DISSERTATION

GENETIC SELECTION FOR RESISTANCE TO BOVINE RESPIRATORY DISEASE USING POOLED DNA APPROACHES.

Submitted by

Ryan J. Boldt

Department of Animal Sciences

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2025

Doctoral Committee:

Advisor: R. Mark Enns
Co-Advisor: Scott Speidel

John Keele
Larry Kuehn
Tara McDaneld
Tim Holt

ABSTRACT

GENETIC SELECTION FOR RESISTANCE TO BOVINE RESPIRATORY DISEASE USING POOLED DNA APPROACHES

Bovine Respiratory Disease (BRD) is the costliest disease that affects the beef cattle industry. However, the only methods that are currently available to reduce the incidence of the microbial organisms (viruses and bacteria) that cause BRD are vaccination and antibiotic treatment. Examples using other species and diseases have shown that the selection for resistance to disease is an effective method to reduce the economic burden of that disease on the industry. Due to the challenge of collection of phenotypes for a trait like BRD resistance, one of the best methods for selection could be genomic selection. To try and capture a representative sample of commercial genetic makeup of the beef industry, samples for the study were collected from a commercial harvest facilities. To reduce overall genotyping costs, samples were genotyped using a pooled DNA approach. While pooled DNA has been used previously to identify genomic regions that differentiate based on disease status, this has not been done for animals that showed symptoms for BRD during the post weaning period. Therefore, the objectives of this research were to, 1) examine different analysis techniques for pooled DNA information, and 2) identify across breed SNP that are significant for identifying animals more likely to develop clinical signs of BRD.

To investigate the first objective of the dissertation, two separate analyses were done. The first analysis evaluated the number of SNPs used to calculate a genomic relationship matrix. While using DNA pooling does reduce the cost of genotyping by grouping samples, the cost could potentially be further reduced by using SNP chips with lower density. For the analysis, 106 pools

comprised of 96 individuals each were genotyped using a high-density genomic panel that contained 777,962 SNP. To evaluate the use of lower density SNP chip on pooled DNA analyses, 50 replications of number of SNP from 500 to 770,000 were sampled randomly. For each level and replication, the resulting genomic relationship matrix was compared to the full relationship matrix calculated from 776,749 SNP, after individual SNP were removed for minor allele frequency <0.05. To calculate the equivalence of the matrices, the genomic relationship matrix calculated from the reduced number of SNP was multiplied by the Eigenvalues and Eigenvectors of the genomic relationship matrix formed from all SNP. After this multiplication, the variance of the Eigenvalues of the reduced matrix was standardized by the full matrix variance of the Eigenvalues of the resulting matrix was calculated. The closer the resulting variance is to 0 both matrices were considered to be proportional to one another. When examining the resulting Eigenvalues variances after 2,000 SNP the reduction of variance decreased in magnitude. These results suggest that a low-density panel may be used for pooled DNA data and for calculating genomic relationship matrices.

The second analysis that was conducted to address the first objective looked at alternative analysis techniques for identification of simulated important SNP at varying levels of allelic prevalence and effect size. For the analysis, 100 random SNP across all chromosomes were selected to act as the significant SNP among the approximately 770,000 SNP available on the BovineHD chip. All SNP pooling allele frequencies (PAF) were simulated using a beta distribution. For the 100 significant SNP, the PAF were then modified based on differing levels of prevalence and the effect that the disease-causing SNP would have. For prevalence levels from 0.10 to 0.90, increments of 0.10 were simulated and for effect of the SNP values from 0.01 to 0.50 were simulated in increments of 0.01. For each of the 450 combinations of prevalence and effect,

