I can't imagine having gotten this far without you cheering me on.

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS iv

LIST OF TABLES vii

LIST OF FIGURES ix

CHAPTER

1 INTRODUCTION 1

2 LITERATURE REVIEW 4

References 15

3 DISSECTION OF THE IMPACT OF PRIORITIZED QTL-LINKED AND -UNLINKED SNP MARKERS ON THE ACCURACY OF GENOMIC SELECTION 19

Abstract 20

Introduction 21

Materials and Methods 26

Results 32

Discussion 38

Conclusions 44

References 45

4 FUZZY LOGIC AS A STRATEGY FOR COMBINING MARKER STATISTICS TO OPTIMIZE PRESELECTION OF HIGH-DENSITY AND SEQUENCE GENOTYPE DATA 62
Abstract 63
Introduction 63
Materials and Methods 66
Results 72
Discussion 76
Conclusion 79
References 80
5 AN ALGORITHM FOR FINE-TUNING OF FUZZY INFERENCE SYSTEM PARAMETERS FOR VARIANT PRESELECTION AND PRIORITIZATION IN GENOMIC SELECTION 96
Abstract 97
Introduction 98
Materials and Methods 100
Results 103
Discussion 105
Conclusion 106
References 107
6 CONCLUSIONS 116

LIST OF TABLES

Page

Table 3.1: Accuracy of genomic predictions under varying number of random-, F_{ST}-, or estimated effect-based preselected markers 49

Table 3.2: Overlap (%) between random-, F_{ST}-, or effect-preselected marker subsets and G_2 SNPs 50

Table 3.3: Proportion of total GV[a] explained by random, effect, and F_{ST}-preselected markers 51

Table 3.4: Correlations between centered and non-centered genomic relationships with QTL relationships for different sets of markers[a] 52

Table 3.5: Correlations between non-centered and centered genomic and QTL relationships for varying numbers of F_{ST}-preselected markers 53

Table 3.6: Correlations between non-centered and centered genomic and QTL relationships for varying numbers of effect-preselected markers 54

Table 3.7: Accuracy after exclusion of different subsets of LQ28 markers from construction of the genomic relationship matrix 55

Table 3.8: Accuracy and percent of genetic variance explained by F_{ST}- and effect-preselected subsets under a simulation design with 200 QTL across all 30 chromosomes 56

Table 4.1: Fuzzy rules and corresponding rule parameters 82

Table 4.2: Simulation parameters for QMSim 83

Table 4.3: The number of unique SNPs and proportion of unique SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 5k causative variants 84
Table 4.4: The number of unique SNPs and proportion of unique SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 1k causative variants 85
Table 4.5: The number of unique SNPs and proportion of unique SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 300 causative variants 86
Table 4.6: The number of shared SNPs and proportion of shared SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 5k causative variants 87
Table 4.7: The number of shared SNPs and proportion of shared SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 1k causative variants 88
Table 4.8: The number of shared SNPs and proportion of shared SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 300 causative variants 89
Table 4.9: The number of SNPs unique to one criteria or shared across two or more criteria for a genetic architecture of 1k causative variants distributed across all chromosomes 90
Table 5.1: Simulation parameters for QMSim.....109

LIST OF FIGURES

Page

Figure 3.1: A general description of the simulation and workflow 57

Figure 3.2: Characterization of the modelling of QTL Mendelian Sampling (MS) using all, HQ2, and LQ28 markers 58

Figure 3.3: Characterization of the modelling of QTL Mendelian Sampling (MS) based on F_{ST}- and estimated-effects-preselected markers 59

Figure 3.4: Errors in the estimation of QTL Mendelian Sampling 60

Figure 3.5: Regression of F_{ST} scores on the absolute estimated effect for a) HQ2 and b) LQ28 markers 61

Figure 4.1: Membership functions for a) F_{ST} scores and b) $-\log_{10}$(p-value) 91

Figure 4.2: The fuzzy output space for the observed range of F_{ST} scores and $-\log_{10}$(p-value) 92

Figure 4.3: Proportion of SNPs linked with causative variants selected by random, F_{ST} scores, p-values, or fuzzy scores 93

Figure 4.4: Trends in accuracy, dispersion, and variance of EBVs for random, F_{ST} score, p-value, or fuzzy score preselection when only 14 of 29 chromosomes harbor causative variants 94

Figure 4.5: Trends in accuracy, dispersion, and variance of EBVs for random, F_{ST} score, p-value, or fuzzy score preselection when all chromosomes harbor causative variants 95

Figure 5.1: A comparison of the distributions of F_{ST} scores and p-values calculated based on true and randomized phenotypes 110

Figure 5.2: Fuzzy output spaces and selected SNPs for FIS parameter optimization based on difference of means, signal noise ratio (SNR), and T-Test statistics.....111

Figure 5.3: The trend in accuracies for preselection based on difference-of-mean-optimized fuzzy scores compared to preselection based on random, F_{ST} scores, p-values, or general fuzzy scores for 6 scenarios varying across heritability and sampling distribution for allelic effects 112

Figure 5.4: The trend in accuracies for preselection based on SNR-optimized fuzzy scores compared to preselection based on random, F_{ST} scores, p-values, or general fuzzy scores for 6 scenarios varying across heritability and sampling distribution for allelic effects 113

Figure 5.5: The trend in accuracies for preselection based on T-Test-optimized fuzzy scores compared to preselection based on random, F_{ST} scores, p-values, or general fuzzy scores for 6 scenarios varying across heritability and sampling distribution for allelic effects 114

Figure 5.6: Comparison of the trend in accuracies for preselection based on different FIS parameter optimization strategies (difference of means, SNR, or T-Test) for 6 scenarios varying across heritability and sampling distribution for allelic effects.... 115

CHAPTER 1

INTRODUCTION

The availability of high-density SNP panels and whole genome sequence data have raised the expectation of animal breeders and geneticists to better understand the genetic basis of complex traits and more importantly to improve the accuracy of genomic selection (GS). Unfortunately, some of those hopes have faded quickly as several studies reported no significant increase in accuracy using HD or sequence data compared to current low density panels. These results have created a certain level of skepticism about the utility of HD and sequence data in GS. However, fundamental questions have to be answered before any final decision. Is the lack of improvement in accuracy is due to 1) the redundancy of HD and sequence data compared to low density panels, 2) the structure, depth and limitations of existing data used to implement GS; and 3) the limitations of current methods used in GS. HD and sequence data dramatically increase the number of variants to be included in the association model. However, these variants, at least those with common minor allele frequencies are at a high linkage disequilibrium (LD) between themselves which will increase co-linearity and ultimately results in a reduction in statistical power. Although a similar increase of LD between markers and QTL is expected and it is positive for the association analysis, it is likely that the gain from increased LD will not offset the loss from the reduction in statistical power. This could be the case because the increase in LD between marker-QTL comes with the cost of splitting the effect of each QTL (often small) between an increasing number of

markers making most of them undetectable in the association analyses. This process will be magnified if the increase in the number of variants is not accompanied by a similar increase in the number of genotyped and phenotyped animals. Thus, even for populations with small effective population size there will still be amble genetic variance that could be captured much better using HD or sequence data than low density panels. The majority of livestock species are under heavy artificial selection resulting in well and highly related populations with very small effective population size. Thus, large segments of the genome are shared between the population members and could be effectively tracked by low density panels. Such argument is in general valid. However, several favorable de novo mutations could reach common level MAF in few generations given the high selection pressure and the extensive use of artificial insemination leading to sufficient genetic variation between families and even branches within families. Such diversity cannot be explained by the small number of large genomic segments. Using current methods, an increase in SNP density, after a certain threshold, seems to not affect the quality of the estimation of the observed relationship matrix G. Because HD and sequence data did not improve the quality of G in any significant way, accuracy of genomic selection did not improve using mixed linear approaches.

Similarly, including all the variants in the association model using regression approaches will severely aggravate the small n large p problem and substantially reduce the statistical power. Current marker prioritization methods based on variable selection do not have the statistical power to discriminate between millions or even hundreds of thousands of variants. Consequently, as the marker density increases, the performance of these

methods decays resulting in no improvement of accuracy. Prioritization based on external biological information suffers from the limitation and variability of the latter.

It is clear that the inclusion of all genotyped variants either in the association model for regression methods or in the calculation of G is not beneficial. Furthermore, current SNP/variant prioritization methods had limited success due to lack of sufficient statistical power for Bayesian variable selection approaches or the unavailability or/and low quality (noise to signal ratio) of external biological data for approaches requiring prior information. Consequently, innovative methods to reduce the dimensionality of the genomic information are needed to harness the benefit of HD and sequence data.

The overall aim of this dissertation work is the development of methodologies and implementation procedures for the for the reduction of the dimensionality of genomic data through the testing of individual marker prioritization criteria and the development of soft classification procedure based on fuzzy logic to integrate multiple individual preselection criteria into a global multicomponent prioritization approach.

CHAPTER 2

LITERATURE REVIEW

The field of animal breeding and genetics is primarily concerned with two general goals: 1) determining how to best leverage information recorded on a population to more accurately predict its future, and 2) how to use predictions for individual genetic merit to realize the most optimal future. It is natural to expect that animals possessing some advantageous quality are likely to produce progeny sharing that quality as well, and that breeding such animals together should yield more favorable results than breeding at random. While this natural intuition resulted in immense genetic progress for production traits in animal species prior to any formalization of animal breeding theory, optimization of animal improvement programs is complicated by such factors as Mendelian sampling of parental alleles and the potential for negative effects of inbreeding as harmful recessive alleles accumulate in a population, among others.

The past approximately 100 years have seen continual progress in formalization of animal breeding theory, leading to better prediction of the top selection candidates in a population, faster rates of genetic progress, and stronger accord concerning selection goals among producers.

The foundation for modern animal breeding and quantitative genetics was laid largely by the landmark papers Fisher (1919) and Wright (1922). Fisher (1919) provided an explanation for how discrete genetic units could act collectively to generate continuous variation at the macroscopic scale, what is called today The Infinitesimal Model. This

effectively consolidated two disjoint theories of genetics and inheritance that existed at the time: the school of thought pioneered by Gregor Mendel that was concerned with how segregating alleles were passed on to progeny to yield discrete variation in phenotypes, and what was then known as biometrics, which was concerned with continuous variation in traits and seemed inherently at odds with the concept of disjoint units of inheritance. Mendel had observed through experiments on *Pisum sativum*, the garden pea, that the expression of some phenotypes could be explained by the inheritance of discrete alleles from parents to progeny. These experiments provided strong evidence of a genetic basis for certain traits that had a straightforward categorical classification. However, this did not provide a clear explanation for the genetic basis underlying phenotypes that were expressed on a continuous scale (e.g. height). Fisher (1919) proposed that consistency between these two schools of thought could be established by way of a large number of segregating genetic units of a discrete Mendelian nature, each individually contributing a small effect to a particular trait, acting collectively to generate the continuous variation observed for many traits.

In this landmark paper, Fisher additionally laid the groundwork for the analysis of variance (ANOVA) as well as coined the term ‘variance’. Fisher proposed that the total phenotypic variance could be partitioned into a portion attributed to genetic factors and another attributed to environmental factors.

While Fisher (1919) laid the groundwork for understanding how the genetic makeup of an individual could generate the continuous variation manifested in many economically important traits, it was Wright (1922) that presented a method for calculating inbreeding and relationships via path coefficients, which were termed the coefficients of

inbreeding and relationships, respectively. The coefficient of inbreeding was formulated to reflect the expected proportion of homozygosity at segregating sites. Wright chose this metric as a measure of inbreeding based on the argument that inbreeding depression was a consequence of the accumulation of harmful recessive alleles in a homozygous state and it was expected that inbreeding depression would be proportional to the rate of homozygosity.

Another method of quantifying the relationship between two individuals was presented by Malécot (1948) as the probability that individuals sharing a common ancestor would inherit the same allele identical by descent (IBD). This quantity is commonly referred to as the co-ancestry coefficient.

Malécot's coancestry coefficient has notably been used since to assemble the numerator relationship matrix based on pedigree information, which is essential for capturing connectedness within a population to best understand how similarity between the performance of relatives and similarity at the genotypic scale are connected. This connectedness based on expected (co)variances allows the data collected on one individual to influence predictions of all their relatives, ultimately yielding better predictions than what could be achieved based only on an individual's own records or records of only immediate relatives (e.g., parents, siblings, and progeny).

The inverse of the numerator relationship matrix is an important component of the mixed model equations (MME) (Henderson 1950, Henderson 1973, Henderson 1975), which allow simultaneous estimation of the fixed and random effects by maximizing the likelihood of their joint distribution. An important advantage of Henderson's MME compared to previously available methods is the ability to predict random effects without

necessitating inversion of the variance-covariance matrix for observations, which is generally extremely dense and computationally intensive to invert. Henderson presented an early derivation of these equations based on maximum likelihood estimation in 1950, but did not present a formal proof of their unbiasedness until 1973. Without inclusion of the NRM inverse, the off-diagonals of the left-hand side of the equations pertaining to animal effects will be zero and connectedness in the population will not be captured; this inverse captures the variance-covariance structure between relatives based on expected proportion of shared alleles.

Prior to 1976, one complication of solving the mixed models was the necessity of using $\boldsymbol{A}^{-1}$; the dimension of the NMR is typically quite large, making direct inversion prohibitively expensive. Alternatively, both sides of the equations could be multiplied by the NMR to eliminate $\boldsymbol{A}^{-1}$, but doing so greatly reduces the efficiency of iterative methods typically used to solve the MME (Henderson 1976).

Obtaining the inverse of the NRM was greatly simplified by Henderson (1976), which derived via LDL decomposition an algorithm for calculating $\boldsymbol{A}^{-1}$ directly. However, Henderson's method still relied on finding the diagonal elements of a matrix $\boldsymbol{L}$ such that $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}'$ in the case of an inbred population and doing so required prohibitively large amounts of computer memory for the time. Later the same year, Quaas (1976) determined a method that greatly reduced the amount of computer memory needed for calculating the diagonal elements of $\boldsymbol{L}$.

A further advantage of building $\boldsymbol{A}^{-1}$ directly besides avoidance of inversion was the fact that $\boldsymbol{A}^{-1}$ is more sparse than $\boldsymbol{A}$ due to the only non-zero off-diagonal elements for a given animal being those pertaining to its parents, progeny, and mates.

An inherent limitation of the NRM is that it captures only expected relationships. Parental gametes are a random sample of 50% of the paired chromosomes present in the parental genome, and so not all progeny will inherit the same alleles. Naturally, this leads to variation around the expected relationship of relatives, a component referred to as the Mendelian sampling variance. Given pedigree information alone, the Mendelian sampling component of a relationship cannot be elucidated and variation in the performance and genetic merit of siblings due to inheritance of different alleles of causative variants from a shared parent(s) cannot be measured until records on self and progeny are available.

Early formulations of relationship matrices based on genotype information for use in Henderson's BLUP equations were proposed by Fernando and Grossman (1989) and Nejati-Javaremi et al. (1997). Both strategies aimed to better model shared inheritance of causative variants indirectly by way of linked genomic markers compared to what could be accomplished by pedigree-estimated relationships alone, though they notably differed in how they defined allelic similarity between individuals. Fernando and Grossman (1989) modeled the variance-covariance structure based on known pedigree, identity-by-descent (IBD) of alleles, and recombination rates between markers and QTL. The use of only IBD information meant that off-diagonals of the matrix corresponding to individuals without a known relationship in the pedigree would be zero. Nejati-Javaremi et al. (1997) differed from this approach by accounting for IBS of all alleles shared between individuals regardless of pedigree information. This approach is more similar to current standards for the construction of genomic relationship matrices (VanRaden 2008) and has the advantage of capturing missing or incorrect pedigree information, though the resulting matrix will be far more dense than that proposed by Fernando and Grossman (1989). They additionally

differ in that Fernando and Grossman (1989) estimates paternal and maternal allelic effects for every animal in the model, making the dimension of the resulting square matrix 2*nm* for *n* animals and *m* markers while Nejati-Javaremi et al. (1997) estimates a single random effect per animal. The amount and type of genomic information available for evaluations has changed considerably since Fernando and Grossman (1989) was published and this paper clearly represents a transition from linkage analysis and marker-assisted selection to the genomic BLUP (GBLUP) models that are routinely used today.

As genotyping technology developed and corresponding costs fell, focus further grew on how such data would be incorporated into routine evaluations for production animals. It was anticipated that soon markers in the order of tens of thousands would become available for prediction of genomic breeding values. Meuwissen et al. (2001) compared the performance of least squares estimation, BLUP, and Bayesian models for estimation of marker haplotype effects in simulated data. The prediction accuracies of the corresponding direct genomic values based on BLUP and Bayesian approaches were approximately 83 to 112% higher than the expected accuracy using pedigree information alone for a heritability of 0.5. These results generated optimism around the potential for genomics in animal genetic improvement programs, but were inflated relative to realized results (VanRaden et al. 2009) likely due to simulation design, fitting of multiallelic marker haplotypes rather than biallelic SNPs in the models, and a genetic architecture with one causative variant explaining more than 40% of the genetic variance.

Despite limitations due to computation and access to genotype data, there were some efforts in the following years to further explore potential genetic gains and which models were best suited when using genome-wide markers in calculations of breeding

values (Fernando et al. 2007, Habier et al. 2007). The results of these studies illustrated the clear influence of genetic architecture on the performance of the different models. BayesB (Meuwissen et al. 2001) tended to outperform alternatives when the genetic architecture was dominated by a small number of loci due to its intrinsic prioritization of markers that improve model fit. Ridge regression BLUP (RRBLUP, often also called SNPBLUP because SNP markers are fitted as predictors) tended to yield the most accurate predictions as the complexity of the genetic architecture approached the infinitesimal model because it penalizes large estimates to control inflation caused by multicollinearity between predictors and leads to estimation of a non-zero effect for all predictors (Hoerl and Kennard 1970). These results have been consistently supported by more recent simulation studies (Daetwyler et al. 2010, Clark et al. 2011) well as real data where the trait architecture is known (Ober et al. 2012, Wolc et al. 2016) .

A formulation of the genomic relationship matrix ($\boldsymbol{G}$) with dimensions $n \times n$ was proposed by Habier et al. (2007) and VanRaden (2008) as,

$$\boldsymbol{G} = \frac{\boldsymbol{Z}\boldsymbol{Z}'}{2\sum_{i=1}^{m} p_i(1-p_i)}$$

Where $\boldsymbol{Z}$ is a $n \times m$ matrix of genotypes, m is the number of SNPs, n is the number of individuals with genotypes, and p_i is the minor allele frequency of the ith SNP. In contrast to the $\boldsymbol{G}$ proposed by Nejati-Javaremi et al. (1997), this formulation is scaled to be equivalent to the NRM and its inverse replaces $\boldsymbol{A}^{-1}$ in Henderson's MME to estimate random animal effects directly. It can be shown that GBLUP using this formulation of $\boldsymbol{G}$ is equivalent to SNPBLUP/RRBLUP under the assumption that $var(\boldsymbol{u}) = \frac{1}{2\sum_{i=1}^{m} p_i(1-p_i)}$, where $\boldsymbol{u}$ is the vector of SNP effects (Habier et al., 2007).

The first routine evaluation using genome-wide SNP data was implemented in Holstein in 2008 using approximately 40k SNPs post-quality control to calculate genomic relationships as in VanRaden (2008) and VanRaden et al. (2009) summarized the outcome of this evaluation for 3,576 training and 1,759 validation bulls. An increase of approximately 23% in reliabilities relative to predictions from parent averages was achieved. These gains are equivalent to adding 11 daughter records per sire to the evaluation and illustrates that the benefit of genomic information is not just an increase in accuracy but also a shortening of the generation interval as genotypes could be collected at birth.

Shortening of the generation interval leads to more rapid genetic progress, a trend that was summarized for the US Holstein population by Garcia-Ruiz et al. (2016). A sharp decrease in the generation interval, particularly for sires of bulls, was shown to occur beginning around 2009 when the first genomic evaluation was implemented in US Holstein and the rate of genetic gain for several yield traits increased considerably around the same time.

An important drawback of the initial formulation of GBLUP with $\boldsymbol{A^{-1}}$ simply replaced by $\boldsymbol{G^{-1}}$ (VanRaden 2008) is that contributions from records to direct predictions of animal effects are limited to those for genotyped individuals. Integrating information from records for ungenotyped individuals required blending DGVs from GBLUP with parent averages from a pedigree BLUP model akin to a selection index approach. Aguilar et al. (2010) and Christensen and Lund (2010) simultaneously published equivalent derivations for construction of a matrix, $\boldsymbol{H}$, that blends information from expected relationships in $\boldsymbol{A}$ and genomic relationships in $\boldsymbol{G}$.

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

where $\boldsymbol{A}_{22}^{-1}$ is the inverse of the block of $\boldsymbol{A}$ corresponding to genotyped individuals. Though not trivial in its derivation, $\boldsymbol{H}^{-1}$ is quite elegant in the simplicity and intuition of its formulation, as it effectively removes double-counting of expected relationships in the summation of $\boldsymbol{A}^{-1}$ and $\boldsymbol{G}^{-1}$. Use of $\boldsymbol{H}^{-1}$ in the MME eliminates the necessity of combining parent averages with DGV to yield genomic EBVs.

The performance of genomic selection has been evaluated in several additional breeds and species (Christensen et al. 2012, Lourenco et al. 2015, Wolc et al. 2016, Garcia et al. 2018) as the cost of genotyping and development of accurate imputation strategies (Berry and Kearney 2011, Sargolzaei et al. 2014, van Binsbergen et al. 2014) has increased the dimensionality of genotype data available in terms of both SNPs and individuals with genotypes. The prevalence of genotyping and size of genomic databases for several breeds and species have increased enough that the dimensions of $\boldsymbol{G}$ using all available genotypes have become too intractably large to allow inversion in a reasonable amount of time. To address this issue, algorithms that avoid inversion of $\boldsymbol{G}$ entirely (Fernando et al. 2014) or require inversion of only a subset of $\boldsymbol{G}$ (Pocrnic et al. 2016) have been developed to take advantage of the increased dimensionality of genotype data while controlling computational costs.

Efforts such as the 1000 Bull Genomes Project (Daetwyler et al. 2014, Hayes and Daetwyler 2019), which has identified more than 86 million segregating variants in domesticated cattle populations based on the genomes of 2,703 whole cattle genomes from 121 breeds, have played a key role in providing a reference for construction of higher-density SNP chips and imputation from low- to mid- or high-density SNP chips. This

increase in the dimensionality of SNP markers and other variants (e.g., indels) has the potential to benefit genome-wide association studies (Mao et al. 2016, Ring et al. 2019, Zhang et al. 2020) and efforts to analyze and maintain genetic diversity in populations with low effective population size (Eynard et al. 2016).

There were also expectations that high-density SNPs and sequence data might increase the accuracy of genomic predictions by increasing LD between markers and causative variants or by providing genotypes of the causative variants themselves (Meuwissen and Goddard 2010, Druet et al. 2013). However, practical results have not met such expectations (van Binsbergen et al. 2015, Heidaritabar et al. 2016).

The failure of high-density and sequence genotypes to meet expectations in genomic selection applications can be easily explained. Models that use hundreds of thousands or millions of variants as predictors will be highly overparameterized and the influence of any individual variant on calculation of relationships and predictions will be negligible. Chang et al. (2019) demonstrated that the elements of $\boldsymbol{G}$ will converge to relatively stable values as markers reach high densities and inclusion of additional highly informative variants will not appreciably change relationships. Considering also the computational cost of inclusion of such large numbers of variants in calculation of genomic predictions, there has so far been little justification for use of the full genotypes rather than lower-density SNP chips.

To take full advantage of the information contained in high-density and sequence genotype data, variants with little or redundant information on causative variants should be identified and excluded from analyses. Effort to identify quantitative trait loci (QTL) and causative variants have been complicated largely by two factors: 1) the small effect size of

the vast majority of causative variants that are difficult to distinguish from random noise, and 2) the long stretches of LD in commercial production animal populations that spread association signals across large genomic regions.

Prior to the widespread availability of SNP markers and implementation of genomic selection in commercial animal populations, analysis of genotype data primarily involved QTL mapping experiments using variants that tend to be highly unique to families such as microsatellite markers (Moore et al. 1994, Ron et al. 1994, Georges et al. 1995). An expectation from these studies was that a large number of QTL could be identified and markers that segregated closely with these QTL could be used in marker-assisted selection (Beckmann and Soller 1983) to identify and select animals that had inherited favorable QTL alleles (Meuwissen and Goddard 1995). Such experiments depended on measuring significant differences in the phenotypes of individuals having alternative homozygote status of markers of interest. As stated previously, this is complicated by small effect sizes that become obscured by statistical noise and long-range LD in commercial livestock populations that make fine-tuning resolution of QTL location in the genome difficult even when the effect size is large enough to detect (Mackay 2001). Ultimately, relatively few QTL were successfully identified, particularly for lowly heritable traits such as reproduction traits (Khatkar et al. 2004), motivating use of genome-wide markers as SNP data became more available.

Efforts to identify causative variants using genome-wide marker data have faced similar complications, with the addition of the high dimensionality of SNP genotypes requiring strict multiple testing correction procedures to control false positive rates such as the Bonferroni correction. While high-density SNP genotypes are expected to vastly

increase the number of markers in high LD with causative variants and sequence data to provide the genotypes of causative variants themselves (Meuwissen and Goddard 2010, Druet et al. 2013), significance thresholds with sufficiently strict correction for multiple testing for such high dimensional data are unlikely to allow many, if any, causative variants to reach significance without training data sizes that will be prohibitively large for many breeds and species.

References

Aguilar, I., et al. (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): 743-752.38

Beckmann, J. S. and M. Soller (1983). "Restriction fragment length polymorphisms in genetic improvement: methodologies, mapping and costs." Theoretical and Applied Genetics **67**: 35-43.10

Berry, D. P. and J. F. Kearney (2011). "Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection." Animal **5**(8): 1162-1169.41

Chang, L. Y., et al. (2019). "A Weighted Genomic Relationship Matrix Based on Fixation Index (FST) Prioritized SNPs for Genomic Selection." Genes (Basel) **10**(11).103

Christensen, O. F. and M. S. Lund (2010). "Genomic prediction when some animals are not genotyped." Genetics Selection Evolution **42**(2).39

Christensen, O. F., et al. (2012). "Single-step methods for genomic evaluation in pigs." Animal **6**(10): 1565-1571.43

Clark, S. A., et al. (2011). "Different models of genetic variation and their effect on genomic evaluation." Genetics Selection Evolution **43**(18).102

Daetwyler, H. D., et al. (2014). "Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle." Nature Genetics **46**(8): 858-865.49

Daetwyler, H. D., et al. (2010). "The Impact of Genetic Architecture on Genome-Wide Evaluation Methods." Genetics **185**(3): 1021-1031.40

Druet, T., et al. (2013). "Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions." Heredity **112**(1): 39-47.50

Eynard, S. E., et al. (2016). "Whole-genome sequence data uncover loss of genetic diversity due to selection." Genetics Selection Evolution **48**(1).61

Fernando, R. L., et al. (2014). "A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses." Genetics Selection Evolution **46**(50).51

Fernando, R. L. and M. Grossman (1989). "Marker assisted selection using best linear unbiased prediction." Genetics Selection Evolution **21**: 467-477.12

Fernando, R. L., et al. (2007). "Genomic selection." Acta Agriculturae Scandinavica, Section A - Animal Science **57**(4): 192-195.28

Fisher, R. A. (1919). "The causes of human variability." Eugenics Review: 213-220.1

Fisher, R. A. (1919). "XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance." Transactions of the Royal Society of Edinburgh **52**(2): 399-433.100

Garcia-Ruiz, A., et al. (2016). "Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection." Proc Natl Acad Sci U S A **113**(28): E3995-4004.62

Garcia, A. L. S., et al. (2018). "Development of genomic predictions for harvest and carcass weight in channel catfish." Genet Sel Evol **50**(1): 66.74

Georges, M., et al. (1995). "Mapping Quantitative Trait Loci Controlling Milk Production in Dairy Cattle by Exploiting Progeny Testing." Genetics **139**: 907-920.15

Habier, D., et al. (2007). "The impact of genetic relationship information on genome-assisted breeding values." Genetics **177**(4): 2389-2397.29

Hayes, B. J. and H. D. Daetwyler (2019). "1000 Bull Genomes Project to Map Simple and Complex Genetic Traits in Cattle: Applications and Outcomes." Annual Review of Animal Biosciences **7**(1): 89-102.78

Heidaritabar, M., et al. (2016). "Accuracy of genomic prediction using imputed whole-genome sequence data in white layers." J Anim Breed Genet **133**(3): 167-179.63

Henderson, C. R. (1950). "Estimation of Genetic Parameters."3

Henderson, C. R. (1973). "Sire Evaluation and Genetic Trends."6

Henderson, C. R. (1975). "Best Linear Unbiased Estimation and Prediction under a Selection Model." Biometrics **31**(2): 423-447.7

Henderson, C. R. (1976). "A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values." Biometrics **32**(1): 69-83.8

Hoerl, A. E. and R. W. Kennard (1970). "Ridge Regression: Applications to Nonorthogonal Problems." Technometrics **12**(1): 69-82.5

Khatkar, M. S., et al. (2004). "Quantitative trait loci mapping in dairy cattle: review and meta-analysis." Genetics Selection Evolution **36**(2): 163-190.22

Lourenco, D. A. L., et al. (2015). "Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus." Journal of Animal Science **93**: 2653–2662.58

Mackay, T. F. C. (2001). "The Genetic Architecture of Complex Traits." Annual Review of Genetics **35**: 303-339.18

Malécot, G. (1948). Les mathématiques de l'hérédité, Masson et Cie.101

Mao, X., et al. (2016). "Genome-wide association studies of growth traits in three dairy cattle breeds using whole-genome sequence data." Journal of Animal Science **94**: 1426-1437.65

Meuwissen, T. and M. Goddard (2010). "Accurate prediction of genetic values for complex traits by whole-genome resequencing." Genetics **185**(2): 623-631.107

Meuwissen, T. H. E. and M. E. Goddard (1995). "The use of marker haplotypes in animal breeding schemes." Genetics Selection Evolution **28**: 161-176.16

Meuwissen, T. H. E., et al. (2001). "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." Genetics **157**: 1819–1829.19

Moore, S. S., et al. (1994). "Characterization of 65 bovine microsatellites." Mammalian Genome **5**: 84-90.13

Nejati-Javaremi, A., et al. (1997). "Effect of Total Allelic Relationship on Accuracy of Evaluation and Response to Selection." Journal of Animal Science **75**: 1738–1745.17

Ober, U., et al. (2012). "Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster." PLoS Genet **8**(5): e1002685.44

Pocrnic, I., et al. (2016). "The Dimensionality of Genomic Information and Its Effect on Genomic Prediction." Genetics **203**(1): 573-581.67

Quaas, R. L. (1976). "Computing the Diagonal Elements and Inverse of a Large Numerator Relationship Matrix." Biometrics **32**(4): 949-953.9

Ring, S. C., et al. (2019). "Variance components for bovine tuberculosis infection and multi-breed genome-wide association analysis using imputed whole genome sequence data." PLoS One **14**(2): e0212067.81

Ron, M., et al. (1994). "Mapping quantitative trait loci with DNA microsatellites in a commercial dairy cattle population." Animal Genetics **25**: 259-264.14

Sargolzaei, M., et al. (2014). "A new approach for efficient genotype imputation using information from relatives." BMC Genomics **15**(478).53

van Binsbergen, R., et al. (2014). "Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle." Genetics Selection Evolution **46**(41): 1-13.54

van Binsbergen, R., et al. (2015). "Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle." Genetics Selection Evolution **47**(1).60

VanRaden, P. M. (2008). "Efficient methods to compute genomic predictions." J Dairy Sci **91**(11): 4414-4423.35

VanRaden, P. M., et al. (2009). "Invited review: reliability of genomic predictions for North American Holstein bulls." J Dairy Sci **92**(1): 16-24.37

Wolc, A., et al. (2016). "Implementation of genomic selection in the poultry industry." Animal Frontiers **6**(1): 23-31.69

Wright, S. (1922). "Coefficients of inbreeding and relationship." The American Naturalist **56**: 330-338.2

Zhang, F., et al. (2020). "Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: I: feed efficiency and component traits." BMC Genomics **21**(1).90

CHAPTER 3

DISSECTION OF THE IMPACT OF PRIORITIZED QTL-LINKED AND -UNLINKED SNP MARKERS ON THE ACCURACY OF GENOMIC SELECTION[1]

[1] Ling, A.S., Hay, E.H., Aggrey, S.E. *et al.* Dissection of the impact of prioritized QTL-linked and -unlinked SNP markers on the accuracy of genomic selection[1]. *BMC Genom Data* **22,** 26 (2021). Reprinted here with permission of the publisher.

Abstract

Use of genomic information has resulted in an undeniable improvement in prediction accuracies and an increase in genetic gain in animal and plant genetic selection programs in spite of oversimplified assumptions about the true biological processes. Even for complex traits, a large portion of markers do not segregate with or effectively track genomic regions contributing to trait variation; yet it is not clear how genomic prediction accuracies are impacted by such potentially nonrelevant markers. In this study, a simulation was carried out to evaluate genomic predictions in the presence of markers unlinked with trait-relevant QTL. Further, we compared the ability of the population statistic F_{ST} and absolute estimated marker effect as preselection statistics to discriminate between linked and unlinked markers and the corresponding impact on accuracy. We found that the accuracy of genomic predictions decreased as the proportion of unlinked markers used to calculate the genomic relationships increased. Using all, only linked, and only unlinked marker sets yielded prediction accuracies of 0.62, 0.89, and 0.22, respectively. Furthermore, it was found that prediction accuracies are severely impacted by unlinked markers with large spurious associations. F_{ST}-preselected marker sets of 10k and larger yielded accuracies 8.97 to 17.91% higher than those achieved using preselection by absolute estimated marker effects, despite selecting 5.1 to 37.7% more unlinked markers and explaining 2.4 to 5.0% less of the genetic variance. This was attributed to false positives selected by absolute estimated marker effects having a larger spurious association with the trait of interest and more negative impact on predictions. The Pearson correlation between F_{ST} scores and absolute estimated marker effects was 0.77 and 0.27 among only linked and only unlinked markers, respectively. The sensitivity

of F_{ST} scores to detect truly linked markers is comparable to absolute estimated marker effects but the consistency between the two statistics regarding false positives is weak. Identification and exclusion of markers that have little to no relevance to the trait of interest may significantly increase genomic prediction accuracies. The population statistic F_{ST} presents an efficient and effective tool for preselection of trait-relevant markers.

Introduction

Whole-genome marker information has been successfully utilized through genomic selection (GS) in many livestock and plant genetic improvement programs for the prediction of genomic merit and has led to a significant increase in the rate of genetic gain in these species (Garcia-Ruiz et al. 2016). This has been partly a result of increased prediction accuracy for selection candidates, particularly for individuals with no phenotypes or progeny of their own (VanRaden et al. 2009). Such improvement in accuracy is due to a better modeling of the Mendelian sampling (MS) using genomic information compared to using only pedigree information.

Though millions of single nucleotide polymorphisms (SNPs) have been discovered in human (Genomes Project et al. 2015), livestock (Daetwyler et al. 2014), and plant (Rimbert et al. 2018) genomes, relatively high accuracies have been achieved using marker panels that utilize just a fraction of these markers (Chang et al. 2018, Chang et al. 2019). The falling costs of full genome sequencing and genotyping combined with more reference genomes and the availability of imputation algorithms have now allowed the regular use of high-density and sequence genotypes in genomic analyses.

It has been suggested that sequence data has the potential to significantly improve the accuracy of genomic predictions by increasing the linkage disequilibrium (LD)

between quantitative trait loci (QTL) and SNPs or even making available the genotypes of causal loci (Meuwissen and Goddard 2010, Druet et al. 2013, Hayes et al. 2014). Early simulation studies found optimistic potential for the use of sequence data in GS. Meuwissen and Goddard (2010) estimated that accuracies could be improved by more than 40% when using sequence data compared to low-density SNP panels, but concluded that this was likely due to the weak relationship structure of the training population and did not expect the same results in real livestock populations due to the long-ranging LD and strong family structures. Druet et al. (2013) found that accuracies could be increased by up to 28% using sequence data compared to the equivalent of a bovine 50k SNP chip when the trait was controlled by rare QTL; however, these gains were largely lost when the sequence genotypes were imputed, likely as a result of lower imputation accuracy of rare markers that would be most effective in tracking causal loci with low minor allele frequencies.

Most results from real data have found little to no improvement in accuracy using high-density and sequence data for genomic prediction (Ober et al. 2012, van Binsbergen et al. 2015, Heidaritabar et al. 2016, Zhang et al. 2018). This lack of improvement has in some cases been attributed to the fact that low- and moderate-density panels are sufficient to capture realized additive relationships across the whole genome. Furthermore, a marginal decline in accuracy with the increase in SNP density was observed in some cases (van Binsbergen et al. 2015, Zhang et al. 2018), which results in part from overparameterization of the model (Hickey et al. 2014). This is not a surprising occurrence, as a disproportional increase in the number of unknown parameters in the

association model relative to the number of observations available in the training set will lead to the well-known small n large p problem.

Models that intrinsically perform variable selection (e.g., BayesB, LASSO, and elastic-net) have been proposed as a way to reduce the dimensionality of genomic data and alleviate the issues associated with the small n large p problem. Daetwyler et al. (2010) showed using a simulation scheme that BayesB (Meuwissen et al. 2001) tends to have an advantage compared to GBLUP when the number of causal loci is less than the estimated number of independent chromosome segments.

In comparisons between GBLUP and BayesB using real data, the latter tends to yield superior results when the trait of interest is under the influence of at least one major gene, such as DGAT1 for fat and protein content in dairy cattle (Croiseau et al. 2011). While BayesB tends to yield predictions that are at least as accurate as GBLUP in most practical analyses, it is computationally demanding, particularly as the number of predictors included in the model increases. Principal component analyses can dramatically reduce the dimensionality of the association model without a substantial loss in the portion of explained genetic variance; however, the estimated effects are linear combinations of the original predictors, thus complicating their interpretation. In general, the gains from using variable selection methods have been modest to nonexistent.

While the presence of causal variant genotypes in sequence information might be expected to give variable selection methods an advantage, this has not been supported by results from real data (van Binsbergen et al. 2015, Calus et al. 2016, Heidaritabar et al. 2016, Zhang et al. 2018), likely due to the high dimensionality of the models and high LD of the causal variants with large numbers of neighboring markers.

Preselection of variants prior to training of the model has been suggested both as an alternative and complement to variable selection methods. Heidaritabar et al. (2016) preselected SNPs based on mutation type (e.g., synonymous, nonsynonymous, and non-coding) from a full set of approximately 4.6 million markers but found no appreciable gain in accuracy. Other studies have attempted to identify the most relevant variants through association statistics such as p-values, absolute estimated effects, or the relative contribution to the genetic variance. Investigating inbred lines of *D. melanogaster*, Ober et al. (2012) selected the top 5% of SNPs ranked either by absolute estimated effect or the proportion of the genetic variance explained and found no significant improvement of accuracy using either preselection criteria. Veerkamp et al. (2016) preselected variants based on p-values in Holstein data and found no improvement in accuracy, with the additional disadvantage of bias in the GEBVs and inflation of the variance component estimates. Frischknecht et al. (2018) used p-values, annotation missense status, or LD-pruning to preselect variants; LD-pruning was the only strategy that did not reduce accuracies. Some studies have combined preselected SNPs with standard medium-density SNP chips to compromise between the potential benefits of each marker set, but few have found any benefit from this approach (Brondum et al. 2015, van den Berg et al. 2016, Wiggans et al. 2016, VanRaden et al. 2017). However, many of these studies performed SNP discovery and training of the prediction model using the same reference data set.

These results are not surprising and are in fact a consequence of the Beavis effect (Xu 2003), a variation of the so-called "winner's curse" phenomenon, where many of the selected SNP effects are overestimated, which will result in biased predictions and reduced accuracies in the validation set. Many studies that have investigated marker

preselection based on association statistics criteria (e.g., p-values, absolute estimated effects) have used the data twice (in preselection and training), and this could be the primary explanation for their failure to improve accuracies. Splitting the data into three non-overlapping sets for discovery, training, and validation may alleviate this bias; however, this is a suboptimal use of an expensive resource and could result in an increase in the standard error of estimates and corresponding decrease in power to detect relevant markers. Additionally, splitting the data may not eliminate the population structure that arises from families or breeds, which can contribute to an erroneously inflated association of markers with the trait (Price et al. 2010).

Toghiani et al. (2017) introduced the population statistic F_{ST}, a measure of deviation in allele frequencies between populations, as a criterion for marker preselection in genomic evaluations of livestock. They showed that by using high- and low-phenotype individuals within a population to calculate F_{ST} scores, historical selection signals could be detected at markers that tag causal loci. Chang et al. (2018) demonstrated that preselection of markers by F_{ST} scores could significantly improve genomic prediction accuracies, and even outperformed BayesB and BayesC as the dimensionality of the model increased. A subsequent study by (Chang et al. 2019) showed that genomic similarity between individuals will be maximized using a highly stringent subset of the top markers as ranked by F_{ST} scores, though accuracies will not be maximized using this subset. They proposed that the highest potential accuracy will be achieved when a balance between high genomic similarity and the proportion of genetic variance explained is achieved.

In this study, we expanded upon these results by investigating how the inclusion of markers in linkage equilibrium with causal loci impact the estimation of genomic relationships and affect prediction accuracies. Additionally, we compared the sensitivity of F_{ST} scores and estimated SNP effects as preselection criteria to discriminate between markers that are linked and unlinked with causal loci and the potential of each to increase accuracies.

Materials and Methods

The simulated genome consisted of 30 chromosomes (100 centiMorgans each in length) that harbored a total of 777k evenly-distributed SNP markers and was generated using QMSim (Sargolzaei and Schenkel 2009). As one of the primary goals of this study was to investigate how markers that segregate independently from QTL affect prediction accuracies, 200 QTL were randomly distributed across 2 of the 30 chromosomes, as illustrated in Fig. 3.1a. While it is an extreme scenario for real distribution of QTL for a complex trait, this design allowed unambiguous classification of approximately 725 k markers as segregating independently from any QTL, with the approximately 52 k remaining markers potentially linked with at least one QTL. QTL were limited to 2 chromosomes to ensure a high LD of the majority of markers on these chromosomes with at least one QTL and to evaluate the behavior of markers that are unlinked with QTL in a prediction model. QTL effects were sampled from a Gamma distribution with a shape parameter of 0.4.

The LD structure was generated through 2070 generations of random mating in a historical population, with a bottleneck occurring at generation 2000. The number of

individuals in the historical population varied between 600 and 4000. The resulting average LD (r^2) was 0.32 and 0.084 between consecutive SNP and QTL, respectively.

A trait with a heritability equal to 0.4 was simulated. The genetic and residual variances were set equal to 0.4 and 0.6, respectively. The model used to simulate the phenotypes included the true breeding value (the cross product of true QTL effects and genotypes) and a normally distributed error term with mean zero and dispersion equal to the residual variance.

All individuals from the last generation of the historical population (500 males and 3500 females) were selected to be founders of a population under selection (SP). An additional seven generations (3500 progeny each) were generated. Pedigree-based estimates of breeding values (pEBVs) were used for selection. Sire and dam replacement rates were set equal to 0.5 and 0.2, respectively. The training population consisted of the first six SP generations. It included 2 k phenotyped individuals; a randomly selected half of these were also genotyped. The seventh generation of SP consisted of 3.5k genotyped individuals, none of which were phenotyped, and was used as the validation population. Ten replicates of the simulated genome and phenotypic data were generated.

Predictions were made using a single-step GBLUP model (ssGBLUP) (Legarra et al. 2009, Aguilar et al. 2010, Christensen and Lund 2010) as implemented in the BLUPF90 software (Aguilar et al. 2018). The genomic relationship matrix (**G**) was constructed according to VanRaden (2008),

$$\boldsymbol{G} = \frac{\boldsymbol{Z}\boldsymbol{Z}'}{2\sum_{i=1}^{n} p_i(1-p_i)} \ ,$$

Where **Z** is a matrix of SNP genotypes, p_i is the minor allele frequency of the i^{th} SNP, and n is the number of SNPs.

In order to evaluate the consequences that SNPs unlinked with QTL have on prediction accuracies, three analyses that differed in the linkage status of the SNPs used to build $\boldsymbol{G}$ were performed. The three SNP subsets considered were 1) only the 51,800 SNP situated on either of the two chromosomes harboring around 100 QTL each (HQ2), 2) only the 725,200 SNP on any of the 28 chromosomes that lack QTL (LQ28) and can be definitively classified as being unlinked with QTL, and 3) the union of the two previous subsets that includes all 777,000 SNP regardless of linkage status.

In the next stage of the study, F_{ST} was used as a criterion for preselection of SNP subsets and was compared to preselection according to absolute estimated marker effect or random subsets. Subsets of 1, 10, 20, 30, 40, and 50 =k SNPs for each preselection criteria were used for the construction of $\boldsymbol{G}$ and corresponding prediction accuracies, defined as the Pearson correlation between true and estimated breeding values of validation individuals, compared.

F_{st} scores were calculated following Nei (1973),

$$F_{ST} = \frac{H_T - H_S}{H_T} \ ,$$

where $H_T = 2 \cdot p_T \cdot q_T$, $H_S = \frac{H_{S1} \cdot n_{S1} + H_{S2} \cdot n_{S2}}{n_{S1} + n_{S2}}$ and $H_{Si} = 2 \cdot p_{Si} \cdot q_{Si}$, where p_T and q_T are major and minor allele frequencies, respectively, in the population; p_{Si} and q_{Si} are major and minor allele frequencies, respectively, in the i^{th} subpopulation; and n_{Si} is the number of individuals in the i^{th} subpopulation. The genotyped and phenotyped individuals in the training population were ranked according to their phenotype and the bottom and top 5% of individuals used to create two subpopulations (S_1 and S_2), as illustrated in Fig. 3.1b. Using these subpopulations, an F_{st} score for each SNP was computed as indicated in

formulae above. SNPs were then ranked based on their F_{st} scores and subsets of the top 1, 10, 20, 30, 40, and 50k markers used to compute $\boldsymbol{G}$.