two different models were applied to the same dataset. The first model type was a GWAS analysis that has previously been applied to this data type. Under this model each SNP is tested via an F-test. The dependent variable for this analysis was the PAF and the fixed effect was a binary classification of if a pool was a case or a control. Additionally, a relationship matrix was calculated to account for any population stratification that was occurring in the simulated dataset. For each F-test, a p-value was calculated. The second type of analysis that was conducted was a Random Forrest analysis. For the Random Forrest the same number of trees, terminal node size, and number of explanatory variables to try at each node were applied to all combinations. The optimal number was determined to be 2,000 trees, a terminal node size of 1, and to try 60,000 explanatory variables. For each of the combinations the results were ranked based on lowest p-value and highest variable importance factor for the GWAS and Random Forrest analysis, respectively. From there, the top 100 most significant SNP were compared, and the number of pre-identified significant SNP were counted within the subset. Across all levels of prevalence each model was able to identify a subset of the most significant SNP. Across all levels of prevalence, the Random Forrest model started? identifying significant SNP at lower levels of effect of the disease-causing allele. Random Forest model started identifying significant SNP at lower levels of the disease-causing allele. At low (0.10, 0.20, 0.30) and high levels (0.70, 0.80, 0.90) prevalence levels the traditional GWAS model was able to identify a higher number of significant SNP at high effect levels. Whereas at moderate prevalence levels (0.40, 0.50, 0.60) the Random Forest model more correctly identified a larger number of the significant SNP.

To address objective two, several analyses were run looking at estimating SNP effects to identify informative variants for selection against development of BRDC. For this analysis samples were collected from three large commercial processing plants in Colorado and Nebraska. DNA

samples were collected from ears when the animals were harvested. Samples for the study were collected over a four-year period. For pooling, punches were removed from each ear, and animals were sorted into either a case or control pool. Within each individual pool 96 animals were represented. For each case a corresponding control from the same group from the feedlot was also collected. In total 106 pools were constructed representing 10,176 animals across all pools with a matching case and control strategy. DNA was extracted using a Quigen Kit and pools were sent to Neogen (Lincoln, NE) for genotyping on a Bovine SNP chip that contained approximately 770,000 individual SNP. For each SNP and each pool, a PAF was calculated. To account for population stratification in the analysis a covariance matrix among pools, PAF was calculated. Mixed model methodology was used to solve for effects in the model. In the first analysis, each individual SNP was examined. For each individual SNP an F-test was performed to test for significance. Additionally, analyses were performed using SNP groups. SNP groups were formed using 100, 500, and 1,000 SNP regions. For each region a distance matrix based on the PAF for SNPs in the region was calculated. This was then used as a response variable for an ANOVA analysis. Fixed effects were the A matrix to account for population stratification as well as 2 x 106 matrix to signify if an animal was either in a case or control pool.

For all analysis types, no significant SNP were discovered. Additionally, several regions that have been previously reported to be significantly associated with BRDC in previous studies were also examined. To see if similar signal was being picked up, SNP were ranked from being estimated as the most significant to least significant and compared to previous results. Among the previously reported results there were regions on BTA16 (70-71), BTA16 (70-71), BTA14 (9-10), and BTA8 (63-64) that were among the top 1% of most significant SNP in the single SNP analyses. However, in the grouped SNP analyses none of these regions were in the top 1% of significant

SNP. Other regions that have been previously identified in other papers were either not in the top 1% of SNP in any analysis or had p-values that were 0.85 or greater.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND OBJECTIVES

1.1 Introduction

One of the diseases that has the largest economic impact on the beef cattle industry is Bovine Respiratory Disease Complex (**BRDC**) (Griffin, 2014). This disease can affect animals of all age and weight classes, but animals entering the feedlot are most susceptible. For this reason, another commonly used terminology for the condition is Shipping Fever. Considerable effort has been invested in decreasing the incidence of this disease, however, the rate of infection has not reduced.

One of the challenges with attempts to reduce the incidence of this disease is its multifactorial nature. There are multiple viruses and bacteria that interact with each other and with environmental factors to cause disease progression. To further complicate the situation, many of these pathogens are ubiquitous in the respiratory tract of cattle. The disease will progress during times of stress due to depression of the immune system resulting in viral pathogens proliferating to create additional suppression of the immune system. This allows for the bacterial pathogens to replicate leading to disease symptoms. The most commonly identified symptoms for animals experiencing BRDC are labored breathing, cough, nasal and or ocular discharge, depression, appetite depression, fever, and death (Duff and Galyean, 2007).

Considerable effort and research have been devoted to the development of improved management and pharmaceutical practices to reduce the incidence of BRDC. The average cost

to treat incidences of BRDC is estimated at $23.60 per treatment (APHIS, 2013). Given that, on average, 16.2% of animals are treated in feedlots with capacity of 1,000 head or more, the cost of the treatment of the disease is a major consideration (NAHMS 2011). In addition to the cost of treatment, there are also additional costs associated with losses in production. Animals that are infected with BRDC have reduced production efficiency the impacts which have been estimated to cost the beef industry over a billion dollars annually (Griffin 2014).