The rationale for forming subpopulations from individuals of extreme phenotype in a single breeding population rather than using separate breeding populations with highly divergent phenotypes (e.g., milk production in Holstein-Friesian and Jersey) is to minimize the likelihood of preselection of adaptive SNP markers that are specific to a population due to natural or artificial selection. By obtaining extreme phenotypes from within a single breeding population, any potential divergence in allele frequency related to traits uncorrelated with the one of interest will be more effectively averaged over.

Estimated genomic breeding values that were obtained using all 777k markers to compute **G** were used to derive SNP effects through the following relationship (Wang et al. 2012),

$$\hat{\boldsymbol{u}} = \boldsymbol{DZ}'[\boldsymbol{ZDZ}']^{-1}\hat{\boldsymbol{a}} \ ,$$

where $\hat{\boldsymbol{u}}$ is the vector of SNP effects, $\hat{\boldsymbol{a}}$ is the vector of estimated genomic breeding values for individuals in the training population with $\boldsymbol{G}$ modeled using all 777k SNPs, $\boldsymbol{Z}$ is a known incidence matrix of SNP genotypes, and $\boldsymbol{D}$ is a diagonal matrix of weights. In this study, D was set equal to the identity matrix to convey equal weight to all SNPs. Estimation of genomic breeding values and the back-calculation of SNP effects were obtained using ssGBLUP (Legarra et al. 2009, Aguilar et al. 2010, Christensen and Lund 2010) as implemented in BLUPF90 (Aguilar et al. 2018) and PreGSF90 (Aguilar et al. 2014), respectively.

Random SNP subsets were generated by random sampling from all available SNPs with no restrictions placed on the number of markers sampled from a particular

chromosome or the proportion that were linked and unlinked with QTL. A generalized outline of the approach to the analysis for Fst and SNP effect preselection and predictions is summarized in Fig. 3.1b.

Genomic information improves prediction accuracies compared to pedigree primarily through a better modeling of the MS of QTL. To dissect how the various marker-estimated genomic relationships capture the true QTL similarity between individuals and how they contribute to the maximization of prediction accuracy, several metrics were used to quantify the agreement between marker- and QTL-based $\boldsymbol{G}$ matrices.

First, a correlation was calculated between all elements of the full marker- and QTL-based $\boldsymbol{G}$ matrices, as suggested by VanRaden (2008); this correlation reflects the adequacy of the estimated genomic relationships to capture both the expected relatedness and MS. To further evaluate how well each set of markers models the MS component specifically, expected relationships were subtracted from each genomic relationship, and a correlation between the resulting centered $\boldsymbol{G}$ matrices was calculated,

$$cor\left(\boldsymbol{G}_M - \boldsymbol{A}_{22}, \boldsymbol{G}_{QTL} - \boldsymbol{A}_{22}\right),$$

where $\boldsymbol{G}_M$ and $\boldsymbol{G}_{QTL}$ are the marker and QTL relationship matrices, respectively, and $\boldsymbol{A}_{22}$ is the matrix of expected relationships based on pedigree information for genotyped individuals. It is possible that certain markers will capture variation in MS that is not consistent with the MS of QTL when LD between markers and QTL is low. If the marker-estimated and QTL relationships fall in opposite directions around the expected relationship, then the expected relationship will in fact be a better estimate than the

marker-estimated genomic relationship. The ability of marker subsets to capture the correct direction of the MS of QTL was determined as,

$$Directional\ MS = \begin{cases} \left|\frac{\boldsymbol{G}_M - \boldsymbol{A}_{22}}{sd(\boldsymbol{G}_M)}\right| & if\ same\ direction\ as\ QTL\ MS \\ -\left|\frac{\boldsymbol{G}_M - \boldsymbol{A}_{22}}{sd(\boldsymbol{G}_M)}\right| & if\ opposite\ direction\ from\ QTL\ MS \end{cases}$$

where $sd(\boldsymbol{G}_M)$ is the standard deviation over all relationships within $\boldsymbol{G}_M$, and the sign reflects whether the MS component for marker-estimated relationships falls in the same (positive) or opposite (negative) direction as the QTL MS relative to the expected relationship.

At a minimum, marker-estimated relationships should capture the correct direction of the MS of QTL in order to improve the modeling of relationships relative to the expectation. An ideal set of markers will additionally minimize the distance between marker- and QTL-based relationships. This distance can be approximated using the following formulae,

$$MS\ Error\ (\%) = \left|\frac{\frac{\boldsymbol{G}_M - \boldsymbol{A}_{22}}{sd(\boldsymbol{G}_M)} - \frac{\boldsymbol{G}_{QTL} - \boldsymbol{A}_{22}}{sd(\boldsymbol{G}_{QTL})}}{\frac{\boldsymbol{G}_{QTL} - \boldsymbol{A}_{22}}{sd(\boldsymbol{G}_{QTL})}}\right| \times 100\%.$$

This puts the discrepancy between marker-estimated and QTL relationships in terms of an error (%) relative to the scale of the MS of QTL. The closer a value is to zero, the less the discrepancy between the marker-estimated and QTL relationship. A value less than one implies that the marker-estimated relationship is a closer approximation of the QTL relationship than the expected relationship, while a value greater than one implies either that the marker-estimated relationship captures the correct direction but overestimates the

QTL MS, or that the marker-estimated relationship has opposite direction of MS than the QTL. Figures were generated using the tidyverse package in R (Wickham et al. 2019).

Results

Accuracy of prediction was 0.37, 0.62, 0.89 and 0.22 using pedigree, all, HQ2, and LQ28 markers, respectively, to model the relationship matrix. As expected, the highest (0.89) and lowest (0.22) accuracies were obtained when the genomic relationship matrix was constructed using only linked (HQ2) or unlinked (LQ28) markers (Fig. 3.1a), respectively. Using the latter, accuracy was 39.6% lower than that achieved using expected relationships despite being based on genomic information. While use of all 777k markers outperformed expected relationships by 70.3%, the accuracy was still approximately 30% lower than that obtained using only HQ2 SNPs.

Accuracies based on marker subsets preselected either randomly, by F_{ST} scores, or by estimated effects are shown in Table 3.1. When markers were preselected randomly, accuracy increased rapidly and plateaued when approximately 20k markers were used. This is similar to the trend observed using commercial genotyping panels, where a subset of reasonably well-distributed markers yielded prediction accuracies similar to much higher density platforms. Although 50 to 60k markers are typically necessary for many livestock species before reaching a plateau in accuracy, the smaller number of SNPs required in this study is likely due to the unconventional simulated genome structure and high LD between markers and QTL.

Use of markers preselected based on F_{ST} scores resulted in a higher accuracy compared to the use of all markers. In fact, accuracy increased between 26.7 and 36.4% across all subsets. Accuracy peaked with the use of the top 10k markers and remained

fairly persistent; the decrease in accuracy was only 7.1% as the number of preselected markers increased to 50k.

For preselection based on SNP effects, accuracy for 1k markers was initially comparable to that achieved using 10k F_{ST}-preselected markers (0.84 and 0.85, respectively); however, accuracies rapidly declined (by 20.2%) with larger subsets and the top 50k markers yielded accuracies that exceeded use of all markers by only 8%.

Table 3.2 shows the percentage of preselected markers that are located on either of the two chromosomes harboring QTL. These percentages are measures of the sensitivity of the preselection criteria to detect markers that are truly linked with causal loci. The top 1k F_{ST}-preselected markers were almost all (99.99%) SNPs in true linkage with QTL. The sensitivity steadily declined as the number of preselected markers increased and reached a minimum of only 28% linked when 50k markers were preselected. Preselection by SNP effects followed a similar trend but had greater sensitivity to detect markers potentially linked with QTL for all subsets compared to F_{ST}.

The proportion of genetic variance explained by preselected marker subsets is shown in Table 3.3. The genetic variance contributed by a particular QTL was considered explained by a marker subset if at least one marker had an r^2 greater than 0.9 with the QTL. As expected, preselection using a random selection criterion explained the least amount of the genetic variance. Preselection by F_{ST} and absolute estimated effects resulted in significantly more genetic variance explained, as much as 40 and 41%, respectively. Yet for neither criteria did maximization of genetic variance explained coincide with maximization of prediction accuracy, likely as a consequence of an increasing proportion of unlinked markers present in larger subsets (Table 3.2).

Genomic information increased accuracy compared to pedigree by improving modeling of the MS. The effectiveness of a set of markers to capture QTL similarity and MS between individuals could be evaluated by assessing the correlation between marker- and QTL-based $\boldsymbol{G}$ matrices. The non-centered $\boldsymbol{G}$ matrix reflects the total QTL similarity while the centered $\boldsymbol{G}$ matrix (Eq. 1) will reflect the MS component only.

Correlations between the marker- and QTL-based $\boldsymbol{G}$ matrices for all, HQ2, or LQ28 markers are listed in Table 3.4. As expected, the non-centered correlations followed the same trend as that observed for the accuracies, with the maximum (0.63) and minimum (0.28) correlation obtained using only HQ2 and LQ28 markers, respectively. When $\boldsymbol{G}$ was centered by the expected relationships, the correlation for LQ28 markers was effectively zero. In contrast, using only linked markers to construct $\boldsymbol{G}$, the correlation decreased by just 8.4% after adjusting for expected relationships.

This independence between the variation of LQ28 markers and QTL is illustrated in Fig. 3.2a, which plots the density of Eq. 2 for all, HQ2, and LQ28 markers. For the LQ28 subset, the distribution of this directional MS component falls evenly around zero; the number of marker-estimated relationships that fail to capture the correct direction of the QTL MS and the number that capture it correctly are approximately equal (Fig. 3.2b). The distribution for HQ2 is shifted towards more positive values, showing that this group of markers estimates the correct direction of the QTL MS more often than not. Interestingly, HQ2 markers still fail to capture the correct direction of the MS of QTL approximately 30% of the time (Fig. 3.2b); this likely occurs primarily when the deviation of the QTL genomic relationship from the expectation is quite small.

Tables 3.5 and 3.6 show the non-centered and centered correlations of the QTL-based **G** with **G** based on F_{ST}- and effect-preselected subsets, respectively. For F_{ST}, the correlation followed a similar trend as that observed for the accuracies (Table 3.1), with the largest correlation for both non-centered and centered **G** matrices achieved using the top 10 k F_{ST}-preselected markers. The correlation for effect also peaked at the top 10k markers, however, this does not coincide with where the accuracy is maximized. The relative decrease in the correlation with centering was smaller for SNP effects than for F_{ST}-score-based prioritization, indicating that marker effects have a slightly better ability to capture the direction of the MS of QTL (Fig. 3.3a). However, both preselection criteria for all subsets considered were more likely than not to identify the true direction of the MS, as presented in Fig. 3.3b and c.

Figure 3.4 presents the distribution of the errors in estimating the MS of the QTL (Eq. 3) using subsets of markers preselected by F_{ST} and absolute estimated effects. For both preselection methods, the error was minimized when only 10k markers were preselected (highest density near zero). This coincides with the subset that maximizes accuracy for F_{ST}, but not for preselection by estimated effects. Preselection based on the magnitude of the estimated effect maximized the accuracy using 1k markers, which actually appears to yield the greatest error in MS estimation among the subsets considered.

When only 1k SNPs were prioritized, the estimated effects preselection method seems to outperform the F_{ST}-score-based approach. However, beyond the top 1k panel, F_{ST} preselection consistently yields significantly higher accuracies. This coincides with when the sensitivity of both preselection methods starts to decrease, and unlinked

markers begin to form part of the preselected subsets. This suggests that the difference between the two approaches is a consequence of the unlinked markers selected. Figure 3.5a and b show the regression of F_{ST} on estimated effect for HQ2 and LQ28 markers, respectively. There is a more consistent trend between the two statistics for HQ2 than for LQ28 markers. The Pearson correlation between F_{ST} and estimated effect is 0.77 and 0.27 for HQ2 and LQ28 markers, respectively. Together these results suggest that the two statistics tend to have high agreement when a prioritized marker is linked with a QTL but less so when the marker is unlinked.

In Fig. 3.5b, the threshold for inclusion in the top 10k marker subsets for F_{ST} and estimated effects are denoted by a yellow and blue lines, respectively. It is clear that more SNPs with a large spurious association are preselected when using estimated SNP effects rather than F_{ST} scores. Without an independent training dataset, these large spurious associations will be re-estimated and exacerbated when training the prediction model and negatively affect the prediction accuracy in the validation set. The higher and more persistent accuracy for larger subsets when using F_{ST} as a preselection tool could be explained by its tendency to select markers that on average have less pronounced spurious associations.

To investigate this further, the top or bottom 50k LQ28 (unlinked) markers as ranked based on F_{ST} scores or absolute estimated effects were excluded from the full panel of 777k SNP markers. The reduced panels of 725k markers were then used for predictions and the resulting accuracies are presented in Table 3.7. Theoretically, given their lack of linkage with any QTL, it is expected that the excluded 50k top or bottom markers should not influence the accuracy. However, that was not the case and exclusion

of certain unlinked markers yielded an increase in accuracy, indicating that the analysis benefits from their absence.

Exclusion of the 50k unlinked markers with the largest estimated effects resulted in the largest increase in accuracy (approximately 8.6%) compared to use of all markers without preselection. In contrast, exclusion of the 50k unlinked markers with the smallest estimated effect led to no change in accuracy relative to use of all markers, as expected given that their estimated effects were close to zero. However, exclusion of the 50k unlinked markers with the largest F_{ST} scores resulted in a smaller increase in accuracy (4.1%), showing the superiority of the F_{ST} method in avoiding the preselection of unlinked markers with pronounced spurious associations.

While the simulation design previously evaluated is convenient for evaluating the behavior of markers that are unlinked with QTL in a prediction model, it would be unreasonable to expect a complex trait in reality to be accurately modeled by such a design. To evaluate whether a similar trend could persist under a more reasonable distribution of QTL across the entire genome, the simulation was repeated with the 200 QTL distributed across all 30 chromosomes. Table 3.8 shows accuracy and percentage of genetic variance explained for F_{ST}- and effect-preselected subsets.

With QTL distributed across all chromosomes, accuracy using all markers was 0.60. Both preselection methods achieve a maximum accuracy of 0.73, though F_{ST} requires a larger number of preselected markers to achieve this. As the panel size increases to 50k, the accuracy for effect- and F_{ST}-preselection decrease by approximately 12.3 and 2.7%, respectively. Despite yielding a lower accuracy for panels of 10k markers and larger, the effect-preselected subsets explain 9.1 to 17.2% more of the genetic

variance than the equivalently-sized F_{ST}-preselected subsets. This demonstrates that the trend in prediction results for F_{ST}- and effect-preselected subsets is consistent even when all chromosomes harbor multiple causal loci.

Discussion

It was shown that the predictive ability of markers that are unlinked with QTL is inferior to even pedigree information, a result that agrees with previous studies (Habier et al. 2007, Goddard 2009, Habier et al. 2013). However, despite their inferior predictive power, accuracies using only unlinked markers were always positive. Habier et al. (2007) attributes this to unlinked markers modeling additive genetic relationships and shows that the accuracy will converge to that of pedigree BLUP as the number of independently segregating markers increases. Regardless of linkage, the distribution of QTL and marker additive relationships for a particular order of kinship will share a mean, the expected relationship. The advantage of using genomic information compared to pedigree is the better modeling of the MS of QTL. However, when markers and QTL segregate independently the covariance of marker and QTL MS is zero (Table 3.4) and the marker-based relationships are noisy estimates of the average additive relationships. While these markers will independently yield positive accuracies, they should not be expected to benefit the analysis when markers in LD with causal loci are available. HQ2 markers also capture the additive relationship with the additional benefit of accounting for some portion of the MS of QTL, as evidenced by the limited decrease in the correlation between the HQ2-marker- and QTL-based **G** matrices after centering with expected relationships (Table 3.4) and the shift of the HQ2 distribution in Fig. 3.2a to more positive values.

Ideally, the effect of unlinked markers on the estimation of the breeding values would be zero when more informative markers are present in the model. However, the inferior accuracy obtained using all markers compared to only HQ2 markers demonstrates that the effect of unlinked markers will not be null. The results of this study demonstrate that in terms of a GBLUP model, allowing unlinked markers to have a nonzero contribution to $\boldsymbol{G}$ adds noise to the estimation of genomic relationships that will not be reflective of true QTL similarity, resulting in lower accuracy relative to that achieved using only linked markers in the validation population. In terms of a SNP-BLUP model, which has been shown to be equivalent to GBLUP (Habier et al. 2007), nonzero estimates will be obtained for unlinked markers that have no association with QTL inheritance in validation individuals. Table 3.4 shows that the MS of QTL and unlinked markers vary around the same average relationship, which creates an association of the unlinked markers with the QTL. The model cannot discriminate spurious marker associations that are a result of this shared expectation and random sampling from associations due to true linkage with a causal locus, particularly when the unlinked markers are themselves used to inform the variance-covariance structure.

These results highlight the motivation and potential for preselection of markers to improve accuracies. Both F_{ST} scores and absolute estimated effect preselection-based methods were able to identify relevant markers with high sensitivity when preselecting a small number of markers and yielded high accuracies. However, the trend in accuracy differed substantially between the two approaches. As the number of preselected markers increased, their sensitivity to detect linked markers decayed, and unlinked markers were incorrectly selected. Preselection by F_{ST} increased accuracy from 1k to 10k markers

while the accuracy for preselection by estimated effects decreased by approximately 7.1% over the same interval. This occurred despite F_{ST} preselection adding 903 more unlinked markers and explaining approximately 5% less of the genetic variance than estimated effects. The accuracy for F_{ST} preselection declined as the number of preselected markers increased beyond 10k, but was more persistent than the accuracy for estimated effects despite consistently selecting more unlinked markers and explaining less of the genetic variance.

There are two important concepts that are illustrated by the behavior of these statistics. First, when the preselection criteria have imperfect sensitivity, accuracy will be maximized by a balance between increasing the genetic variance explained and minimizing deleterious contributions from poorly informative markers. F_{ST} added a large number of unlinked markers when the number of preselected SNPs increased from 1k to 10k, but the genetic variance explained was also significantly increased, resulting in an overall improvement in accuracy. As long as the beneficial contribution to the genetic variance explained by linked markers exceeds the negative effects of the association noise added by unlinked markers, the accuracy will increase. The decline in accuracy for F_{ST} when the number of preselected markers increased from 10k to 20k is explained by the fact that the genetic variance explained increased by only 2.6% while approximately 73% of added markers were unlinked with QTL; this likely contributed significant noise to estimation of genomic relationships. This is in concordance with Chang et al. (2019), who concluded that a balance is needed between genomic similarity and the proportion of genetic variance explained by the preselected markers in order to maximize accuracies. While in the current study we make only a distinction between linked and unlinked

markers, markers that are linked to but in low LD with a QTL will also contribute noise to the model and the negative impact of this noise may outweigh the benefit of any genetic variance they explain.

Second, the noise contributed by unlinked markers is not necessarily equal between both preselection methods. Estimated-effects-based approach consistently showed a greater sensitivity to detect linked markers than F_{ST}, yet yielded significantly lower accuracies, except in the case of the 1k panel where it selected no unlinked markers. For panel sizes of 10k and larger, the accuracy for the estimated-effects-based approach was lower than for F_{ST} scores largely because the unlinked markers selected by the approach have a greater detrimental effect.

When the 50k most spuriously associated unlinked markers were excluded from the analysis (Table 3.7), accuracies improved significantly. These markers have a large spurious association with the trait and the analysis benefits from their exclusion. While the complications that such markers present are often considered in the context of marker preselection, this result shows that such markers will have an appreciable negative impact even in the absence of preselection. There is therefore an incentive to identify and filter spuriously associated markers if a reliable and efficient method for distinguishing them from true associations can be developed.

Excluding the 50k LQ28 markers with the largest F_{ST} scores from the full panel also resulted in the accuracy increasing, but this increase was not as pronounced as when the LQ28 markers with largest estimated effect were excluded. This indicates that when the training data is also used to calculate F_{ST} for preselection, there will be some tendency to select irrelevant markers with a spurious association, but that the spurious associations

will on average be less severe than when preselecting by the absolute estimated effects. This could explain why accuracies are more persistent for preselection by F_{ST} scores than estimated marker effects even when the F_{ST} preselection criteria selects more unlinked markers and explains less of the genetic variance.

Both F_{ST} and marker effects were estimated using some portion of the training data rather than an independent dataset. While partitioning of the training data into two subsets, one for estimation of preselection statistics and one for training of the prediction model, may alleviate some bias, it will decrease the size of the data available for training the model and therefore increase the standard error in estimation of the statistics anyway. Splitting of the training data will not be a feasible option for most analyses, and the literature shows that several analyses that consider preselection by association statistics in genetic improvement programs have chosen to reuse the SNP discovery data for training of the model.

In contrast to marker effect estimation, calculation of F_{ST} used just 10% of the training data (Fig. 3.1). Spurious associations present in the full training data may be less extreme in subsets of that data, which could explain why F_{ST} is less affected by the bias that results from using the same data for both preselection and model training. F_{ST} then has the potential to be a simple and efficient preselection tool that can reduce the bias associated with preselection by association statistics without requiring an inefficient partitioning of the training data or expensive collection of new independent data.

F_{ST} scores and association statistics could potentially be combined into an index to harness the benefits of both preselection statistics. The Pearson correlation between F_{ST} scores and estimated effects was 0.78 and 0.28 for HQ2 and LQ28 markers,

respectively. This suggests that there is high agreement among the two statistics when markers are linked with QTL, but much less so among unlinked markers. Spuriously associated markers could possibly be identified and excluded when there is large disagreement between the two statistics.

An additional benefit of F_{ST}-based prioritization is that it is not affected by an increase in the number of markers included in the model due to the independence in calculating the score of each marker. As the number of markers in the association model increases, estimation variance for estimated effects of markers will increase without a corresponding increase in the size of the training data set. Furthermore, the estimated effect of each marker will be further regressed toward zero as QTL effects become distributed over correlated blocks of the predictors (Ishwaran and Rao 2010). This will further complicate disentangling true from spurious associations as both take a similar magnitude of estimated effect. In contrast, F_{ST} scores will remain constant regardless of the number of markers, correlated or uncorrelated, that enter jointly into the analysis. This does carry the drawback that highly correlated markers will have similar F_{ST} scores and so selecting only by top F_{ST} score will select all correlated markers in a block, which could cause bias (Frischknecht et al. 2018) and inflation of variance estimates (Veerkamp et al. 2016) due to multicollinearity. While not evaluated in this study, these issues could be avoided through LD-pruning of F_{ST}-selected markers or similar filtering measures.

Variable selection models are a conceptually similar but fundamentally different approach to marker preselection for reduction of the parameter space. While we do not explore a comparison of F_{ST} and variable selection models in this study, Chang et al. (2018) compared F_{ST} preselection implemented in a BayesA-like regression with BayesB

and BayesC. They found that while F_{ST} preselection did not outperform the Bayesian variable selection models in all scenarios, it did tend to have an advantage as the density of the full panel increased. In general, they found that BayesB and BayesC accuracies decreased with increased density of available markers, while accuracy using F_{ST} preselection tended to improve. This seems to be a result of the decreased statistical power to identify relevant markers flanking QTL of low effect as the number of parameters in the model increases. The benefits of both approaches might be harnessed by using F_{ST} to preselect markers with a generous threshold followed by implementation in a variable selection model that includes the prioritized markers.