Beyond the economic burden that this disease complex adds to the beef industry, there are also social aspects associated with improving disease resistance. Beef consumers are becoming increasingly concerned with animal welfare. Globally, a growing expectation is that a larger percentage of animals will be raised without the use of antimicrobial and therapeutic drugs. This viewpoint is largely driven by the concept that the use of these tools in animal production are leading to increases in antibiotic resistance in human medicine (Zhou et al., 2020).

With the growing concern and lack of improvement using previous managerial based approaches, genetic selection for increased disease resistance is a viable option to reduce the incidence of BRDC. Previous research has illustrated that selection for increased disease resistance in beef, dairy, sheep, poultry, and swine industries can have successful outcomes (Kuhnlein, Ni et al. 1997, Bishop and Morris 2007, Cardoso et al. 2015, Popescu et al. 2016, Prather et al. 2017, Martin et al. 2018). These examples provide the evidence that inclusion of traits that improve health are worth including in modern breeding objectives. Given the economic costs associated with BRDC, it should be considered an economically relevant trait.

1.2 Objective

The objectives of this research were to, 1) examine different techniques for analyzing pooled DNA information, and 2) identify across breed SNP that are significant for identifying animals more likely to develop clinical signs of BRDC.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

Bovine Respiratory Disease Complex (BRDC) is the single most costly disease that affects beef cattle production. It has been estimated that the annual cost in treatment and lost productivity is over one billion annually (Griffin, 2014). Cattle of all ages and classes can be affected with BRDC, however, the cattle entering the feedlot system are most affected. Therefore, another term used for animals experiencing symptoms of BRDC is Shipping Fever.

From an economic perspective, there are many factors that contribute to the economic impact this disease has on the beef industry. It is estimated that of all cattle that enter a feedlot with a minimum of 1,000 head capacity, 16.2% of the cattle developed respiratory disease (NAHMS, 2011). Required treatment for symptomatic cattle for BRDC has been previously reported as $23.60 per case, on average (APHIS, 2013). The treatment cost for BRDC is comparative to interstitial pneumonia ($21.70 per case) and central nervous system diseases ($20.10 per case), but more expensive than treatments for lameness ($13.40 per case) and digestive problems ($9.90 per case) (APHIS, 2013). Given the level of incidence and cost of treatment, BRDC is expensive to treat compared to other commonly treated ailments.

Cost of treatment is not the only consideration when accounting for the impact of BRDC on the beef industry. Estimates include the expense due to loss of performance for animals exhibiting symptoms of the disease and the financial consequence of death loss, however, there is strong evidence that many animals with BRDC but do not show physical symptoms. These

nonsystematic animals would be expected to have some loss in performance that would lead to economic inefficiencies (Griffin, 2014).

One of the biggest challenges when studying BRDC is that there are multiple viruses and bacteria that are commonly associated with disease progression. A common mode of respiratory disease progression is an initial infection via a viral pathogen. The most common viral pathogens associated with initial infection includes the following: Bovine herpesvirus-1 (BVH-1), parainfluenza virus-3 (PI-3), bovine respiratory syncytial virus (BRSV), and bovine viral diarrhea virus (BVDV), Bovine Corona Virus (BCV) (Panciera and Confer, 2010, Fulton et. al, 2011). These viruses do not generally cause symptoms in the infected animal but do facilitate a depression in the immune system that leads to a secondary bacterial infection which causes disease symptoms. The bacteria that are commonly associate with respiratory disease include the following: *Mannheimia haemolytica, Pasturella multocida, Histophilus somni, Truperella pyogenes, Mycoplasma bovis,* and *Bibersteninia trehalosi* (Panciera and Confer, 2010). Each of these bacteria are ubiquitous in the nasal passages of cattle (Allen et al., 1992, Fulton et al., 2002). Therefore, it is often difficult to prescreen animals for these viruses and bacteria to diagnosis their risk for developing BRDC.

Traditionally, it has been thought that stressful events such as transportation, weather, commingling, dehydration, hypoxia, and metabolic disturbances are major factors promoting development of BRDC (Irwin et al., 1979, Taylor et al., 2010). These different factors then interact with the normal viral and bacterial flora contributing to the multi-factorial nature of the disease complex. Generally, the animal is initially infected with one of the viral pathogens. These pathogens do not always cause symptoms of respiratory disease, instead it weakens the host's immune system allowing a predisposition for bacterial infection. This can be accomplished in two

ways: 1) the viral agents cause damage to respiratory clearance mechanisms and lung parenchyma, allowing for the bacteria to move from the upper respiratory tract and into the lungs (Taylor et al., 2010); or 2) the viral agents interfere with the immune system's ability to respond to the bacterial infection (Martin and Bohac, 1986, Czuprynski et al., 2004).

One of the most common methods to identify animals as being infected with any of the pathogens in BRDC is through observation of symptoms. Symptomatic observation most commonly occurs during feeding periods when specially trained personnel are able to identify animals exhibiting symptoms. The most common signs and symptoms that are associated with BRDC are:

1. Fever over 40˚C
2. Difficulty or labored breathing
3. Nasal and or ocular discharge
4. Depression
5. Diminished or no appetite
6. Shallow, rapid breathing
7. Coughing
8. Death
9. Chronic or Acute Bloat

When any combination of the above symptoms is present the animal should be considered to be infected with one of the pathogens associated with BRDC (Duff and Galyean, 2007). Once an animal is identified as symptomatic in many commercial feedlots they are removed from the pen and administered treatment immediately. A further complicating factor of this system is the animals with subclinical symptoms. These are animals that have an active infection with BRDC-

associated organisms however never express physical symptoms. It is impossible to accurately identify and diagnose this group of sub-acute animals given the typical observation method (Timsit et al., 2016).

There are currently many available tools for the identification of different BRDC pathogens. Table 2.1 presents a generalized list of various tools available for identification of different viral and bacterial pathogens associated with BRDC. A challenge that is often encountered with testing for different pathogens is the specificity and sensitivity of the test, as well as the cost to perform the test. Tests could provide inconclusive or different results based on the specific test performed or may show multiple disease-causing pathogens instead of identifying the main pathogenic causing the animal to be symptomatic.

Table 2.1. Summary of available diagnostic tools for identification of different pathogens in the Bovine Respiratory Disease Complex adapted from Pardon and Buczinski (2020).			
Test	Use	Positives	Negatives
Serology	Antibody Detection	Detect vaccine response and past infections.	Titers do not necessarily infer resistance and are not able to differentiate vaccine-induced anti-bodies from infection-acquired anti-bodies.
Culture-nasal, nasopharynx, trachea	Detect bacteria and viruses	Demonstrate the presence of colonization or active infection.	Positive culture does not necessarily mean lung infection or disease causation. Times for results to be obtained are days to weeks.
Culture- Lung Lesion	Detect bacteria and viruses	Require active replication of the agent in the tissue at time of death, so isolation usually indicates that high concentrations are in tissue. Antimicrobial resistance can be determined.	Sensitivity is not great and may miss true positives due to concurrent infections and antimicrobial therapy. Time for results to be obtained are days to weeks.
Immunohistochemistry- Lung Lesion	Detects antigen in lung lesions	One can localize the infections agent with the lesion. Strong evidence the infectious agent is related to disease.	Sensitivity and specificity depend on available monospecific immune serum or monoclonal antibodies to specific infectious agent.
In-situ hybridization - Lung Lesion	Detects region of genome of agent in lesion	One can localize the infections agent with the lesion. Strong evidence the infectious agent is related to disease. Monospecific antiserum or monoclonal antibodies not needed.	Depends on known, pathogen-specific genomic region for development of specific oligonucleotide primers.
Single PCR - nasal, nasopharynx, trachea, BAL swabs or collection	Detects genetic material of agent in sample	Provides specific evidence that infectious agent is in or recently has been in sample.	Cannot differentiate subclinical or incidental concurrent infection from natural exposure or vaccinations. Does not always detect infectious material. Cannot determine antimicrobial resistance.