Conclusions

In this study, F_{ST} was shown to be an efficient criterion for preselection of trait-relevant markers that can improve modeling of QTL similarity between individuals, increase prediction accuracies, and maintain more stable prediction accuracies than comparative association statistics like absolute estimated marker effects. While association statistics are powerful tools for identifying loci associated with a particular trait, disentangling spurious associations from weak but true signals is often not possible within the constraints of the data available. We showed that the more persistent prediction accuracy using F_{ST}-score-prioritized markers was the result of the ability of the F_{ST}-score-based method to select unlinked markers with weaker spurious associations with the trait compared to preselection based on absolute estimated effects. While this study only explored F_{ST} scores as an independent preselection statistic, we showed that F_{ST} and absolute estimated effect approaches preselected highly correlated sets of markers linked with QTL but significantly less so among unlinked markers. This

highlights the potential for the possibility of combining F_{ST} scores (or similar population statistics) and association statistics into a powerful preselection index to reduce inclusion of nonrelevant markers with a large spurious association.

References

Aguilar, I., et al. (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): 743-752.38

Aguilar, I., et al. (2014). PREGSF90 – POSTGSF90: Computational Tools for the Implementation of Single-step Genomic Selection and Genome-wide Association with Ungenotyped Individuals in BLUPF90 Programs. World Congress on Genetics Applied to Livestock Production (WCGALP), Vancouver, Canada.124

Aguilar, I., et al. (2018). BLUPF90 suite of programs for animal breeding with focus on genomics. Proceedings of the World Congress on Genetics Applied to Livestock Production, New Zealand.121

Brondum, R. F., et al. (2015). "Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction." J Dairy Sci **98**(6): 4107-4116.56

Calus, M. P., et al. (2016). "Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection." Genet Sel Evol **48**(1): 49.111

Chang, L.-Y., et al. (2019). "Increasing accuracy of genomic selection in presence of high density marker panels through the prioritization of relevant polymorphisms." BMC Genetics **20**(1).118

Chang, L. Y., et al. (2018). "High density marker panels, SNPs prioritizing and accuracy of genomic selection." BMC Genet **19**(1): 4.106

Christensen, O. F. and M. S. Lund (2010). "Genomic prediction when some animals are not genotyped." Genetics Selection Evolution **42**(2).39

Croiseau, P., et al. (2011). "Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm." Genet Res (Camb) **93**(6): 409-417.110

Daetwyler, H. D., et al. (2014). "Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle." Nature Genetics **46**(8): 858-865.49

Daetwyler, H. D., et al. (2010). "The Impact of Genetic Architecture on Genome-Wide Evaluation Methods." Genetics **185**(3): 1021-1031.40

Druet, T., et al. (2013). "Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions." Heredity **112**(1): 39-47.50

Frischknecht, M., et al. (2018). "Short communication: Genomic prediction using imputed whole-genome sequence variants in Brown Swiss Cattle." Journal of Dairy Science **101**(2): 1292-1296.73

Garcia-Ruiz, A., et al. (2016). "Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection." Proc Natl Acad Sci U S A **113**(28): E3995-4004.62

Genomes Project, C., et al. (2015). "A global reference for human genetic variation." Nature **526**(7571): 68-74.104

Goddard, M. (2009). "Genomic selection: prediction of accuracy and maximisation of long term response." Genetica **136**(2): 245–257.126

Habier, D., et al. (2007). "The impact of genetic relationship information on genome-assisted breeding values." Genetics **177**(4): 2389-2397.29

Habier, D., et al. (2013). "Genomic BLUP decoded: a look into the black box of genomic prediction." Genetics **194**(3): 597-607.127

Hayes, B. J., et al. (2014). Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project, Vancouver, Canada.52

Heidaritabar, M., et al. (2016). "Accuracy of genomic prediction using imputed whole-genome sequence data in white layers." J Anim Breed Genet **133**(3): 167-179.63

Hickey, J. M., et al. (2014). "Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation." Crop Science **54**(4): 1476-1488.109

Ishwaran, H. and J. S. Rao (2010). Generalized ridge regression: geometry and computational solutions when p is larger than n.128

Legarra, A., et al. (2009). "A relationship matrix including full pedigree and genomic information." J Dairy Sci **92**(9): 4656-4663.120

Meuwissen, T. and M. Goddard (2010). "Accurate prediction of genetic values for complex traits by whole-genome resequencing." Genetics **185**(2): 623-631.107

Meuwissen, T. H. E., et al. (2001). "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." Genetics **157**: 1819–1829.19

Nei, M. (1973). "Analysis of Gene Diversity in Subdivided Populations." Proceedings of the National Academy of Sciences **70**(12): 3321-3323.122

Ober, U., et al. (2012). "Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster." PLoS Genet **8**(5): e1002685.44

Price, A. L., et al. (2010). "New approaches to population stratification in genome-wide association studies." Nature Reviews Genetics **11**(7): 459-463.116

Rimbert, H., et al. (2018). "High throughput SNP discovery and genotyping in hexaploid wheat." PLoS One **13**(1): e0186329.105

Sargolzaei, M. and F. S. Schenkel (2009). "QMSim: a large-scale genome simulator for livestock." Bioinformatics **25**(5): 680-681.119

Toghiani, S., et al. (2017). "Genomic differentiation as a tool for single nucleotide polymorphism prioritization for Genome wide association and phenotype prediction in livestock." Livestock Science **205**: 24-30.117

van Binsbergen, R., et al. (2015). "Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle." Genetics Selection Evolution **47**(1).60

van den Berg, I., et al. (2016). "Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle." Genet Sel Evol **48**(1): 83.112

VanRaden, P. M. (2008). "Efficient methods to compute genomic predictions." J Dairy Sci **91**(11): 4414-4423.35

VanRaden, P. M., et al. (2017). "Selecting sequence variants to improve genomic predictions for dairy cattle." Genet Sel Evol **49**(1): 32.113

VanRaden, P. M., et al. (2009). "Invited review: reliability of genomic predictions for North American Holstein bulls." J Dairy Sci **92**(1): 16-24.37

Veerkamp, R. F., et al. (2016). "Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle." Genet Sel Evol **48**(1): 95.68

Wang, H., et al. (2012). "Genome-wide association mapping including phenotypes from relatives without genotypes." Genet Res (Camb) **94**(2): 73-83.123

Wickham, H., et al. (2019). "Welcome to the Tidyverse." Journal of Open Source Software **4**(43).125

Wiggans, G. R., et al. (2016). "Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle." Journal of Dairy Science **99**(6): 4504-4511.114

Xu, S. (2003). "Theoretical basis of the beavis effect." Genetics **165**: 2259–2268.115

Zhang, C., et al. (2018). "Genomic evaluation of feed efficiency component traits in Duroc pigs using 80K, 650K and whole-genome sequence variants." Genetics Selection Evolution **50**(1).108

Tables

Table 3.1 Accuracy of genomic predictions under varying number of random-, F_{ST}-, or estimated effect-based preselected markers

Selection Method[a]	**Number of Preselected SNPs (in thousands)**					
	1	10	20	30	40	50
Random	0.27	0.51	0.57	0.59	0.59	0.60
F_{ST}	0.81	0.85	0.83	0.81	0.80	0.79
Effect	0.84	0.78	0.72	0.69	0.68	0.67

[a] SNPs were preselected either randomly, based on their F_{ST} scores, or based on the absolute value of their estimated effect

Table 3.2 Overlap (%) between random-, F_{ST}-, or effect-preselected marker subsets and G_2 SNPs

Selection Method[a]	**Number of Preselected SNPs (in thousands)**					
	1	10	20	30	40	50
Random	6.75	6.65	6.63	6.64	6.63	6.69
F_{ST}	99.99	67.04	47.19	37.84	32.26	28.45
Effect	100.00	76.07	54.13	43.13	36.42	31.89

[a]SNPs were preselected either randomly, based on their F_{ST} scores, or based on the absolute value of their estimated effect

Table 3.3 Proportion of total GV[a] explained by random, effect, and F_{ST}-preselected markers

Selection Method[b]	**Number of Preselected SNPs (in thousands)**					
	1	10	20	30	40	50
Random	0.0041	0.018	0.082	0.11	0.10	0.14
F_{ST}	0.31	0.38	0.39	0.39	0.39	0.40
Effect	0.33	0.40	0.41	0.41	0.41	0.41

[a]GV Genetic Variance [b]SNPs were preselected either randomly, based on their F_{ST} scores, or based on the absolute value of their estimated effect

Table 3.4 Correlations between centered and non-centered genomic relationships with QTL relationships for different sets of markers[a]

	All	HQ2	LQ28
Non-Centered	0.35	0.63	0.29
Centered	0.16	0.58	0.0018
Relative Decrease (%)	0.54	0.084	0.99

[a]All = all markers; HQ2 = markers on the two chromosomes harboring the QTL; LQ28 = markers on the 28 chromosomes lacking QTL

Table 3.5 Correlations between non-centered and centered genomic and QTL relationships for varying numbers of F_{ST}-preselected markers

Selection Method[b]	**Number of Preselected SNPs (in thousands)**					
	1	10	20	30	40	50
Non-Centered	0.34	0.54	0.54	0.53	0.51	0.50
Centered	0.32	0.48	0.47	0.45	0.43	0.42
Relative Decrease (%)	0.073	0.12	0.14	0.15	0.15	0.16

Table 3.6 Correlations between non-centered and centered genomic and QTL relationships for varying numbers of effect-preselected markers

Selection Method[b]	Number of Preselected SNPs (in thousands)					
	1	10	20	30	40	50
Non-Centered	0.38	0.61	0.60	0.58	0.55	0.53
Centered	0.35	0.55	0.55	0.53	0.51	0.48
Relative Decrease (%)	0.074	0.094	0.090	0.083	0.082	0.084

Table 3.7 Accuracy after exclusion of different subsets of LQ28 markers from construction of the genomic relationship matrix

Exclusion Criteria[a]	**Excluded Markers[b]**		
	None	Top 50k	Bottom 50k
Effects	0.62	0.68	0.62
F_{ST} Scores	0.62	0.65	0.63

[a]Markers were excluded from the LQ28 subset based either on their F_{ST} scores or effects;
[b]All markers were included (None), top 50k markers excluded (Top 50k), and bottom 50k markers excluded (Bottom 50k)

Table 3.8 Accuracy and percent of genetic variance explained by F_{ST}- and effect-preselected subsets under a simulation design with 200 QTL across all 30 chromosomes

		Number of Preselected SNPs (in thousands)					
		1	10	20	30	40	50
Accuracy	F_{ST}	0.66	0.73	0.73	0.72	0.71	0.71
	Effect	0.73	0.70	0.67	0.66	0.65	0.64
GV Explained (%)	F_{ST}	0.21	0.29	0.31	0.32	0.32	0.33
	Effect	0.21	0.34	0.35	0.35	0.35	0.36

Figures

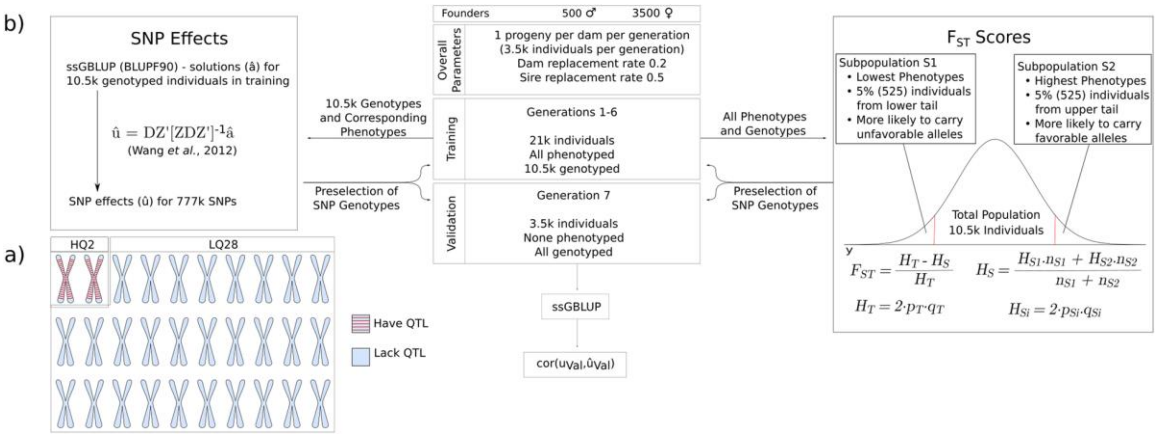

Figure 3.1 A general description of the simulation and workflow: **a**) A 30-chromosome genome was simulated with 200 QTL randomly distributed across 2 chromosomes and the remaining 28 chromosomes harboring no QTL. **b** A schematic representation of the pedigree simulation (7 generations of 3.5 k individual each). The first six generations (21 k phenotyped individuals and half of them genotyped) were used for training. The last generation consisting of 3.5 k genotyped and non-phenotyped individuals was used as validation set. Preselection of SNPs was based either on the absolute estimated marker effects or F_{ST} scores calculated using data from the training population

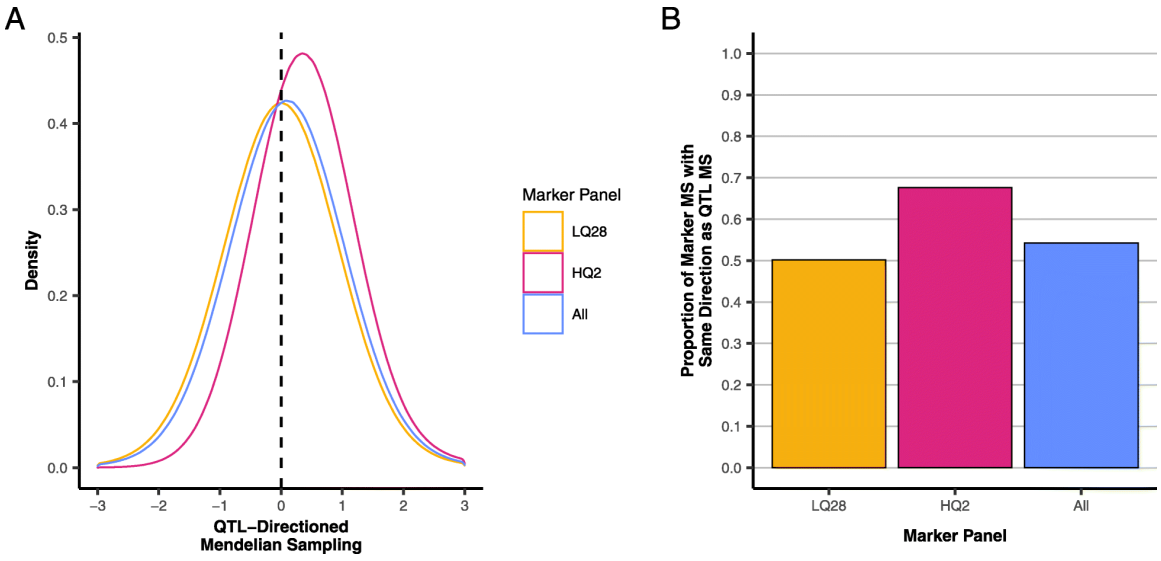

Figure 3.2 Characterization of the modelling of QTL Mendelian Sampling (MS) using all, HQ2, and LQ28 markers: **a**) The distribution of marker-estimated MS for relationships among training individuals with sign reflecting whether marker-estimated and QTL MS fall in the same (+) or opposite (−) direction relative to the expected additive relationship. **b** The proportion of relationships among training individuals for which marker-estimated and QTL MS fall in the same direction relative to expected additive relationships

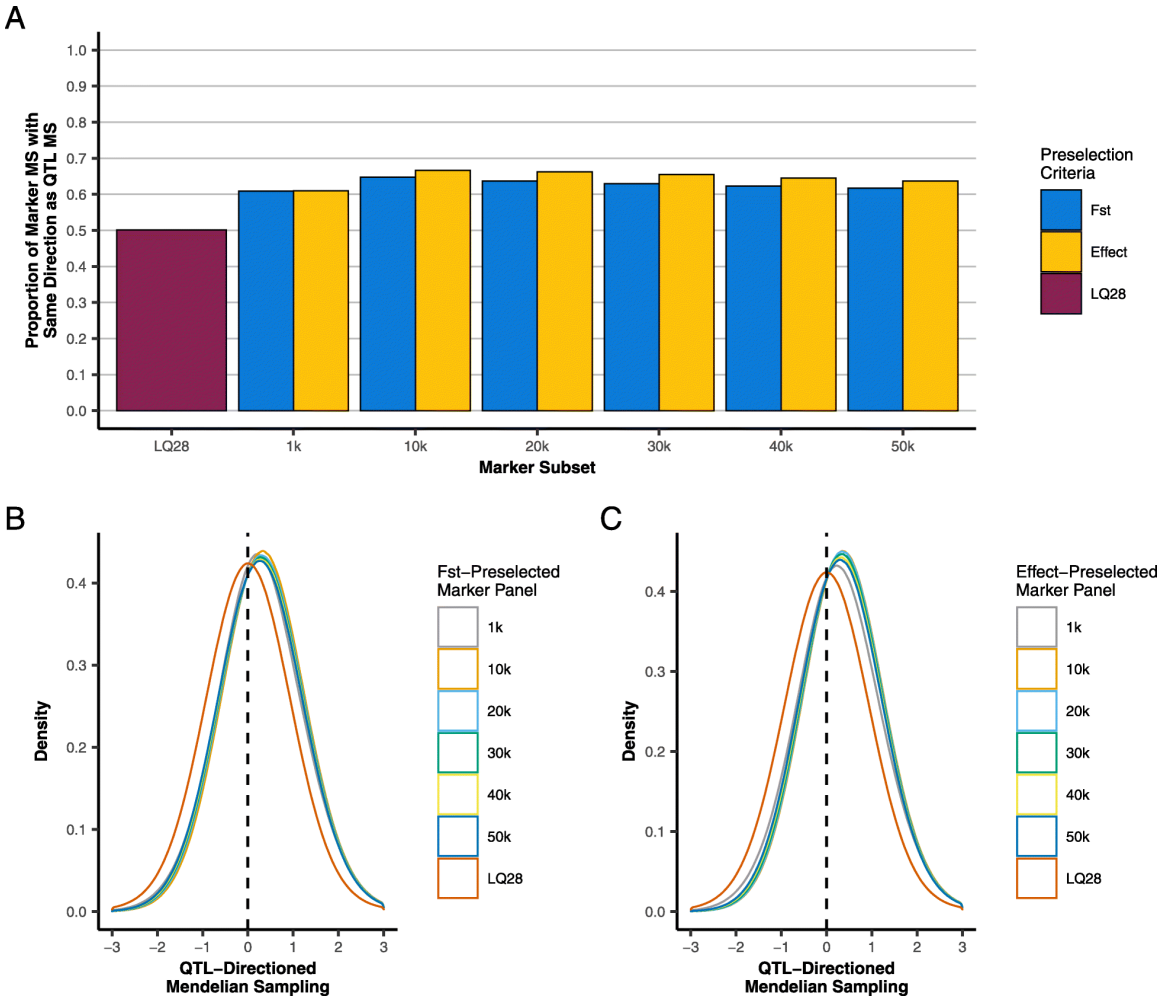

Figure 3.3 Characterization of the modelling of QTL Mendelian Sampling (MS) based on F_{ST}- and estimated-effects-preselected markers: **a**) The proportion of relationships among training individuals for which marker-estimated and QTL MS fall in the same direction relative to expected additive relationships. **b)** and **c)** The distribution of marker-estimated MS for relationships among training individuals with sign reflecting whether marker-estimated and QTL MS fall in the same (+) or opposite (−) direction relative to the expected additive relationship

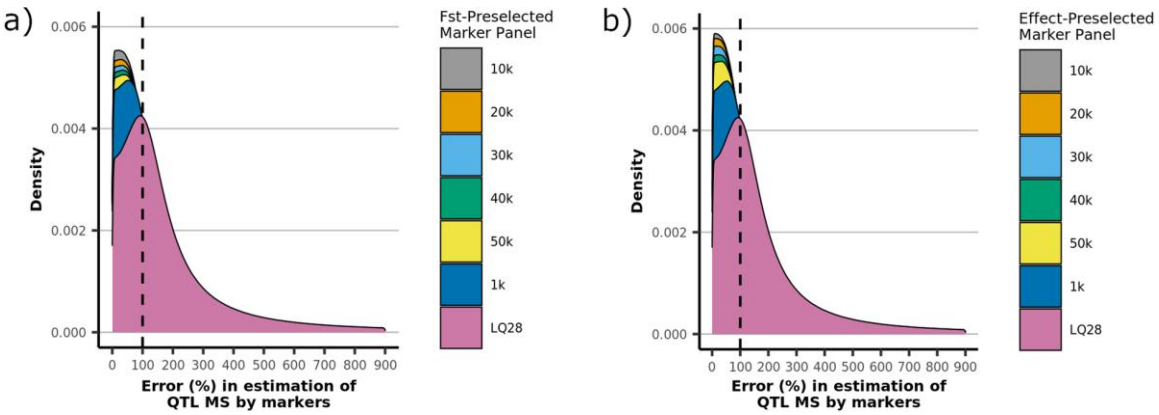

Figure 3.4 Errors in the estimation of QTL Mendelian Sampling: Distribution of error terms (%) in the estimation of genomic relationships (Eq. 3) for a) F_{ST} - and b) estimated effect-preselected marker subsets

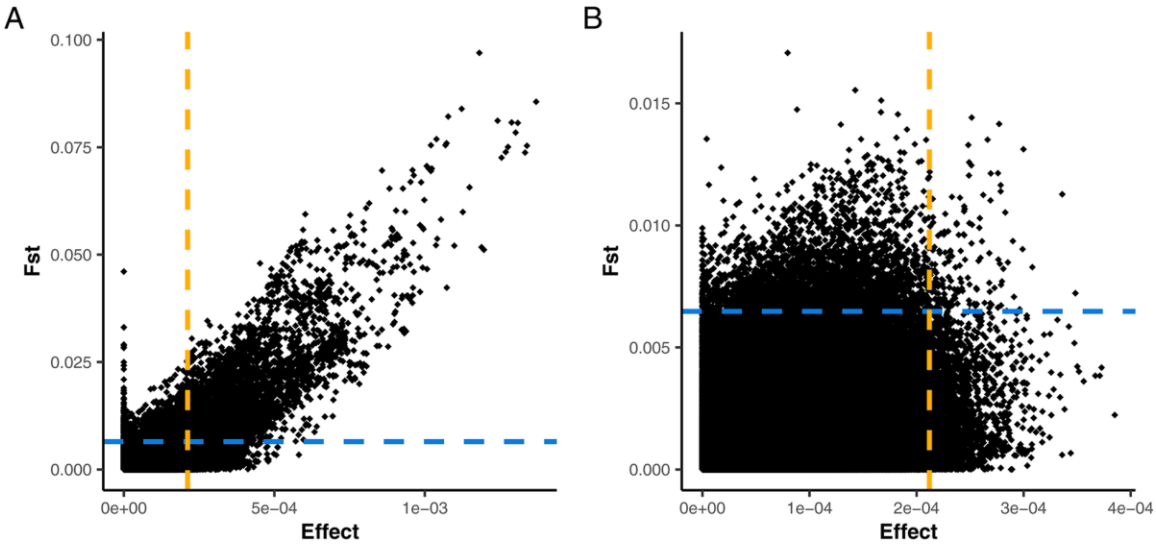

Figure 3.5 Regression of F_{ST} scores on the absolute estimated effect for **a**) HQ2 and **b**) LQ28 markers: The blue and yellow dashed lines denote the thresholds for selection of the top 10 k markers among all markers for F_{ST} and absolute estimated effects, respectively

CHAPTER 4

FUZZY LOGIC AS A STRATEGY FOR COMBINING MARKER STATISTICS TO OPTIMIZE PRESELECTION OF HIGH-DENSITY AND SEQUENCE GENOTYPE DATA[2]

[2] Ling, A.S., E.H. Hay, S.E. Aggrey, and R. Rekaya. To be submitted to *Genes*.