Single PCR – Lung lesion from supernatant of tissue homogenate	Detects region of agent genome	Provides evidence of specific infectious agent is associated with disease.	May not represent causative infectious agent with diseased tissue or differentiate natural infection versus MLV vaccine.
Multiplex PCR- nasal, nasopharynx, trachea, BAL swabs or collection	Detects region of several agents' genome	With a single test, potential evidence of one or more infectious agent associated with disease can be determined. Test provides more information that single PCR.	May not represent causative agents with diseased tissue or differentiate natural infection versus MLV vaccine.
Multiplex PCR - Lung lesion from supernatant of tissue homogenate	Detects region of several agents' genome	With a single test, potential evidence of one or more infectious agent associated with disease can be determined. Test provides more information that single PCR.	May not represent causative agents with diseased tissue or differentiate natural infection versus MLV vaccine.

Prevention of BRDC has been focused on disease prevention through vaccination and metaphalaxis. Considerable effort has been applied to identify the ideal combination of these therapies, but the rate of the disease has not seen considerable change across the beef industry (Callan and Garry, 2002). Alternative methods of reducing the incidence of BRDC would be beneficial to the beef industry.

2.2 Genetic Selection Against BRDC

A practice that has commonly been theorized as a method to reduce BRDC is genetic selection, but one of the barriers that has routinely been encountered for the development of genetic selection tools is data collection and the specificity (or lack therefore) of the data required. These types of records are collected on the commercial level, but without direct information to relate these records back to nucleus parents (Bell et al., 2017). Traits that likely fall under this same scenario would be carcass merit, disease incidence, female fertility, and growth traits (Baller et al., 2020). Strategies that can bring information from the commercial side of the production system and relate it back to seed stock animals offers the opportunity to use the data to influence genetic selection decisions. One of the strategies that shows promise is the use of pooled DNA data. This has an advantage over traditional data collection methods because on the commercial level animals are traditionally managed as groups and not individually identified.

The heritability of a trait is an important factor to establish and gain an understanding on how much of the phenotypic performance is controlled by genetic influences and is an indication

of the observable differences due to phenotype (Bourdon, 1997). Heritability is expressed as a ratio of variances:

Equation 2.1 $$h^2 = \frac{\sigma_a^2}{\sigma_p^2}$$

where σ_a^2 was the additive genetic variance and σ_p^2 was the phenotypic variance. The importance of heritability for a trait is determining the pace that genetic improvement will be achieved for polygenetic traits. These traits are under the control of multiple genes across many loci in the genome that affect phenotypic performance. The heritability of a trait can be classified as either low, moderate, or high. Examples of lowly heritability traits include fertility and disease resistance, compared to traits in beef cattle such as weight and height which are moderate and highly heritable, respectively. The following section will summarize previous research findings for BRD resistance or susceptibility in beef cattle.

Table 2.2. Previous Literature Estimates of Heritability of Development of Symptoms Consistent with Bovine Respiratory Disease Complex.

Heritability	Breeds	Age	Animals, n	Reference
0.11 ± 0.06	Multi-breed	Pre-weaning	1,519	(Schneider et al. ,2010)
0.07 ± 0.04	Angus	Post-weaning	3,277	(Schneider et al., 2009)
0.13	Holstein	Pre-weaning	2,763	(Neibergs et al., 2014)
0.13-0.25	Multi-breed	Post-weaning	1,866	(Seabury et al., 2016)
0.22 ± 0.01	Multi-Breed	Pre-weaning	18,740	(Snowder et al., 2005)
0.05	Norwegian Red	All	250,212	(Heringstad et al., 2008)
0.10 ± 0.02	Multi-Breed	Pre-weaning	10,142	(Muggli-Cockett et al., 1992)
0.06 ± 0.07	Multi-Breed	Post-weaning	10,142	(Muggli-Cockett et al., 1992)
0.08 ± 0.01	Multi-Breed	Pre-weaning	18,112	(Snowder et al., 2006)
0.08 ± 0.01	Multi-breed	Post-weaning	18,112	(Snowder et al., 2007)
0.17 ± 0.08	Multi-breed	Post-weaning	2,869	(Cockrum et al., 2016)

Across previous literature estimates, the heritability of developing BRDC symptoms was low, ranging from 0.05 to 0.22 with many of the estimates for heritability are calculated using multi-breed datasets leading to higher heritability estimates than those estimated within a single

breed. In addition, heritability estimates from different ages and stages of production were similar. Heritability estimates for observations collected pre-weaning ranged from 0.08 to 0.22 and estimates from data collected during the post-weaning phase ranged from 0.06 to 0.13. Based on these results there does not appear to be an advantage to collecting phenotypic information at a certain stage of production.

Another important consideration when looking at making genetic improvement in a trait is the relationship the trait has with other traits in the breeding objective through genetic correlations. A genetic correlation is defined as the relationship between breeding values in one trait and the breeding values of another trait (Bourdon, 1997). Traits can be favorably correlated where both traits can be moved in a favorable direction simultaneously. Alternatively, traits can also be antagonistically correlated where improvement in one trait leads to unfavorable movement in the second. Traits can also show no relationship to one another and therefore can be selected for without any effect on one another.

Multiple studies have examined the genetic and environmental correlations between BRDC and other performance traits. Genetic and environmental correlations between BRDC and carcass traits were examined by Snowder et al. (2007). These studies found that genetic relationships were not different from zero for many of the carcass traits, including hot carcass weight, backfat, marbling score, rib eye area, retail product, fat trim, fat in ribs, and juiciness score. Other carcass traits had genetic correlation estimates that were different from zero to BRDC including average daily gain, Kidney Pelvic Heart Fat, total bone weight, and shear force. The strongest genetic relationship found was with percent of bone in the carcass. Although genetic relationships were generally small, environmental correlations between traits were strong. Environmental correlations were moderate to high and positive between BRDC and average daily gain, hot carcass

weight, KPH, retail product, fat trim, and bone weight. Conversely, Reinhardt et al. (2009) found that there was a negative relationship between average daily gain and carcass weight, and number of treatments for BRDC. However, results of this study do suggest that genetic selection for decreasing BRDC would not have adverse effects on other economically relevant traits such as growth and carcass quality and yield during the feedlot and harvest phase of production.

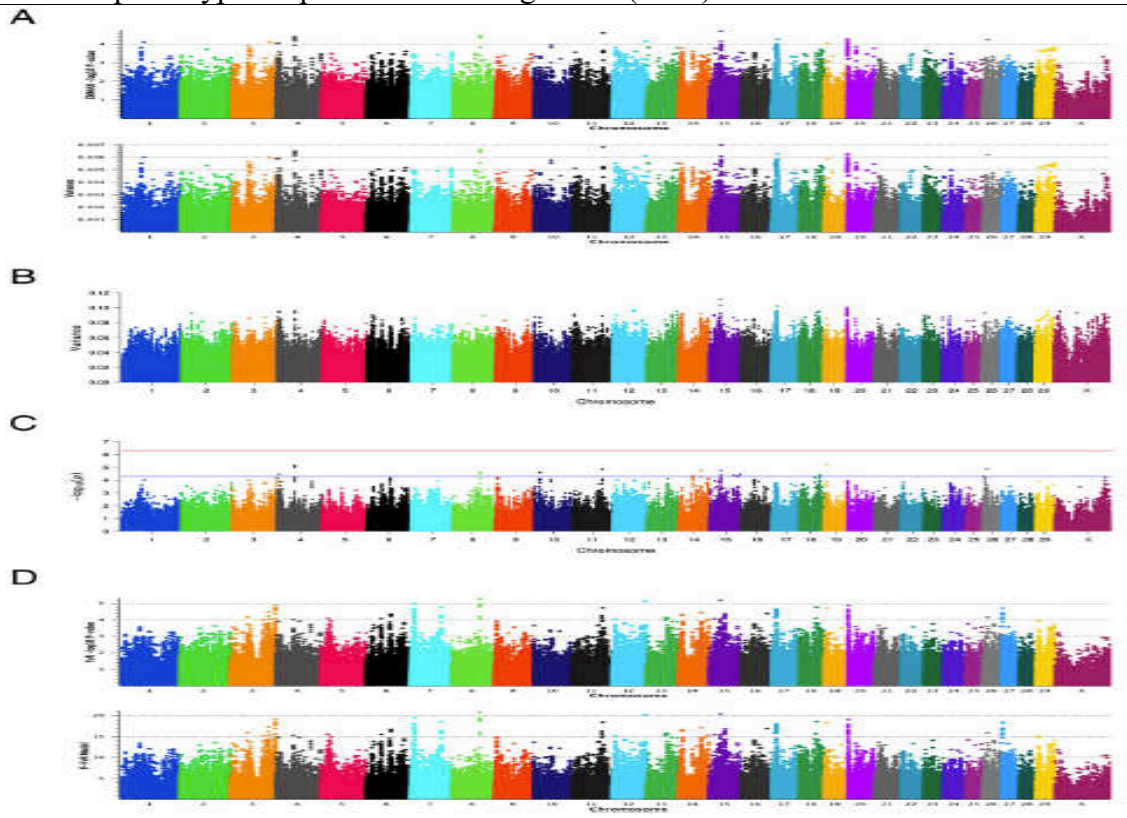
Genetic relationships between BRDC diagnosis and immune response traits at receiving were estimated by Cockrum et al. (2016). This study found that the immune response trait that had the strongest genetic relationships with BRDC was total IgG (0.42 ± 0.28) and IgG1 (0.36 ± 0.32). These relationships would suggest that animals that tended to have genetic propensity to have higher levels of total IgG and IgG1 at receiving also had a higher propensity to develop BRDC symptoms and require treatment. These results may also suggest that since BRDC symptoms are commonly caused by stressful events that may trigger immune response, that animals that are more tolerant of these events are less likely to develop the disease in the future.