Abstract

The high dimensionality of genotype data available for genomic evaluations today has presented a motivation for development of strategies for identifying subsets of SNPs that are as or more predictive of economically-important traits than use of commercial SNP chips in genomic prediction models. In this simulation study, an algorithm for combining statistics used in the preselection and prioritization of SNP data from high-density genotype data (1.3 million SNPs) into a composite "fuzzy" ranking score based on a Sugeno-type fuzzy inference system was developed and evaluated for performance in preselection for genomic predictions. F_{ST} scores and p-values were evaluated as inputs for the fuzzy inference system. The accuracy of genomic predictions for fuzzy-score-preselected panel sizes of 1-50k ranged from -0.4-7.8, 1.0-11.7, and 0.9-7.7% higher than F_{ST} score and -0.3-3.8, 0.1-1.8, and 0.7-1.9% higher than p-value preselection of equivalent-sized panels for genetic architectures of 300, 1k, and 5k causative variants, respectively. Though gains in prediction accuracies using only two inputs to the fuzzy inference system were modest, preselection based on fuzzy scores yielded more accurate predictions than both F_{ST} scores and p-values for the majority of evaluated panel sizes under all genetic architectures. Fuzzy inference systems have potential to aggregate information from multiple criteria that reflect SNP-trait associations and biological relevance in a flexible and efficient way to yield higher quality genomic predictions.

Introduction

The affordability and accessibility of high-density and whole genome sequence data has increased in recent years to allow practical use in genomic evaluations. It was

initially expected that genomic prediction accuracy would increase using such data as a result of a better tracking of the sources of genetic variation (Meuwissen and Goddard 2010), but practical results show little to no advantage relative to low- and moderate-density single nucleotide polymorphism (SNP) panels currently used in genomic evaluations (van Binsbergen et al. 2015). This failure is likely due to a combination of excessive multicollinearity between SNP/variant predictors and overparameterization of the association model.

SNP marker preselection may have the potential to improve the quality of genomic predictions by reducing multicollinearity and overparameterization and increasing the biological relevance of markers included in the association model or in the calculation of the genomic relationship matrix. Several published studies have evaluated single preselection criteria, including p-values (Veerkamp et al. 2016), estimated marker effects (Ober et al. 2012), F_{ST} scores (Toghiani et al. 2017, Chang et al. 2018), and annotation missense status (Frischknecht et al. 2018). In each of these studies, prediction accuracies outcompeted genome-wide SNP panels only when one or a few variants had a large contribution to the genetic variance of the trait of interest for both simulated and real data.

The limited success of single-criteria preselection to improve prediction accuracies is likely to be explained by the limited sensitivity of any single criteria to distinguish between relevant and non-relevant markers. All criteria will be subject to false positives and the impact of such markers on predictions is unlikely to be null. Ling et al. (2021) showed that while SNP preselection by estimated effect results in less false positives and more genetic variance explained than F_{ST} preselection, the false positives

selected based on estimated effect have a greater negative impact on predictions and can ultimately offset the benefit from capturing more genetic variance. Based on these observations, it is reasonable to hypothesize that a composite score combining these two preselection criteria may harness the benefits of both to ultimately yield higher accurate genomic predictions. Xiang et al. (2019) proposed a composite score for ranking variants from sequence data based on association and biological criteria and found that when preselected variants were combined with a genome-wide 630k SNP panel, differences in genomic prediction accuracies between high- and low-ranking variants were statistically significant for several economically important milk traits in Holstein.

Fuzzy logic models are a computationally efficient and flexible tool for decision making based on multiple criteria. They have been successfully implemented in a variety of engineering applications, such as self-driving vehicles and automatic rice cookers (Refs). In this simulation study, we evaluate the potential of a Sugeno fuzzy logic model (Takagi and Sugeno 1985) to combine p-values and F_{ST} scores into a composite score for SNP preselection for use in genomic prediction.

It has been shown that the increase in the accuracy of genomic predictions in the presence of high-density data requires a balance between the fraction of the genetic variance explained by the prioritized SNPs and the number of preselected markers. Increasing the number of the latter will often increase the portion of genetic variance explained but not the accuracy of the genomic prediction, Thus, if whole genome sequence variants fails to significantly increase prediction accuracies by solely through the increase the percentage of genetic variance captured, then increasing accuracies must rely instead on decreasing the bias in genomic predictions.

Fuzzy logic was initially proposed by (Zadeh 1965) as a strategy to mimic the human approach in understanding data generating processes and logical decision making, where factors in the decision-making process may be subject to vagueness and uncertainty. It has been subsequently been applied to a variety of practical applications, from inventory control (Aengchuan and Phruksaphanrat 2015) to analysis of landslide susceptibility (Lucchese et al. 2021).

A fuzzy inference system (FIS) may be particularly suitable for genotype data and genomic prediction analyses as it is generally assumed that the vast majority of markers available are in imperfect linkage disequilibrium (LD) with causative variants rather than they themselves having a causative effect on the trait(s) of interest. While thresholds for assessing the amount of useful LD between a SNP and causative variant have been proposed (Du et al. 2007), these are clearly somewhat arbitrary and the usefulness of information carried by a particular SNP marker with regards to causative variants is more likely to fall in a gray area between true information and noise. However, it is clear that if the loci and genotypes of many causative variants were known then the problem that originally motivated the use of SNPs would be solved and there would be no purpose for SNP preselection. Instead, criteria that reflect association of SNPs with a particular trait, and by extension the potential LD with a causative variant, could be used in a fuzzy model.

Materials and Methods

Fuzzy logic is often used in assessing uncertainty associated with a decision-making process. It provides computers with the ability to compare alternatives in a similar manner to the reasoning process of humans. For example, when comparing the

daily milk production of two cows producing 40 and 41 kg, a crisp classification system based solely on phenotypic selection for high milk yield will conclude that the daily productions are different and the cow producing 41 kg is a better selection candidate. Realistically, a reasonable farmer would be more likely to use a "soft" approach by balancing the milk production with additional important factors that are not fully assessed (e.g., feed efficiency and health factors) and the final decision would potentially be different from the crisp classification. A fuzzy inference system is a process that formalizes this strategy and consists of three main components: 1) fuzzification of the inputs, 2) rule evaluation, and 3) defuzzification to achieve a final output. Conceptually, the final fuzzy outputs in this example would not be dissimilar to a selection index where performance across multiple traits is taken into account, though the fuzzy algorithm notably differs from how selection indices are constructed.

Fuzzification of the inputs involves constructing for each input class a fuzzy set, $A = (U, m)$, where U is a set that falls within the sample space of a given input, x , and m is a membership function that maps x to membership in U with $m(x) \in [0,1]$. Figure 4.1 illustrates the mapping of the inputs used in this study (p-values associated with estimated SNP effects and marker's F_{ST} scores) using piecewise linear functions as a membership function (m). It is worth noting that a given input value may have partial membership in adjacent classes, hence the fuzziness of the set, under the constraint that the sum of all memberships are less than or equal to one.

The rule evaluation step maps the underlying association between the input and output of the underlying system. Each set of inputs must be evaluated according to a rule base. The rule base follows the following generalized structure of,

$$if(Input\ 1\ is\ x_{i1} \Lambda\ Input\ 2\ is\ x_{i2}\ \Lambda \ldots \Lambda\ Input\ t\ is\ x_{it})\ then\ z_i = \boldsymbol{a}_i' \boldsymbol{x}_i$$

where z_i is the true output for individual i, modeled as a linear combination of the vector of input parameters ($\boldsymbol{x}_i$), $\boldsymbol{a_i}$ is a vector of unknown coefficients, and Λ stands for the Boolean operator AND. Following basic fuzzy logic approach formulation, the rulebase is constructed according to principles of Boolean logical operators. The choice of a specific Boolean operator has an additional impact on how the truth values are aggregated in the defuzzification step. Each input that factors into a particular rule is associated with a membership in the membership function (m) for that rule, yet only one these membership values will be associated with the resulting truth output. Which of the membership values is associated with the truth value depends on the choice of the Boolean operator. The minimum (maximum) membership value will be selected using the Boolean operator AND (OR). Table 1 shows the rules and corresponding parameters used in this study, where b_0 is the intercept and $b_1$and $b_2$are the coefficients corresponding to F_{ST} scores and p-values, respectively.

The defuzzification step is the process used to identify the best crisp outcome given the fuzzy set. It consists of generating a single score (y) through the aggregation of the output of the different fuzzy sets.

$$y = \frac{\sum_{i=1}^{p} m_i z_i}{\sum_{i=1}^{p} m_i},$$

where p is the number of rules in the rulebase. It can be seen here how the construction of the rules and choice of Boolean operator ultimately influence the final output; AND will result in a particular value being weighted by the smallest membership among the inputs for that rule, making its potential contribution to the final score potentially quite low, while OR will result in it being weighted by the largest of the input

memberships, which can could result in a large contribution of a truth value even when most of the inputs have low memberships in the rulebase for that truth value. It would be reasonable to suggest that alternative strategies to collapsing the input memberships into a membership for the truth value could be used, for example an average, but this was not investigated in the present study.

Figure 4.2 shows the output space of the crisp fuzzy scores for the range of inputs observed in this study. It can be seen that the fuzzy algorithm is able to accommodate a highly nonlinear output space for the fuzzy scores despite the membership functions and rulebase having linear components.

F_{ST} scores and p-values were chosen as the preselection criteria and fuzzy inputs in this study due to the straightforwardness of calculating these statistics in a simulation environment, but it should be noted that in real data there are many additional criteria that have sound justification for being indicative of biological relevance and could potentially be integrated into a fuzzy model for preselection.

P-values were calculated jointly using a single-step GWAS approach with the BLUPF90 suite of programs as described in Aguilar et al. (2018), where genomic predictions are first calculated using a single-step GBLUP model and SNP effects are then jointly calculated from the genomic predictions,

$$\hat{\boldsymbol{a}} = \frac{1}{2\sum_{i=1}^{m} p_i q_i} \boldsymbol{Z}'\boldsymbol{G}^{-1}\hat{\boldsymbol{u}}$$

where $\boldsymbol{Z}'$ is the transpose of the $n \times m$ matrix of genotypes, $\boldsymbol{G}^{-1}$ is the inverse of the genomic relationship matrix, $\hat{\boldsymbol{u}}$ is the vector of solutions of genotyped individuals from a ssGBLUP model, and p_i and q_i

And SNP variances are calculated according to,

$$Var(\hat{a}_i) = \frac{1}{\sum_{i=1}^{m} 2p_i q_i} \boldsymbol{z}_i \boldsymbol{G}^{-1} (\boldsymbol{G}\sigma_u^2 - \boldsymbol{C}^{u_2 u_2}) \boldsymbol{G}^{-1} \boldsymbol{z}_i \frac{1}{\sum_{i=1}^{m} 2p_i q_i}$$

It should be noted that while the equivalency of GBLUP and SNPBLUP has been shown under certain assumptions (Habier et al. 2007), SNPBLUP and single-step GBLUP are not strictly equivalent due to the inclusion of information from non-genotyped animals in the latter.

F_{ST} scores were calculated as in Toghiani et al. (2017). With this strategy, subpopulations are formed from the extreme tails of the distribution of phenotypes. Groups of individuals with highly divergent phenotypes are expected on average to carry different alleles at loci that contribute variation to the phenotype of interest and F_{ST} is a natural measure of such differences.

$$F_{ST} = \frac{H_T - H_S}{H_T},$$

where $H_T = 2 \cdot p_T \cdot q_T$, $H_S = \frac{H_{S1} \cdot n_{S1} + H_{S2} \cdot n_{S2}}{n_{S1} + n_{S2}}$ and $H_{Si} = 2 \cdot p_{Si} \cdot q_{Si}$,

where p_T and q_T are major and minor allele frequencies, respectively, in the combined populations; p_{Si} and q_{Si} are major and minor allele frequencies, respectively, in the i^{th} subpopulation; and n_{Si} is the number of individuals in the i^{th} subpopulation.

While separate breeding populations are traditionally used for calculations of F_{ST} and it would seem natural to choose distinct breeding populations with divergent phenotypes (e.g., breeds with historically different selection goals), such an approach has the drawback of capturing deviations in allele frequencies at loci affecting all traits for which the populations differ on average. By grouping individuals from a single population by extreme phenotype, deviations in allele frequencies at unrelated variants will be driven only by correlated traits and random sampling.

Simulated datasets were generated using QMSim (Sargolzaei and Schenkel 2009) for three scenarios that differed in the number of causative loci contributing variation to the trait of interest (300, 1k, or 5k). Table 4.2 shows parameters specified for the simulation. Each simulation consisted of ten replicates and three sequential populations. First, a historical population under random mating was generated to establish mutation-drift balance and linkage disequilibrium. None of these individuals had data or pedigree recorded. For the second population, 500 males and 4,000 females were randomly selected from the historical population to be founders. This population underwent phenotypic selection and the number of dams retained expanded by 5% each generation. All individuals had data and pedigree recorded without error. From this population 100 males and 6,000 females were selected to be founders of the most recent population. It consisted of 5 non-overlapping generations under selection and 3,500 individuals per generation. Selection was based on pedigree EBVs. All individuals had data and pedigree and a random 5,000 individuals from generations 1 to 4 and 1,000 individuals from generation 5 had genotypes.

Preselection was based on random selection, p-values, F_{ST} scores, or fuzzy scores as criteria. The predictive power of preselected SNP panels of size 1, 2, 3, 4, 5, 10, 20, 30, 40, and 50k was evaluated using a ssGBLUP model as implemented in BLUPF90 (Aguilar et al. 2018).

The simulated genome consisted of 1.3 million SNPs randomly distributed across 29 chromosomes. There were either 300, 1k, or 5k causative variants randomly distributed across only 14 of the chromosomes, meaning that SNPs on 15 chromosomes did not segregate with any causative loci. The simulated genome was designed in such a

way to allow investigation of how noninformative SNPs affect genomic predictions and sensitivity of different preselection criteria.

Genomic prediction accuracies were calculated as the correlation between true ($\boldsymbol{u}$) and estimated ($\widehat{\boldsymbol{u}}$) breeding values. Bias of SNP associations was measured as the deviation of the regression coefficient of $\boldsymbol{u}$ on $\widehat{\boldsymbol{u}}$ ($\boldsymbol{b}_{1,u|\hat{u}}$) from one, with values less and greater than one indicating over- and underdispersion, respectively. If the estimated effect of a SNP (or its association with the trait) is not consistent from the training to the validation population, then its contribution to the validation genomic predictions will create variation that is not present in the TBVs and result in overdispersion of the estimates ($b_{1,\boldsymbol{u}|\widehat{\boldsymbol{u}}} < 1$).

The amount of genetic variance explained (GVE) by preselected SNPs was quantified as the variance of validation predictions, $var(\widehat{\boldsymbol{u}})$.

Genomic prediction accuracy can be expressed as a function of the regression of true on estimated breeding values, $b_{1,\boldsymbol{u}|\widehat{\boldsymbol{u}}}$, and the ratio of the standard deviations of $\boldsymbol{u}$ and $\widehat{\boldsymbol{u}}$,

$$\boldsymbol{cor}(\boldsymbol{u}, \widehat{\boldsymbol{u}}) = \boldsymbol{b}_{\boldsymbol{1},\boldsymbol{u}|\widehat{\boldsymbol{u}}} \frac{\boldsymbol{sd}(\widehat{\boldsymbol{u}})}{\boldsymbol{sd}(\boldsymbol{u})}. \quad \textbf{(4.1)}$$

Results

In the simulated genomes, only 14 of the 29 chromosomes harbored causative variants, leaving the SNPs on 15 of the chromosomes definitively unlinked with any causative loci. Figure 4.3 shows the proportion of SNPs that are linked (blue) or not linked (red) selected by each of the marker prioritization criteria. As expected, random

preselection results in the most unlinked markers for all genetic architectures. F_{ST} scores preselected more unlinked markers than p-values for all genetic architectures and SNP panel sizes. Fuzzy preselection resulted in the least number of unlinked markers for all preselected panels and genetic architectures, though the differences between p-value and fuzzy preselection are relatively modest. Interestingly, the advantage of the fuzzy scores relative to p-values in discriminating between linked and unlinked markers seems to increase as the genetic architecture increases in complexity, with the percent difference in number of linked markers selected by fuzzy scores ranging from 0.17 to 2.77, 0.67 to 2.57, and 2.49 to 5.40% higher than p-values for 300, 1k, and 5k causative variants, respectively.

It may be useful to further compare how markers that are and are not linked with causative variants are commonly (or not) selected across preselection criteria. Tables 4.3, 4.4, and 4.5 show the number of SNP markers that are unique to a particular criteria and the proportion of those markers that are linked with causative variants. Tables 4.6, 4.7, and 4.8 show the same information but for SNP that are shared between two or more criteria. Between 15 to 25, 14 to 32, and 16 to 31% of the SNPs preselected based on fuzzy scores for 300, 1k, 5k causative variants were not present in the equivalent-sized panels for F_{ST} scores and p-values and a higher proportion of these SNPs were linked with causative variants for all genetic architectures. As expected, a high proportion of the SNPs that were selected by all criteria for a given panel size were linked with causative variants; while the proportion linked decreased as the panel size increase due to an expected loss of sensitivity, the proportion linked was still considerably higher than among SNPs selected by only one or two of the criteria, with the notable exception of

markers that were selected by both F_{ST} scores and p-values but not fuzzy scores. Greater than 95% of SNPs were linked with causative variants for all genetic architectures for the latter case, indicating that there may be a more optimal set of fuzzy parameters that would include these SNPs in the fuzzy-preselected panels.

While it is not straightforward how to quantify the relevance of a particular SNP in a scenario where all markers are potentially linked and in LD with a causative variant(s), if the overlap in SNP selected by each criteria is similar to the scenarios where a portion of the SNPs are definitively not linked with any causative variants then it may be reasonable to expect that the fuzzy scores are robust across genetic architectures for preselection. Table 4.9 shows the number of SNPs unique to one criteria or shared across 2 or all criteria for a genetic architectures of 1k causative variants that are distributed across all 29 chromosomes. While this scenario is expected to be more realistic with regards to the genetic architecture of complex traits in real populations, it is not straightforward how to define the relevance of the preselected SNPs. However, it can be observed that the distribution of SNPs across each criteria in Table 4.9 is quite similar to that observed under the scenarios where a portion of the markers are definitively not linked with causative variants (and therefore have minimal relevancy for trait-specific predictions), so it may be reasonable to expect that the relevancy of SNPs selected under the more realistic genetic architectures may also be similar to what was measured under the less realistic genetic architectures.

Figure 4.4a presents the trends in accuracy for each of the preselection criteria across the different SNP panel sizes. For each of the genetic architectures, fuzzy score-based prioritization yielded higher accuracies compared to the other preselection criteria.

In fact, the accuracy using the fuzzy approach was up to 15.3, 7.9, and 2.4 (under the 5k causative variant scenario), 29.9, 8.9, and 3.5 (under the 1k causative variant scenario), and 44.5, 10.6, and 4.57% (under the 300 causative variant scenario) higher relative to random, F_{ST}, and p-value preselection methods, respectively. The largest advantage for fuzzy score preselection relative to random and F_{ST} preselection was when the size of the prioritized marker set (panel size) was 1k and this difference tended to decrease as the panel size grew with the exception of the scenarios of 5k causative variants for random preselection (fuzzy preselection accuracy was 0.5% lower) and 300 causative variants for F_{ST} score-based prioritization (fuzzy approach accuracy was 0.7% lower). Accuracies for p-value preselection did not exceed those of fuzzy preselection for any of the genetic architectures or evaluated panel sizes.

Figure 4.4b shows $b_{1,\boldsymbol{u}|\hat{\boldsymbol{u}}}$ on the horizontal axis, var($\hat{\boldsymbol{u}}$) on the vertical axis, and a gradient showing values of accuracy that correspond to the range of values. Additionally, the values of $b_{1,\boldsymbol{u}|\hat{\boldsymbol{u}}}$ and var($\hat{\boldsymbol{u}}$) that are observed for each preselection criteria and panel size are plotted to show how each preselection criteria moves through the accuracy parameter space as the SNP panel size increases. For all genetic architectures, random preselection increased the variance of the validation predictions slowly but the overdispersion decreased rapidly as the panel size increased. The opposite trend was observed for p-values; it tended to initially capture more variance in the predictions and increase this variance at a faster rate, but predictions based on p-value preselection are the most overdispersed. In the case of 300 causative variants, this overdispersion even increased, explaining why under this genetic architecture the p-value accuracy trends downward for panel sizes larger than 3k SNPs. F_{ST} scores tended to initially capture

similar genetic variation as random preselection but be less overdispersed; while this dispersion did not decrease as quickly as observed with random preselection with increasing SNP panel size, the variation of the predictions tended to increase more quickly, resulting in higher accuracies for F_{ST} scores under 300 and 1k but not 5k causative variants.

Similar trends in accuracies, $b_{1,\boldsymbol{u}|\hat{\boldsymbol{u}}}$, and var($\hat{\boldsymbol{u}}$) for fuzzy scores, F_{ST} scores, and p-values were observed under the more realistic scenario where causative variants were distributed across all chromosomes, as shown in Figure 4.5. Fuzzy preselection outperformed random, F_{ST}, and p-value preselection by up to 39.7, 7.8 and 3.8, respectively, for 300 causative variants; 24.6, 11.7 and 1.9 for 1k causative variants; and 6.15, 7.7 and 1.9 for 5k causative variants. When there were 5k causative variants, random preselection yielded notably better predictions than the other preselection criteria, outperforming fuzzy preselection for all panels larger than 1k SNPs by up to 5.3%.

Discussion

Much of the optimism around the use of high-density marker panels and whole genome sequence data was motivated by expectations that such data would vastly increase the percentage of the genetic variation captured and ultimately the accuracy of genomic selection. However, both simulation and real-data-based results have shown little benefit using such data in genomic predictions. Equation 4.1 shows that genomic prediction accuracies are influenced by 2 factors: 1) the proportion of the true genetic variance captured in the validation predictions, and 2) the dispersion of validation predictions relative to the true additive genomic values. Increasing the proportion of true

genetic variance captured will not necessarily improve genomic prediction accuracies if the validation predictions become more overdispersed ($b_{1,\boldsymbol{u}|\hat{\boldsymbol{u}}} < 1$).

This is evidenced by the trends of accuracy, GVE, and overdispersion under the 5k QTL genetic architecture. In such case, p-values based preselection yielded lower genomic prediction accuracies than all other criteria for SNP panels larger than 20k despite explaining the largest percentage of the genetic variation. In fact, random preselection explained the least genetic variation of all criteria yet yielded the highest accuracies for SNP panels larger than 10k. These results show that maximizing the genetic variation captured by a set of SNPs is not necessarily ideal if there is an accumulation of bias in the SNP associations, as p-value preselection has a tendency to do given that it consistently yielded the greatest overdispersion of all criteria evaluated.

The performance of random preselection relative to the non-random preselection criteria improved as the number of causative variants increased. It would be reasonable to expect that random preselection might capture more genetic variance by more evenly covering the genome, where non-random criteria might instead saturate regions with causative variants with the strongest signals and consequently fail to capture genetic variance from causative variants with a smaller contribution; However, Figure 4.2 shows that random preselection in fact captures the least amount of genetic variation across all genetic architectures.

As the complexity of the genetic architecture increases, the maximum contribution to the genetic variance of a single variant is reduced and the sensitivity of non-random preselection criteria to distinguish true from spurious signals becomes more limited.

P-values and F_{ST} scores followed opposing trends in accuracy, genetic variance explained, and SNP preselection bias. P-values tended to capture larger portion of the genetic variance but result in more overdispersed genomic predictions while F_{ST} captured consistently less genetic variation but yielded genomic predictions that were less overdispersed.

A primary objective of designing a composite preselection score using FIS was to increase accuracies relative to those of the individual preselection criteria used as inputs to the FIS. While fuzzy preselection did not universally yield higher accuracies than p-value and F_{ST}, there were certain points at which it performed better and this demonstrates that it is feasible to combine multiple criteria in an index to capture the strengths of each and improve predictions. Accomplishing this depends on preserving or increasing genetic variation captured while minimizing accumulation of bias in SNP associations.

Though ultimately none of the preselection criteria yielded better predictions than use of all 1.3 million SNPs, very similar accuracies were yielded using preselected panels of 50k SNPs and smaller, which may be useful in analyses when it is not feasible to routinely use SNP panels of such large dimension.

Although a high density panel was used in this simulation study (1.3 million markers), the reality of the genomic field in livestock and plant applications is that several million markers/variants are available from whole genome sequencing at different depths. This dramatic increase in the number of variants will undoubtably increase the accumulated bias in variant associations using single criteria for prioritization. The results of this study seem to indicate that an aggregated preselection criteria, such as the FIS

approach, may be adequate for such data, especially if further prioritization criteria such annotation, functional, and expression information associated with the considered SNPs/variants is available.

Conclusions

A large amount of sequence data is being generated in several livestock applications largely for potential use in genomic evaluations. Currently, such information did not lead to meaningful increase in accuracy compared to low/moderate commercial marker panels. Several marker prioritization were proposed but their performance was inconsistent across applications with slight to no improve in accuracy. In spite of our understanding of the relationship between the portion of the genetic variance explained by the preselected markers and the bias induced by the prioritization of false positives (markers non-linked to causal variants), we were unable to find the optimum balance between these two opposing forces to maximize accuracy. In this study, a fuzzy logic approach was presented to concatenate two prioritization criteria into a unique score to prioritize relevant markers with the expectation of reducing the number false positives tagged by each individual criteria separately. Under varying marker panel size and genetic complexity (number of causal variants), the FIS approach seems to perform generally better than the single criteria approaches. We believe that the performance of the FIS approach could be improved with the inclusion of already available additional prioritization information such as annotation, functional and expression data. The FIS approach presents a flexible method to combine noisy/uncertain information to derive reasonable decision-making outputs.

References

Aengchuan, P. and B. Phruksaphanrat (2015). "Comparison of fuzzy inference system (FIS), FIS with artificial neural networks (FIS + ANN) and FIS with adaptive neuro-fuzzy inference system (FIS + ANFIS) for inventory control." Journal of Intelligent Manufacturing **29**(4): 905-923.137

Aguilar, I., et al. (2018). BLUPF90 suite of programs for animal breeding with focus on genomics. Proceedings of the World Congress on Genetics Applied to Livestock Production, New Zealand.121

Chang, L. Y., et al. (2018). "High density marker panels, SNPs prioritizing and accuracy of genomic selection." BMC Genet **19**(1): 4.106

Du, F., et al. (2007). "Characterizing Linkage Disequilibrium in Pig Populations." International Journal of Bioogical Sciences **3**.139

Frischknecht, M., et al. (2018). "Short communication: Genomic prediction using imputed whole-genome sequence variants in Brown Swiss Cattle." Journal of Dairy Science **101**(2): 1292-1296.73

Habier, D., et al. (2007). "The impact of genetic relationship information on genome-assisted breeding values." Genetics **177**(4): 2389-2397.29

Ling, A. S., et al. (2021). "Dissection of the impact of prioritized QTL-linked and -unlinked SNP markers on the accuracy of genomic selection(1)." BMC Genom Data **22**(1): 26.129

Lucchese, L. V., et al. (2021). "Mamdani fuzzy inference systems and artificial neural networks for landslide susceptibility mapping." Natural Hazards **106**(3): 2381-2405.138

Meuwissen, T. and M. Goddard (2010). "Accurate prediction of genetic values for complex traits by whole-genome resequencing." Genetics **185**(2): 623-631.107

Ober, U., et al. (2012). "Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster." PLoS Genet **8**(5): e1002685.44

Sargolzaei, M. and F. S. Schenkel (2009). "QMSim: a large-scale genome simulator for livestock." Bioinformatics **25**(5): 680-681.119

Takagi, T. and M. Sugeno (1985). "Fuzzy Identification of Systems and its Applications to Modeling and Control." IEEE Transactions on Systems, Man, and Cybernetics **SMC-15**(1).130

Toghiani, S., et al. (2017). "Genomic differentiation as a tool for single nucleotide polymorphism prioritization for Genome wide association and phenotype prediction in livestock." Livestock Science **205**: 24-30.117

van Binsbergen, R., et al. (2015). "Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle." Genetics Selection Evolution **47**(1).60

Veerkamp, R. F., et al. (2016). "Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle." Genet Sel Evol **48**(1): 95.68

Xiang, R., et al. (2019). "Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits." Proceedings of the National Academy of Sciences **116**(39): 19398-19408.85

Zadeh, L. A. (1965). "Fuzzy Sets." Informatio and Control **8**: 338-353.131

Tables

Table 4.1 – Fuzzy rules and corresponding rule parameters

Input Category		Rule Parameters		
Fst	$-\log_{10}$(p-value)	b0	b1	b2
1	1	-25.68	15.32	8.20
1	2	15.36	-16.79	-9.49
1	3	17.66	-20.66	-5.43
2	1	-0.14	11.18	-6.28
2	2	-20.81	2.73	20.80
2	3	-21.33	-18.67	24.89
3	1	12.81	23.25	-7.86
3	2	-1.82	-5.71	-25.57
3	3	57.42	59.51	57.58

Table 4.2 Simulation parameters for QMSim

N QTL	300 or 1,000 or 5,000
N SNP	1,300,000
N Chromosomes	29
Have QTL	14 or 29
No QTL	15 or 0
Distribution of QTL Allele Effects	Gamma, shape = 0.4
Trait Heritability (h^2)	0.3
Historical Population (HP)	
Number of Generations	4,010
Number of Individuals in First 4,000 Generations (Males/Females)	400/400
Number of Individuals in Last Generation (Males/Females)	500/4,000
Population to Establish Linkage Disequilibrium (LDP)	Founders from last generation of HP
Number of Generations	20
Number of Founder Individuals (Males/Females)	500/4,000
Sire Replacement Rate	0.5
Dam Replacement Rate	0.2
Selection Design	Phenotype
Culling Design	Age
Mating Design	Random
Number of Progeny per Mating	1
Population Under Evaluation (EP)	Founders from second-to-last and last generations of LDP
Number of Generations	5
Number of Founder Individuals (Males/Females)	100/6,000
Sire Replacement Rate	0.5
Dam Replacement Rate	0.2
Selection Design	EBVs
Culling Design	Age
Mating Design	Random
Number of Progeny per Mating	1

Table 4.3 The number of unique SNPs and proportion of unique SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 5k causative variants

Panel Size (k)	Unique to One Criteria					
	F_{ST} Scores		p-values		Fuzzy Scores	
	N	Proportion Linked	N	Proportion Linked	N	Proportion Linked
1	640	0.780	383	0.823	163	0.857
2	1,323	0.753	899	0.773	527	0.824
3	2,006	0.732	1,466	0.748	921	0.808
4	2,621	0.712	2,046	0.734	1,266	0.792
5	3,121	0.699	2,621	0.720	1,486	0.781
10	4,849	0.636	5,560	0.681	1,958	0.738
20	8,980	0.588	10,936	0.646	3,400	0.689
30	13,113	0.564	15,893	0.622	4,727	0.650
40	17,150	0.551	20,581	0.604	5,928	0.626
50	21,075	0.540	25,062	0.591	7,033	0.604

Table 4.4 The number of unique SNPs and proportion of unique SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 1k causative variants

Panel Size (k)	Unique to One Criteria					
	F_{ST} Scores		p-values		Fuzzy Scores	
	N	Proportion Linked	N	Proportion Linked	N	Proportion Linked
1	641	0.759	226	0.844	190	0.899
2	1,345	0.713	631	0.769	543	0.932
3	2,043	0.686	1,182	0.752	964	0.803
4	2,630	0.666	1,738	0.736	1,271	0.777
5	3,111	0.647	2,288	0.723	1,463	0.762
10	4,800	0.584	5,173	0.677	1,896	0.719
20	9,114	0.550	10,373	0.631	3,434	0.665
30	13,390	0.533	15,210	0.605	4,838	0.628
40	17,408	0.520	19,815	0.589	5,987	0.611
50	21,284	0.512	24,240	0.577	7,019	0.595

Table 4.5 The number of unique SNPs and proportion of unique SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 300 causative variants

Panel Size (k)	Unique to One Criteria					
	F_{ST} Scores		p-values		Fuzzy Scores	
	N	Proportion Linked	N	Proportion Linked	N	Proportion Linked
1	534	0.791	235	0.884	157	0.893
2	1,115	0.729	601	0.786	398	0.805
3	1,707	0.695	1,108	0.762	710	0.784
4	2,264	0.668	1,639	0.739	989	0.773
5	2,752	0.645	2,201	0.718	1,218	0.759
10	4,813	0.585	5,037	0.666	1,947	0.705
20	9,044	0.539	10,398	0.618	3,446	0.643
30	13,445	0.523	15,732	0.591	4,971	0.618
40	17,660	0.513	20,132	0.575	6,319	0.598
50	21,738	0.507	24,695	0.562	7,511	0.581

Table 4.6 The number of shared SNPs and proportion of shared SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 5k causative variants

Panel Size (k)	Shared by Two Criteria						Shared by All Criteria	
	F_{ST} and p-values		F_{ST} and Fuzzy		p-values and Fuzzy			
	N	PL[a]	N	PL[a]	N	PL[a]	N	PL[a]
1	38	0.96	258	0.91	515	0.88	65	0.99
2	90	0.95	462	0.88	886	0.86	126	0.97
3	105	0.95	649	0.86	1,189	0.85	241	0.96
4	110	0.94	891	0.85	1,465	0.84	378	0.93
5	109	0.92	1,244	0.83	1,745	0.83	526	0.92
10	81	0.93	3,683	0.77	2,972	0.80	1,388	0.88
20	52	0.95	7,587	0.71	5,631	0.75	3,381	0.84
30	35	0.96	11,201	0.68	8,421	0.71	5,651	0.81
40	31	0.96	14,684	0.66	11,254	0.69	8,135	0.79
50	29	0.96	18,058	0.64	14,071	0.67	10,838	0.77

[a]The proportion of shared SNPs linked with QTL

Table 4.7 The number of shared SNPs and proportion of shared SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 1k causative variants

Panel Size (k)	Shared by Two Criteria						Shared by All Criteria	
	F_{ST} and p-values		F_{ST} and Fuzzy		p-values and Fuzzy			
	N	PL[a]	N	PL[a]	N	PL[a]	N	PL[a]
1	100	0.99	136	0.92	551	0.92	123	0.99
2	169	0.97	257	0.86	971	0.90	230	0.98
3	177	0.97	396	0.83	1,257	0.88	384	0.96
4	172	0.96	640	0.80	1,532	0.86	558	0.95
5	160	0.96	984	0.77	1,806	0.84	746	0.94
10	107	0.97	3,385	0.72	3,012	0.79	1,708	0.89
20	85	0.97	7,024	0.67	5,764	0.74	3,778	0.84
30	78	0.98	10,450	0.64	8,630	0.71	6,083	0.80
40	72	0.98	13,900	0.63	11,492	0.68	8,620	0.77
50	69	0.98	17,290	0.61	14,334	0.67	11,357	0.75

[a]The proportion of shared SNPs linked with QTL

Table 4.8 The number of shared SNPs and proportion of shared SNPs that are linked with causative variants for F_{ST} score, p-value, and fuzzy score preselection under a genetic architecture of 300 causative variants

Panel Size (k)	Shared by Two Criteria						Shared by All Criteria	
	F_{ST} and p-values		F_{ST} and Fuzzy		p-values and Fuzzy			
	N	PL[a]	N	PL[a]	N	PL[a]	N	PL[a]
1	88	0.99	165	0.95	464	0.94	214	0.99
2	165	0.99	368	0.89	882	0.92	353	0.99
3	189	0.98	587	0.84	1,186	0.90	517	0.98
4	199	0.98	848	0.81	1,474	0.88	689	0.96
5	202	0.97	1,186	0.79	1,737	0.86	860	0.95
10	155	0.97	3,246	0.71	3,022	0.81	1,786	0.90
20	117	0.99	7,069	0.67	5,716	0.74	3,770	0.84
30	108	0.99	10,508	0.64	8,582	0.70	5,939	0.81
40	103	0.99	13,916	0.62	11,444	0.67	8,321	0.77
50	100	0.99	17,284	0.60	14,327	0.65	10,878	0.75

[a]The proportion of shared SNPs linked with QTL

Table 4.9 The number of SNPs unique to one criteria or shared across two or more criteria for a genetic architecture of 1k causative variants distributed across all chromosomes

	Unique to One Criteria			Selected by Two Criteria			
Panel Size (k)	F_{ST} Scores	p-values	Fuzzy Scores	F_{ST} Scores and p-values	F_{ST} and Fuzzy Scores	p-values and Fuzzy Scores	Selected by All Criteria
1	711	205	172	80	113	619	97
2	1,420	592	498	140	235	1,062	206
3	2,099	1,119	872	147	394	1,374	361
4	2,662	1,672	1,165	151	658	1,647	530
5	3,137	2,232	1,366	146	1,011	1,917	706
10	4,815	5,075	1,789	92	3,378	3,118	1,714
20	8,937	10,110	3,133	73	7,050	5,877	3,940
30	12,957	14,736	4,300	66	10,502	8,723	6,475
40	16,826	19,136	5,333	63	13,867	11,557	9,244
50	20,500	23,347	6,213	60	17,194	14,348	12,245

Figures

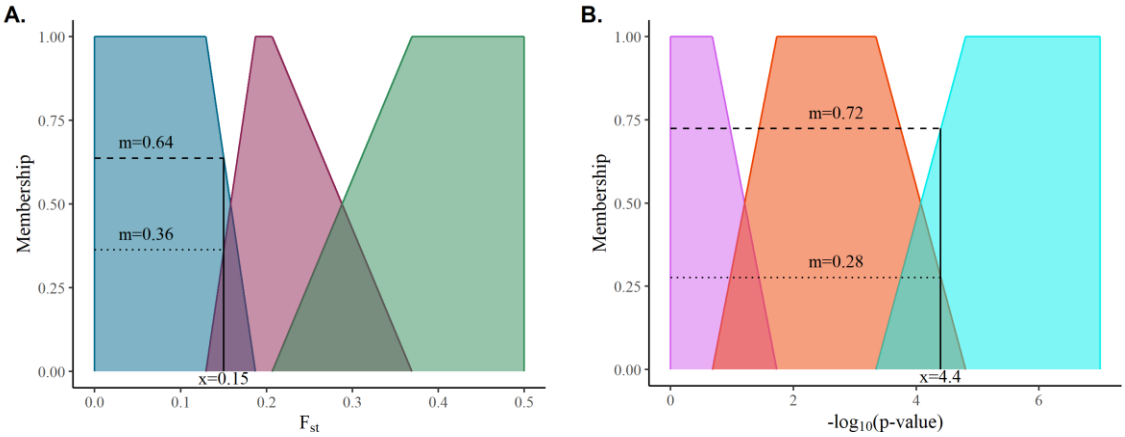

Figure 4.1 Membership functions for a) F_{ST} scores and b) $-\log_{10}$(p-value)

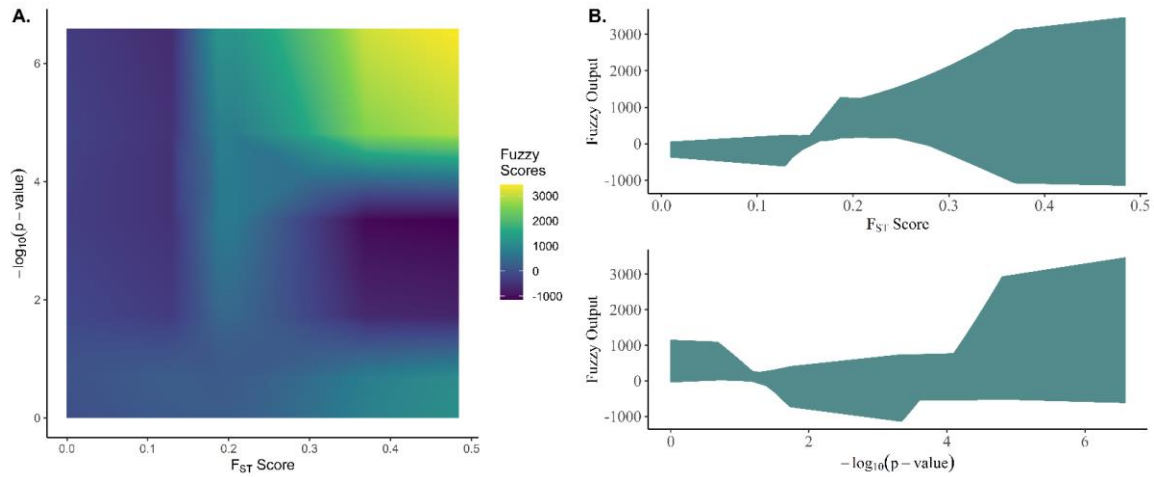

Figure 4.2 The fuzzy output space for the observed range of F_{ST} scores and $-\log_{10}$(p-value)

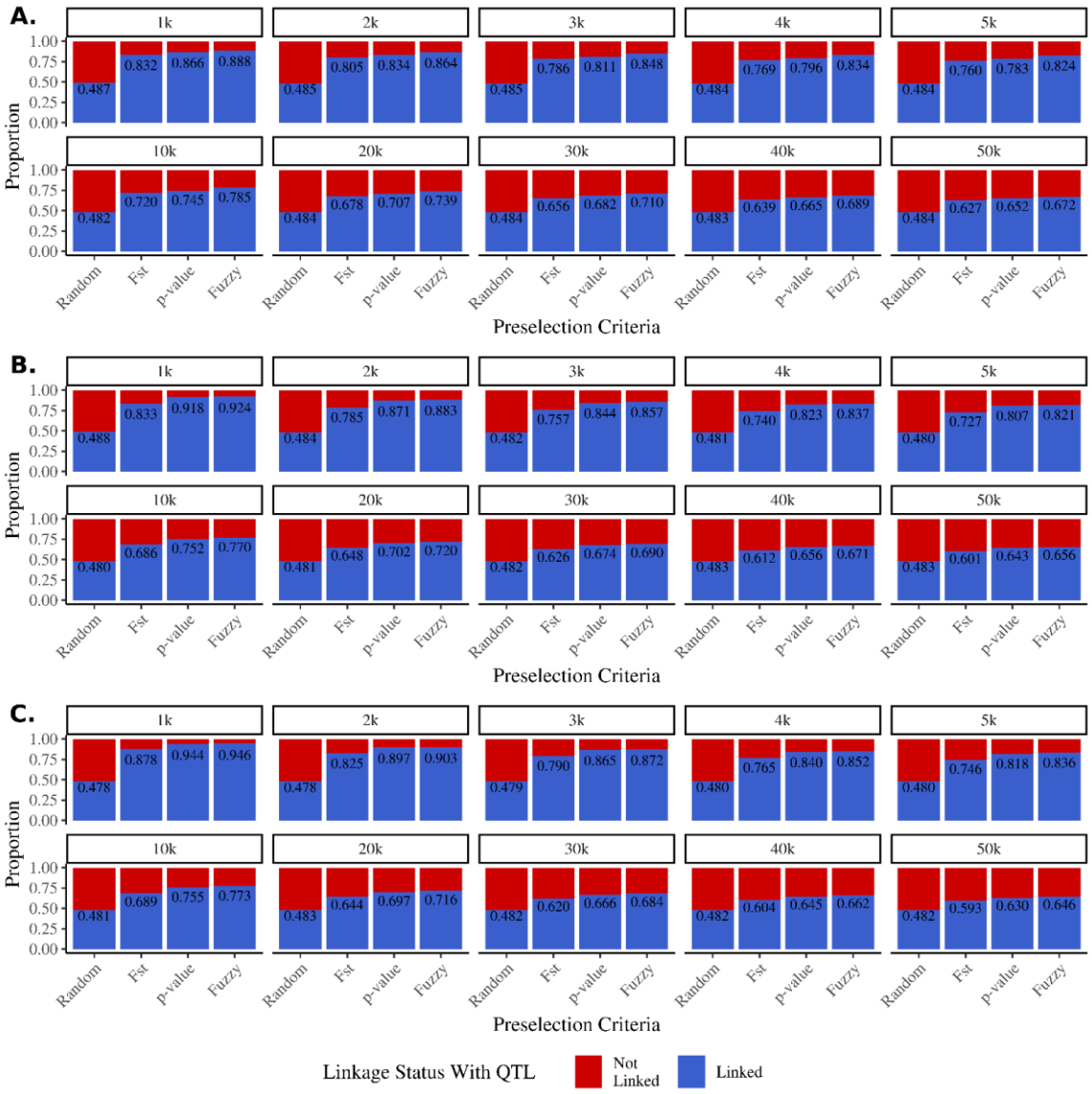

Figure 4.3 Proportion of SNPs linked with causative variants selected by random, F_{ST} scores, p-values, or fuzzy scores

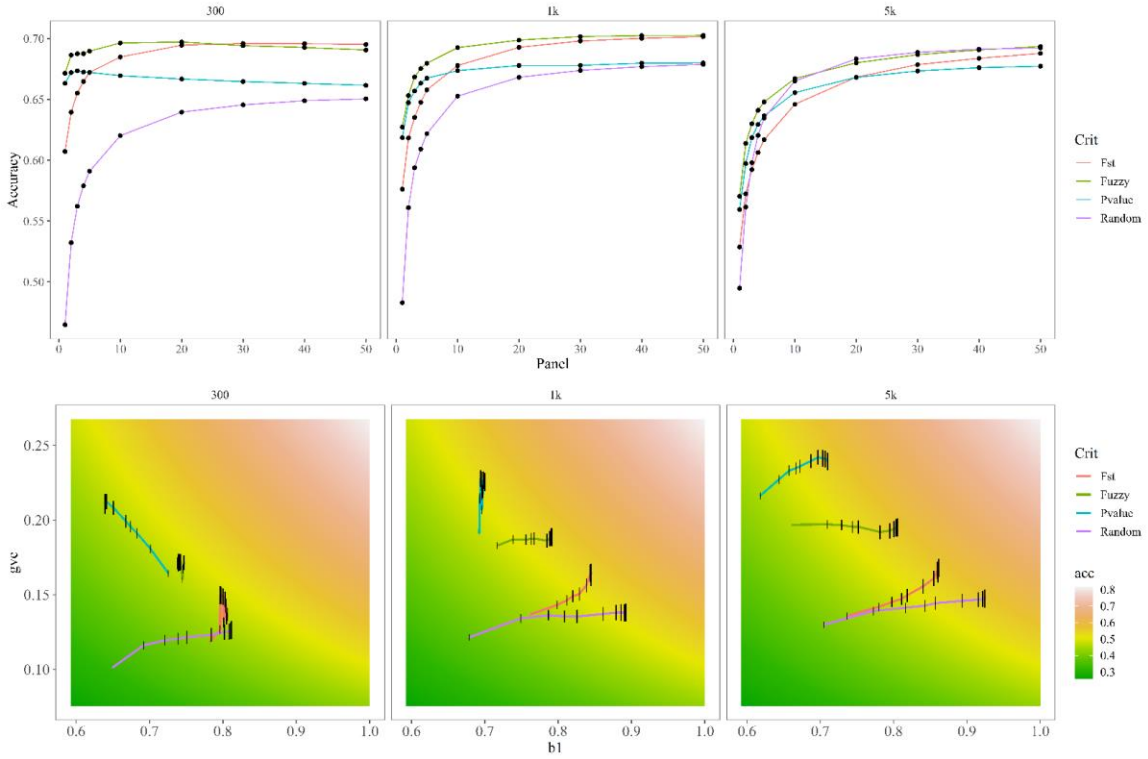

Figure 4.4 Trends in accuracy, dispersion, and variance of EBVs for random, F_{ST} score, p-value, or fuzzy score preselection when only 14 of 29 chromosomes harbor causative variants

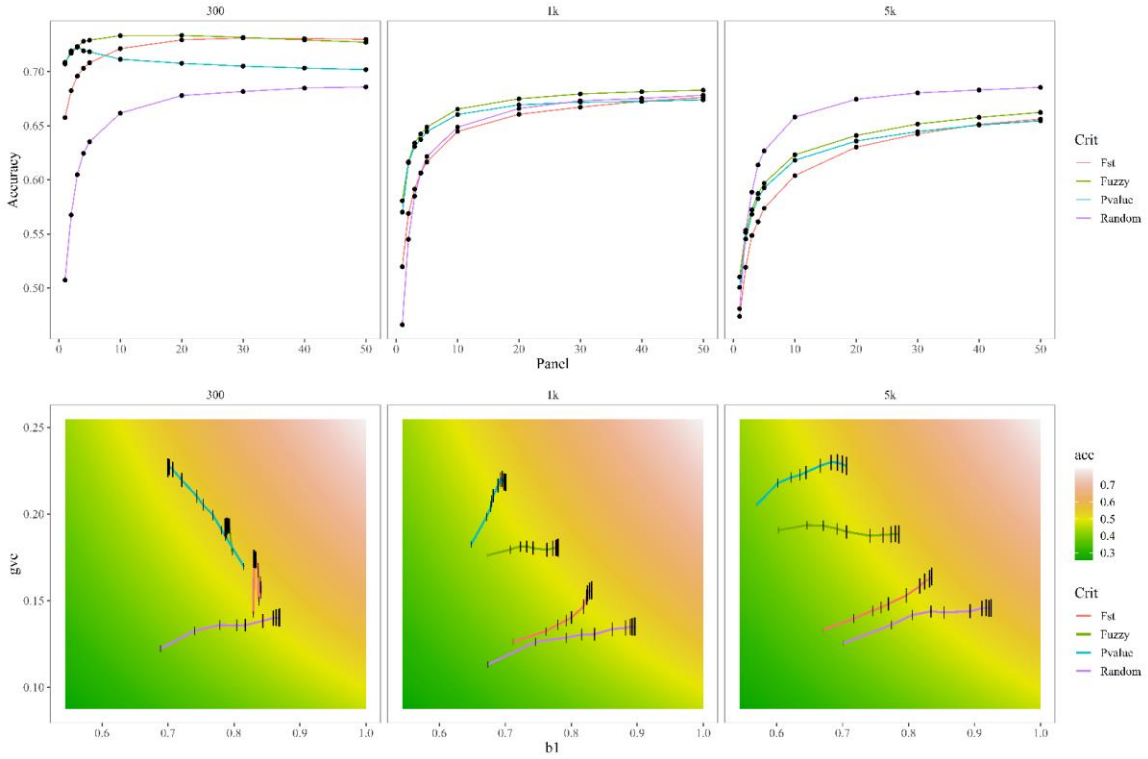

Figure 4.5 Trends in accuracy, dispersion, and variance of EBVs for random, F_{ST} score, p-value, or fuzzy score preselection when all chromosomes harbor causative variants

CHAPTER 5

AN ALGORITHM FOR FINE-TUNING OF FUZZY INFERENCE SYSTEM PARAMETERS FOR VARIANT PRESELECTION AND PRIORITIZATION IN GENOMIC SELECTION[3]

[3] Ling, A.S., E.H. Hay, S.E. Aggrey, and R. Rekaya. To be submitted to *Genetics*.

Abstract

Fuzzy inference systems may offer an effective strategy for combining criteria that reflect SNP relevance into a composite score for more accurate ranking of SNPs for preselection for use in genomic predictions. However, construction of the fuzzy inference system requires specifying several subjective parameters and it is not clear how to estimate the optimal values as "true" measures of SNP relevance cannot be known. In this study, an algorithm for identifying the optimal fuzzy inference system parameters based on maximizing distance between true and null distributions of the input criteria (F_{ST} scores and p-values) to the fuzzy inference system was proposed and evaluated using simulated data with variation in the parameters for heritability (0.1, 0.3, 0.6) and sampling distribution of allelic effects of causative variants (uniform, standard normal, gamma with shape parameter of 0.4 or 4.0). Null distributions of the input criteria were approximated by calculating them based on randomized phenotypes. A random walk was performed through the fuzzy parameter space and the set of parameters that maximized true and null distribution distance based on difference of means, signal-noise ratio, or a t-test statistic were further evaluated for ranking and preselection of SNPs and genomic prediction accuracies. Genomic prediction accuracies based on the optimized fuzzy preselection in general did not outperform preselection based on the individual input criteria or fuzzy preselection using a general set of fuzzy parameters. The algorithm as implemented here is likely complicated by a random walk not being able to fully explore the parameter space of the fuzzy parameters due to high dimensionality. Furthermore, it

may be possible to identify better metrics for optimizing the parameters than the distance metrics evaluated here.

Introduction

Since 2009, the US Holstein population has seen significant gains in genetic progress coinciding with the incorporation of genome-wide genotypes into routine evaluations (VanRaden et al. 2009) and there is evidence of similar trends for several other species and breeds under genomic selection (Abdollahi-Arpanahi et al. 2021). This increase in the rate of genetic progress can be partly attributed to a better modeling of the Mendelian sampling variance relative to the expected additive relationships based on pedigree due to linkage disequilibrium (LD) between single nucleotide polymorphism (SNP) markers and causative loci (Goddard and Hayes 2007).

High-density (HD) and sequence data, which has become more widely available, are expected to increase the number of causative loci or/and the SNPs in high linkage disequilibrium (LD) with causative loci genotyped (Druet et al. 2013). However, efforts to determine which genotyped variants are causal or in high LD with QTL have achieved limited success, particularly in the context of genomic selection, as the majority of causative variants are expected to explain only an exceedingly small portion of the total genetic variance (Hayes and Goddard 2010). Consequently, many studies have endeavored to increase the amount of genetic variance captured and the number of causative variants tagged by markers or genotyped by including all available variants in a prediction model that either fits all predictors simultaneously or performs intrinsic variable selection (van Binsbergen et al. 2015, Heidaritabar et al. 2016, Meuwissen et al. 2021).

It is plausible that the vast majority of SNPs/variants present in HD genotypes and whole-genome sequence data are poorly informative and/or redundant regarding inheritance of alleles at causal loci. This is problematic due to its impact on the overparameterization of the association models leading to overfitting as small effects are attributed to variants with largely spurious and inconsistent association with the trait of interest and a decrease in the accuracy of genomic predictions (Ling et al. 2021).

Several variable selection methods that use a mixture distribution to discriminate between variants of high and low predictive benefit have been proposed (Meuwissen et al. 2001, Habier et al. 2011, Moser et al. 2015). As suggested by Chang et al. (2018), the success of these methods depends on both the genetic variance captured and the bias of the individual predictions (reflected in the dispersion of the estimated relative to the true breeding values). Preselection based on non-random criteria may capture more genetic variance with fewer markers, but, as shown in the previous chapter, tends to increase the bias/dispersion of predictions due to accumulation of false positives.

Fuzzy inference systems (FIS) (Zadeh 1965, Takagi and Sugeno 1985) are a group of flexible and dynamic algorithms for mapping multiple inputs to an output space under conditions of uncertainty regarding input classification and relationships between inputs. A FIS approach was shown in the previous chapter to have the potential to strengthen true signals and decrease rates of false positives of SNP-trait associations by combining multiple and varied genomic variant preselection criteria into a composite score. This was demonstrated using a Sugeno-type FIS and the preselection statistics p-values and F_{ST} scores as inputs to derive a single preselection score that yielded better genomic prediction accuracies under certain parameters.

In the previous chapter, a single set of FIS parameters was identified that performed robustly and was used across all the evaluated scenarios for fuzzy preselection. However, it would be preferable to be able to adjust a set of FIS parameters to suit the unique characteristics of a particular dataset and to optimize marker preselection. In engineering applications where FIS are routinely used, the desired outputs of the FIS tend to be explicitly known and this makes estimation of the most suitable FIS parameters using least squares or supervised ANN-like estimation procedures straightforward. In the case of SNP preselection, causal loci are not known and deciding how to define the relevance of a particular marker is not clear. This makes the use of traditional strategies for the determination of the optimum of FIS parameters problematic and alternative approaches need to be evaluated to accomplish the same goal. The overall main of the present study is to develop a new strategy for optimizing FIS parameters to a dataset that requires no knowledge of causal loci and is generalizable to any dataset typical of livestock genetic improvement programs.

Materials and Methods

The program QMSim (Sargolzaei and Schenkel 2009) was used to generate a series of simulated datasets designed to mimic a cattle genome and population under selection. Several scenarios that varied in the heritability (0.1, 0.3, or 0.6) of the trait of interest or the sampling distribution (gamma, normal, or uniform) of the allelic effects of causative loci were analyzed. Ten replicates per scenario were examined to average over sampling variation of the results. A full summary of the simulation parameters used in this study is shown in Table 5.1.

Calculation of p-values was performed using PREGSF90 from the BLUPF90 suite of programs (Aguilar et al. 2018). First, animal solutions ($\hat{\boldsymbol{u}}$) were calculated using a single-step GBLUP model (Aguilar et al. 2010, Christensen and Lund 2010). SNP effects ($\hat{\boldsymbol{a}}$) were calculated from $\hat{\boldsymbol{u}}$ based on the following,

$$\hat{\boldsymbol{a}} = \frac{1}{2\sum_{i=1}^{m} p_i q_i} \boldsymbol{Z}'\boldsymbol{G}^{-1}\hat{\boldsymbol{u}}$$

where $\boldsymbol{Z}'$ is the transpose of the $n \times m$ matrix of genotypes, $\boldsymbol{G}^{-1}$ is the inverse of the genomic relationship matrix, $\hat{\boldsymbol{u}}$ is the vector of solutions of genotyped individuals from a ssGBLUP model, and p_i and q_i are allele frequencies of the i^{th} SNP. SNP variances were calculated according to,

$$Var(\hat{a}_i) = \frac{1}{\sum_{i=1}^{m} 2p_i q_i} \boldsymbol{z}_i \boldsymbol{G}^{-1}(\boldsymbol{G}\sigma_u^2 - \boldsymbol{C}^{u_2 u_2})\boldsymbol{G}^{-1}\boldsymbol{z}_i \frac{1}{\sum_{i=1}^{m} 2p_i q_i}.$$

F_{ST} scores were calculated as in Toghiani et al. (2017) and Chang et al. (2018). Individuals were ranked based on their records for the phenotype of interest and the 2.5% highest- and 2.5% lowest-ranked individuals were selected to form subpopulations H_{S1} and H_{S2}, respectively.

$$F_{ST} = \frac{H_T - H_S}{H_T},$$

where $H_T = 2 \cdot p_T \cdot q_T$, $H_S = \frac{H_{S1} \cdot n_{S1} + H_{S2} \cdot n_{S2}}{n_{S1} + n_{S2}}$ and $H_{Si} = 2 \cdot p_{Si} \cdot q_{Si}$,

where p_T and q_T are major and minor allele frequencies, respectively, in the combined populations; p_{Si} and q_{Si} are major and minor allele frequencies, respectively, in the i^{th} subpopulation; and n_{Si} is the number of individuals in the i^{th} subpopulation.

A Sugeno-type FIS was used to combine F_{ST} scores and p-values into a composite index for preselection of SNP markers.

A random walk search was conducted through the FIS parameter space for number of classes, boundaries of the membership functions, and coefficients of the rule base. At each iteration, fuzzy scores were calculated based on the current parameters and these scores were evaluated based on how well they discriminated between two sets of inputs, one set calculated based on true phenotypes and a second set of equal size that was calculated based on randomized phenotypes. As both F_{ST} scores and p-values rely on capturing associations between phenotypes and genotypes, it was expected that the set calculated based on randomized phenotypes would capture random variation based on these statistics due to spurious associations and noise and could be used as a null distribution to identify false positives among the true set.

Two statistics were presently evaluated as criteria for measuring distance between the null and true distributions of the fuzzy scores, a two-sample Student's *t*-test statistic and a modification of the signal-to-noise ratio.

The Student's *t*-test statistic for two samples is used as a significance measure of the difference between the means of two samples.

$$t = \frac{\bar{x}_{True} - \bar{x}_{Null}}{\sqrt{\frac{s^2_{True}}{n_{True}} + \frac{s^2_{Null}}{n_{Null}}}}.$$

Rather than as a test of significance, the resulting *t* statistics were used to rank the FIS parameter sets and the set with the largest *t* statistic was used to calculate fuzzy scores for preselection.

The second statistic used to evaluate the FIS parameter sets was a modification of the signal-to-noise ratio used in some engineering applications to measure the background noise relative to true signals,

$$D = \frac{\mu_{True} - \mu_{Null}}{\sigma_{Null}}.$$

As the difference between the mean of the true and null distributions gets larger and the standard deviation of the null distribution gets smaller, the value of *D* will increase.

Results

Successful tuning of the FIS parameters using the proposed approach depends on the marginal and joint distributions of the null and the true input sets that can systematically discriminate between them. Figure 5.1a shows the scaled densities for F_{ST} scores and p-values calculated based on randomized (red) and true (blue) phenotypes. For both statistics, there is a clear shift in the density based on true phenotypes towards higher values relative to the scenario when the phenotypes were randomized.

The joint distributions of F_{ST} scores and p-values based on randomized and true phenotypes are shown in Figure 5.1b. The distribution based on true phenotypes shows greater dispersion than the distribution based on randomized phenotypes. While some portion of the SNPs from the true phenotype distribution clearly fall outside the boundaries of the joint randomized distribution, these SNPs alone are unlikely to account for enough genetic variance to yield reasonable prediction accuracy. The dashed red lines show the lower boundaries of the top 50k SNPs selected by F_{ST} scores (vertical) and p-values (horizontal) using the true phenotype set. This clearly shows that the values of the top SNPs from the true set overlap largely with the distribution of the randomized set. Though only one replicate is shown here, this is typical of all the replicates and scenarios evaluated in the present study.

Figure 5.2 shows two potential fuzzy output spaces relative to the F_{ST} score and p-value inputs and where the top 5k SNPs (shown in red) selected under each fuzzy set fall

within the expected space. The overlap across the top 5k SNPs selected by the fuzzy sets is 12.6%. The selected SNPs yielded accuracies of 0.69, 0.65, and 0.70 for difference of mean, SNR, and T-Test optimization methods, respectively. The horizontal and vertical dashed red lines indicate the lower boundary for the top 5k SNPs selected by p-values and F_{ST} scores, respectively. It can be seen that the FIS algorithm is highly flexible in producing a complex fuzzy output space that can select quite disparate sets of SNPs. The trend in accuracies using SNPs preselected based on difference-of-means-, SNR-, and T-Test-optimized fuzzy scores are shown in Figures 5.3, 5.4, and 5.5, respectively. The trends in the optimized fuzzy preselection were compared to accuracies obtained when SNPs were preselected based on random, Fst, p-value, and classic fuzzy approach (using the generalized FIS parameters) methods. For all scenarios considered, the optimized fuzzy preselection yielded comparable accuracies to the other preselection criteria but in general did not outperform preselection based on F_{ST} scores or even the random approach in some cases. Consistent with previously observed trends (Ling et al. 2021), p-values tended to yield lower accuracies than alternative preselection criteria as the panel size increased, likely due to an accumulation of poorly informative SNPs that are spuriously associated with the trait of interest. Preselection based on the optimized fuzzy parameters tended to yield higher accuracies than F_{ST} preselection for smaller panels and comparable accuracies for larger panels. This indicates that despite ultimately not yielding higher accuracies than alternative preselection methods, the fuzzy scores are able to accumulate information from both inputs to strengthen true signals that capture genetic variance while excluding spurious associations that contribute to the slower gains in accuracy with increasing panel size that p-values suffer from.

The only scenario under which the optimized FIS parameters outperformed the general fuzzy parameters was when the heritability was equal to 0.1. As heritability decreases, signals from causative variants will be confounded and the dispersion of the true relative to the null distribution will decrease. Under such conditions, a general set of FIS parameters may be less robust and fine-tuning the parameters through optimization may become more advantageous.

Figure 5.6 presents the comparison across the different optimized fuzzy preselection approaches. Optimization based on difference of means seemed to be more advantageous relative to SNR and T-Test methods as the heritability decreased and the effect-sampling distribution favored fewer causative variants having an individually large contribution to the genetic variance. When allelic effects of causative variants were sampled from a gamma distribution with shape parameter 0.4, optimization based on difference of means initially yielded higher accuracies than SNR and T-Test optimization for a heritability of 0.1, but this advantage was lost with increasing panel sizes and SNR and T-Test optimization yielded higher or equal accuracies under heritabilities of 0.3 and 0.6.

Discussion

The systematic differences in the overlap between the distributions of p-values and F_{ST} scores calculated based on true and randomized phenotypes show an incentive for fine-tuning the FIS parameters to better use the characteristics of a particular dataset. While simulation offers the advantage of knowing the true underlying differences in genetic and population parameters between datasets, such information is often non available and at best can be approximated. Being able to adjust the FIS model to suit the

characteristics of a particular dataset without having explicit knowledge of the underlying causes of these characteristics would be ideal for yielding optimal predictions in practical applications.

The flexibility of FIS in generating highly non-linear fuzzy output spaces, as shown in this study, makes this method a good candidate for blending individual preselection criteria such as p-values and F_{ST} scores into a composite criteria with the potential to amplify the strengths of each while balancing their weaknesses. In theory, a set of FIS parameters should exist that maximizes some measure of difference between fuzzy scores based on true and randomized input criteria to identify poorly informative variants among the true set and allow reduction in the parameter space of the prediction model without significant loss of information.

The functions evaluated in the present study to measure differences between the null and true distributions did not succeed in optimizing the FIS parameters as they pertain to increasing prediction accuracies through preselection. This can likely be contributed to at least two factors. First, it is likely that maximizing distance between the null and true distributions is just one facet that should be considered when optimizing the FIS parameters; a metric that reflects the genetic variance that would be captured by the top markers for a given fuzzy set should be additionally considered. Second, a random walk through the FIS parameter space is not ideal despite the efficiency of the Sugeno FIS. Even under the relatively simple configurations of the FIS that were considered here, full exploration of the parameter space is effectively impossible under a random walk and such a strategy makes identifying the optimal parameters for fuzzy preselection exceedingly improbable.

Conclusions

Despite the failure to increase accuracies relative to alternate preselection strategies, the optimized fuzzy preselection did strike a balance between the performance of the inputs under most parameters. A non-supervised learning process within a parametric or non-parametric (e.g., neural networks) could provide a robust approach to optimize the input parameters of the FIS approach. There is an extensive literature in using these methods in optimization problems and that will be a research venue to explore in the future.

References

Abdollahi-Arpanahi, R., et al. (2021). "Detecting effective starting point of genomic selection by divergent trends from best linear unbiased prediction and single-step genomic best linear unbiased prediction in pigs, beef cattle, and broilers." J Anim Sci **99**(9).132

Aguilar, I., et al. (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): 743-752.38

Aguilar, I., et al. (2018). BLUPF90 suite of programs for animal breeding with focus on genomics. Proceedings of the World Congress on Genetics Applied to Livestock Production, New Zealand.121

Chang, L. Y., et al. (2018). "High density marker panels, SNPs prioritizing and accuracy of genomic selection." BMC Genet **19**(1): 4.106

Christensen, O. F. and M. S. Lund (2010). "Genomic prediction when some animals are not genotyped." Genetics Selection Evolution **42**(2).39

Druet, T., et al. (2013). "Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions." Heredity **112**(1): 39-47.50

Goddard, M. E. and B. J. Hayes (2007). "Genomic selection." Journal of Animal Breeding and Genetics **124**: 323–330.133

Habier, D., et al. (2011). "Extension of the bayesian alphabet for genomic selection." BMC Bioinformatics **12**(186).135

Hayes, B. and M. Goddard (2010). "Genome-wide association and genomic selection in animal breeding." Genome **53**(11): 876-883.134

Heidaritabar, M., et al. (2016). "Accuracy of genomic prediction using imputed whole-genome sequence data in white layers." J Anim Breed Genet **133**(3): 167-179.63

Ling, A. S., et al. (2021). "Dissection of the impact of prioritized QTL-linked and -unlinked SNP markers on the accuracy of genomic selection(1)." BMC Genom Data **22**(1): 26.129

Meuwissen, T., et al. (2021). "On the use of whole-genome sequence data for across-breed genomic prediction and fine-scale mapping of QTL." Genet Sel Evol **53**(1): 19.96

Meuwissen, T. H. E., et al. (2001). "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." Genetics **157**: 1819–1829.19

Moser, G., et al. (2015). "Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model." PLoS Genet **11**(4): e1004969.136

Sargolzaei, M. and F. S. Schenkel (2009). "QMSim: a large-scale genome simulator for livestock." Bioinformatics **25**(5): 680-681.119

Takagi, T. and M. Sugeno (1985). "Fuzzy Identification of Systems and its Applications to Modeling and Control." IEEE Transactions on Systems, Man, and Cybernetics **SMC-15**(1).130

Toghiani, S., et al. (2017). "Genomic differentiation as a tool for single nucleotide polymorphism prioritization for Genome wide association and phenotype prediction in livestock." Livestock Science **205**: 24-30.117

van Binsbergen, R., et al. (2015). "Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle." Genetics Selection Evolution **47**(1).60

VanRaden, P. M., et al. (2009). "Invited review: reliability of genomic predictions for North American Holstein bulls." J Dairy Sci **92**(1): 16-24.37

Zadeh, L. A. (1965). "Fuzzy Sets." Informatio and Control **8**: 338-353.131

Tables

Table 5.1 Simulation parameters for QMSim

N QTL	1,000
N SNP	1,300,000
N Chromosomes	29
Distribution of QTL Allele Effects	Gamma, shape = 0.4 Gamma, shape = 4.0 Normal(0,1) Uniform
Trait Heritability (h^2)	0.1 or 0.3 or 0.6
Historical Population (HP)	
Number of Generations	4,010
Number of Individuals in First 4,000 Generations (Males/Females)	400/400
Number of Individuals in Last Generation (Males/Females)	500/4,000
Population to Establish Linkage Disequilibrium (LDP)	Founders from last generation of HP
Number of Generations	20
Number of Founder Individuals (Males/Females)	500/4,000
Sire Replacement Rate	0.5
Dam Replacement Rate	0.2
Selection Design	Phenotype
Culling Design	Age
Mating Design	Random
Number of Progeny per Mating	1
Population Under Evaluation (EP)	Founders from second-to-last and last generations of LDP
Number of Generations	5
Number of Founder Individuals (Males/Females)	100/6,000
Sire Replacement Rate	0.5
Dam Replacement Rate	0.2
Selection Design	EBVs
Culling Design	Age
Mating Design	Random
Number of Progeny per Mating	1

Figures

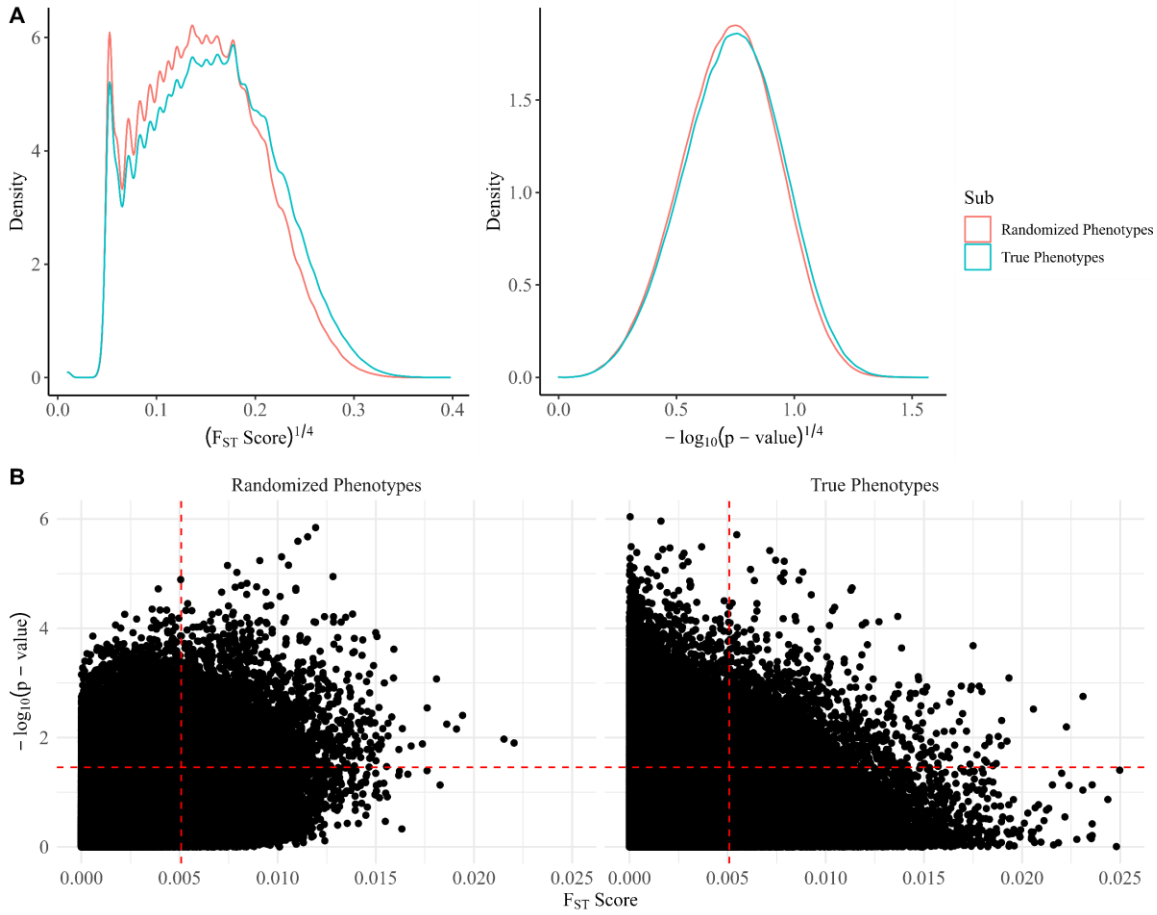

Figure 5.1 A comparison of the distributions of F_{ST} scores and p-values calculated based on true and randomized phenotypes: (a) Densities of scaled F_{ST} scores and p-values calculated based on true (blue) and randomized (red) phenotypes. (b) Scatterplot of paired F_{ST} scores and p-values calculated using randomized and true phenotypes. The red dashed lines show lower thresholds for the top 50k SNPs based on F_{ST} score (vertical) or p-value (horizontal) preselection.

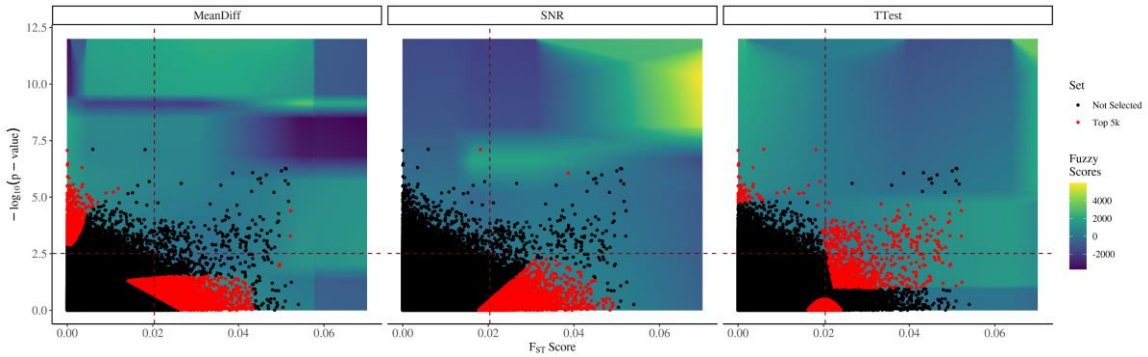

Figure 5.2 Fuzzy output spaces and selected SNPs for FIS parameter optimization based on difference of means, signal noise ratio (SNR), and T-Test statistics. Scatterplot of F_{ST} scores (horizontal axis) and p-values (vertical axis). The top 5k SNPs selected based on difference of mean, SNR, or T-Test FIS parameter optimization are shown in red. The background gradients show the fuzzy output space for each FIS parameter optimization method.

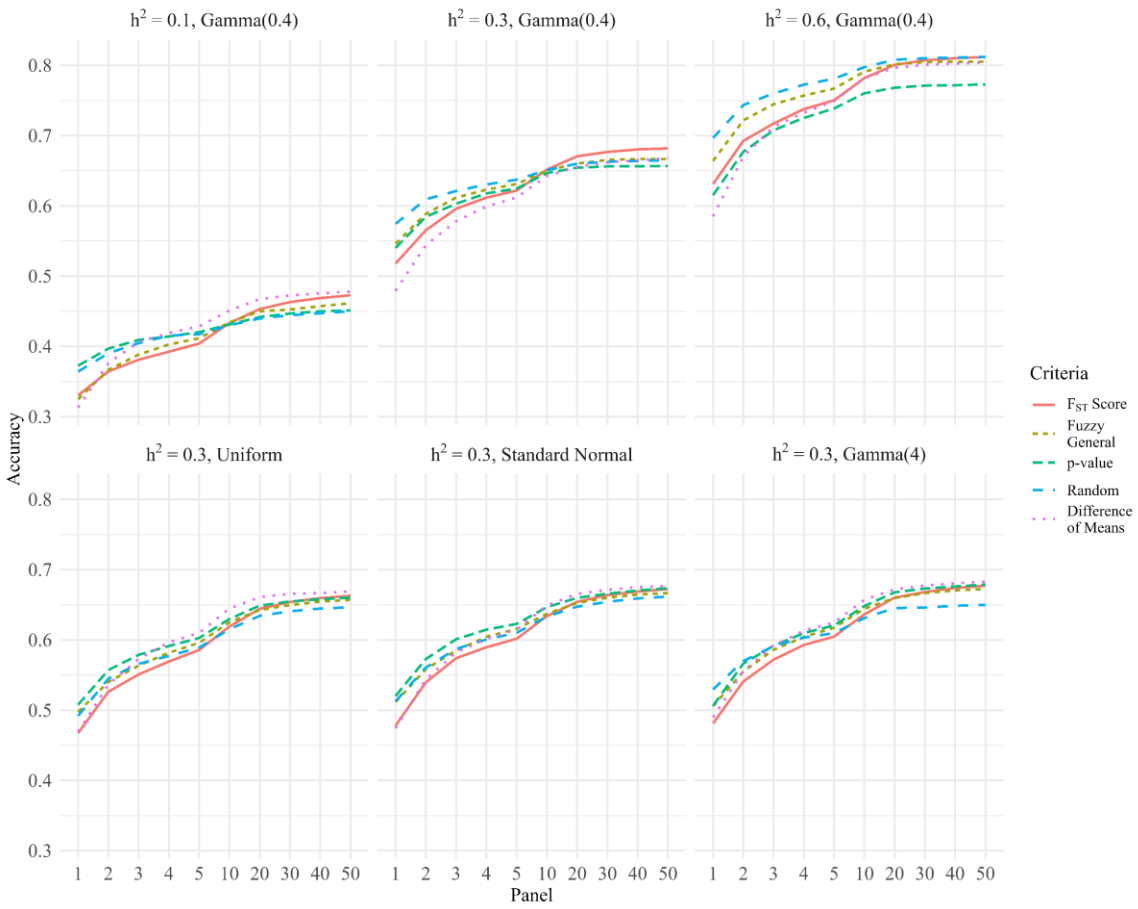

Figure 5.3 The trend in accuracies for preselection based on difference-of-mean-optimized fuzzy scores compared to preselection based on random, F_{ST} scores, p-values, or general fuzzy scores for 6 scenarios varying across heritability and sampling distribution for allelic effects.

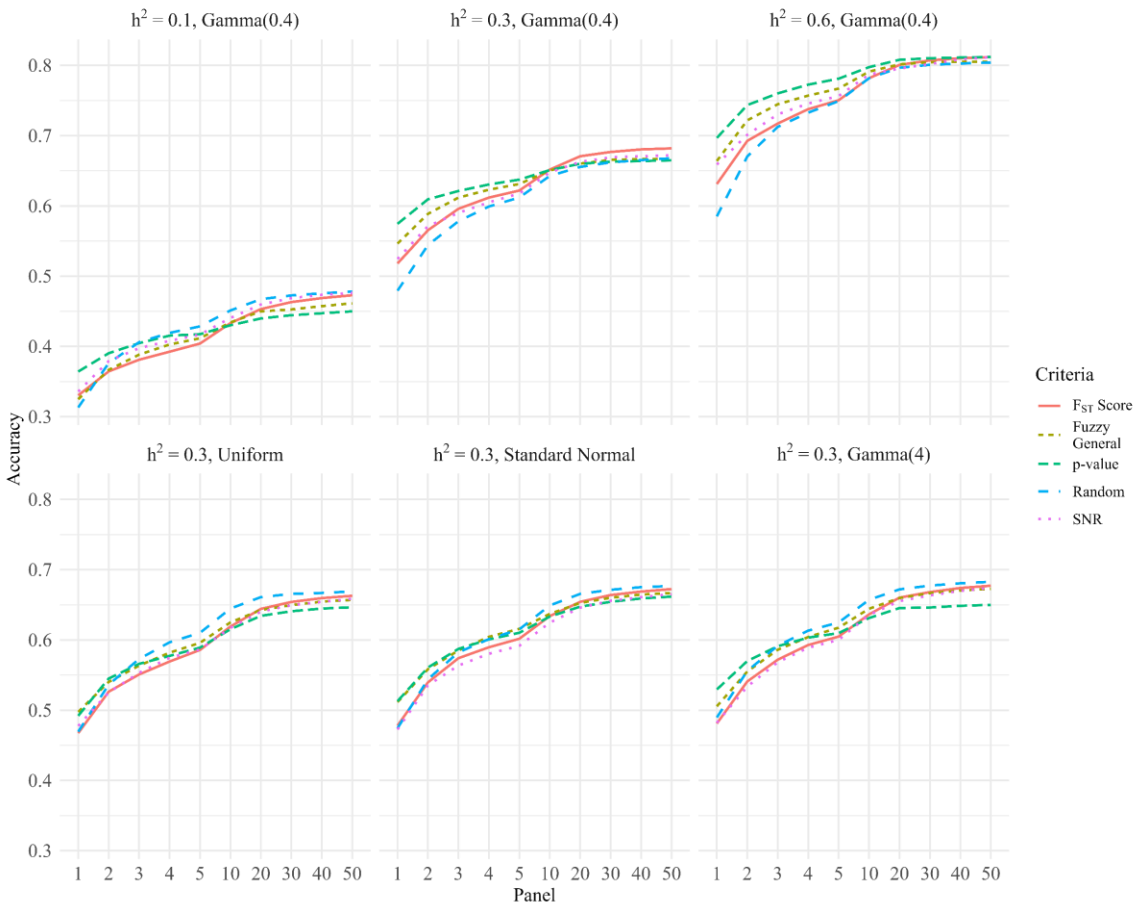

Figure 5.4 The trend in accuracies for preselection based on SNR-optimized fuzzy scores compared to preselection based on random, F_{ST} scores, p-values, or general fuzzy scores for 6 scenarios varying across heritability and sampling distribution for allelic effects.

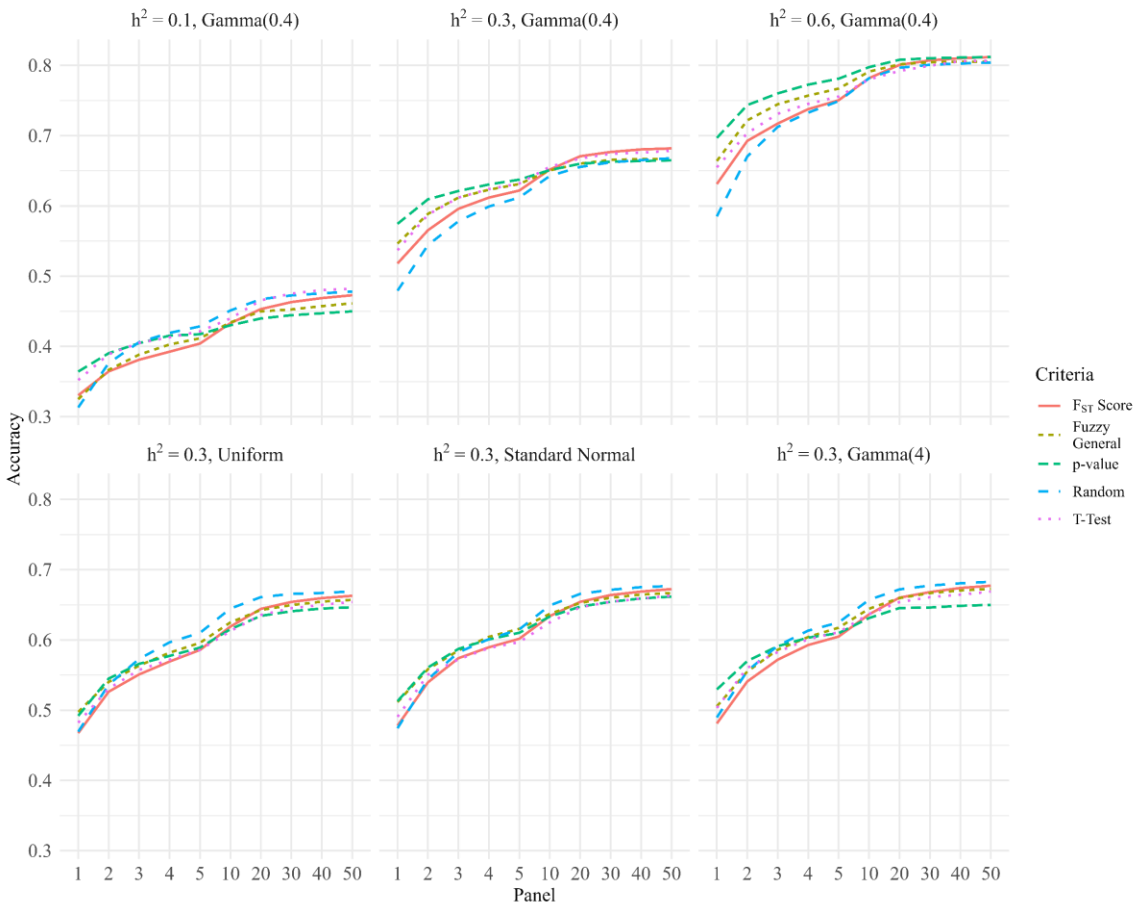

Figure 5.5 The trend in accuracies for preselection based on T-Test-optimized fuzzy scores compared to preselection based on random, F_{ST} scores, p-values, or general fuzzy scores for 6 scenarios varying across heritability and sampling distribution for allelic effects.

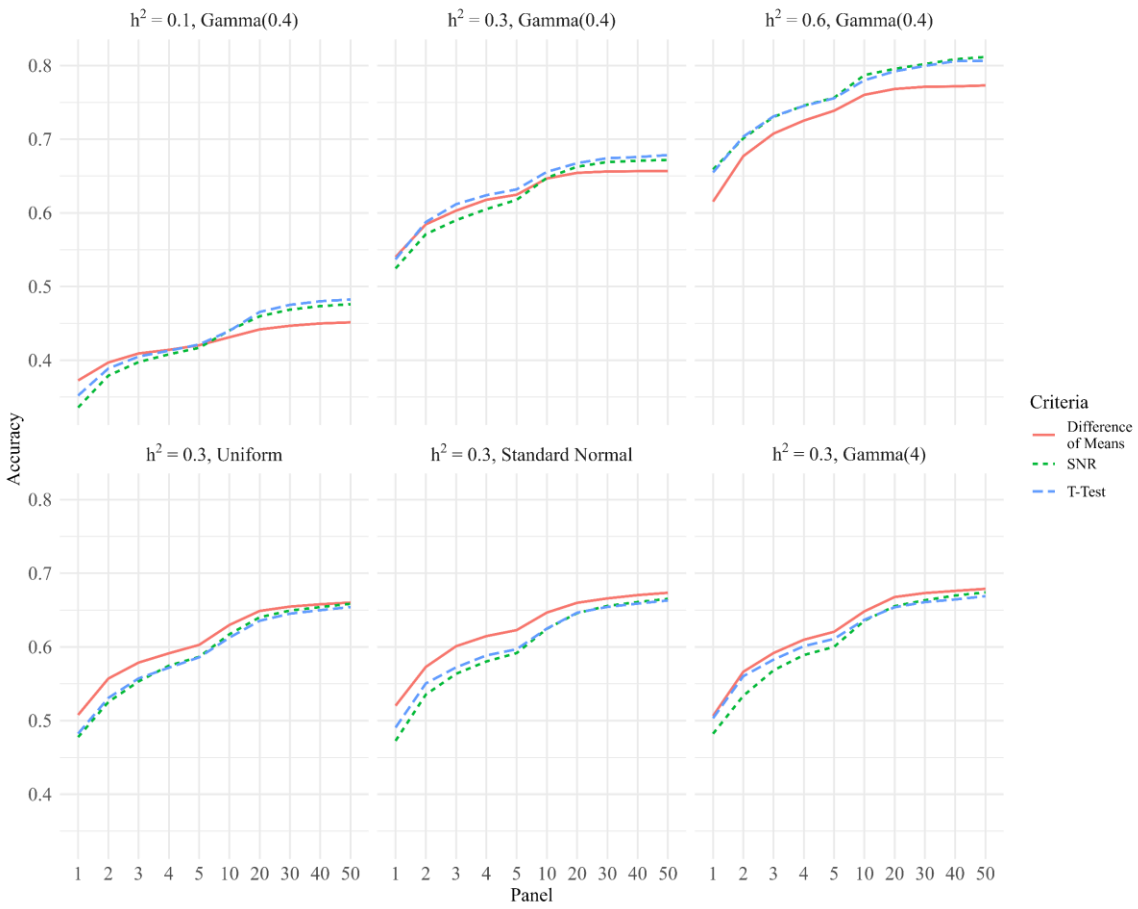

Figure 5.6 Comparison of the trend in accuracies for preselection based on different FIS parameter optimization strategies (difference of means, SNR, or T-Test) for 6 scenarios varying across heritability and sampling distribution for allelic effects.

CHAPTER 6

CONCLUSIONS

The use of genomic information in the genetic evaluations has undoubtedly led to increases in prediction accuracy and rate of genetic gain in livestock and plant applications. Furthermore, it has provided unique opportunities for better understanding of the genetic basis of complex traits and to devise optimized management programs. However, these gains in the accuracy of predictions tend to stagnate at relatively low-density marker panels that are unlikely to contain many, if any, of the causative variants. While the lack of further improvement in accuracy with the use of HD marker panels and whole genome sequence data is generally considered a consequence of the strong LD structure present in many commercial populations, we showed in this dissertation work that the presence of a large number of nonrelevant markers can lead to similar behavior of prediction accuracies as the marker density increases. As expected, the removal of nonrelevant markers has the potential to significantly increase the prediction accuracy by preventing attribution of QTL effects to markers that have an inconsistent association with functional variants.

The population statistic F_{ST} was shown to be an efficient criterion for preselection of trait-relevant markers that can improve modeling of QTL similarity between individuals, increase prediction accuracies, and maintain more stable prediction accuracies than

comparative association statistics like absolute estimated marker effects. While association statistics are powerful tools for identifying loci associated with a particular trait, disentangling spurious associations from weak but true signals is often not possible within the constraints of the data available. We showed that the more persistent prediction accuracy using F_{ST}-score-prioritized markers was the result of the ability of the F_{ST}-score-based method to select unlinked markers with weaker spurious associations with the trait compared to preselection based on absolute estimated effects. While this study only explored F_{ST} scores as an independent preselection statistic, we showed that F_{ST} and absolute estimated effect approaches preselected highly correlated sets of markers linked with QTL but significantly less so among unlinked markers. This highlights the potential for the possibility of combining F_{ST} scores and association statistics into a powerful preselection index.

Several marker prioritization approaches were proposed but their performance was inconsistent across applications with slight to no improvement in accuracy. In spite of our understanding of the relationship between the portion of the genetic variance explained by the preselected markers and the bias induced by the prioritization of false positives (markers non-linked to causal variants), we were unable to find the optimum balance between these two opposing forces to maximize accuracy. We proposed a fuzzy logic approach to concatenate two prioritization criteria into a unique score to prioritize relevant

markers with the expectation of reducing the number false positives tagged by each individual criteria separately. Under varying marker panel size and genetic complexity (number of causal variants), the fuzzy inference system approach seems to perform generally better than the single criteria approaches. Additionally, we developed a new method to optimize the parameters of the fuzzy inference system. We believe that the performance of the fuzzy approach could be improved with the inclusion of already available additional prioritization information such as annotation, functional and expression data. The fuzzy approach presents a flexible method to combine noisy/uncertain information to derive reasonable decision-making outputs.