Due to the challenges of large-scale data collection required for genetic evaluations, research has focused more recently on identifying genomic regions that appear to influence the susceptibility to respiratory disease with the goal of using these in selection programs. In a research setting, genomic regions can be identified and then applied more generally. This approach has been applied in the dairy industries of the USA, New Zealand, and the Netherlands (Hayes et al., 2009) by calculating prediction equations using a reference population then summing the equations across the genomes of subsequent generations to determine genomic estimated breeding values (GEBV) for other animals in the population.

Neibergs et al. (2014) used pre-weaned dairy calves from California and New Mexico to rank SNP loci using multiple statistical approaches. This experiment used a case/control approach

where animals were identified as cases using the McGuirk Scoring System (McGuirk, 2008), and SNP were identified in each independent location, as well as, across both locations. In Figure 2.2 the Manhattan plots from using data from both locations is presented. To identify the most significant regions across the different analysis techniques, regions were ranked based on significance, where those with the lowest rankings across the different analysis techniques were considered to be significant. The locations listed based on the criteria of an average ranking of less than 1,000, indicated three regions located on BTA3, two regions on BTA4, one region on BTA7, one region on BTA10, one region on BTA11, one region on BTA12, three regions on BTA14, three regions on BTA15, on region on BTA16, three regions on BTA17, one region on BTA18, one region on BTA19, three regions on BTA20, and one region on BTA29. Identification of these regions show evidence that genetic selection based on molecular information can be viewed as a viable option to making genetic progress in disease resistance although to this point nothing is currently implemented in the industry.

Figure 2.1. Manhattan plots representing four genome wide association analyses for a binary case-control phenotype adapted from Neibergs et al. (2014).



Panel A represents the results of EMMMAX-GRM model, panel B is the results of using a GBLUP model, panel C represents EIGENSTRAT case control principle component corrected model using first 100 principal comonents, panel D repernets an FvR analysis corrected for the forst 53 prionciple componets and covariates of sex and age.

2.2.1 Pooled DNA

To date, DNA pooling is a mechanism that has gained momentum as an opportunity to link different segments of the beef industry together. A major advantage to using this approach is that it dramatically reduces the cost of obtaining genomic information (Sham et al., 2002). The average cost of genotyping individual animals ranges from $30-$90 based on the cost of the genotype test. Individually genotyping commercial cattle quickly becomes cost prohibitive. The first application of using pooled DNA in literature was a study that used a case-control strategy to investigate type I diabetes mellitus (Arnheim et al., 1985). Since that time the strategy has been used via

microsatellites and SNP genotypes in humans (Pacek et al., 1993, Barcellos et al., 1997, Daniels et al., 1998, Shaw, Carrasquillo et al., 1998, Krumbiegel et al., 2011, Rivas et al., 2011), as well as in animals (Taylor and Phillips, 1996, Gonda et al., 2004, Huang et al., 2010, McDaneld, et al., 2014, Keele et al., 2015, Keele et al., 2016).

One of the major considerations for pooled DNA analyses is the pool construction. Pools can be constructed using a case-control strategy, similar background, or environments, or based on known genetic makeup (breeds/race) of individuals in the pool. DNA samples from each individual in the pool is then mixed so that each individual in the pool contributes equally. The main difference between a pooled DNA analysis and DNA on individuals are the results used from the array. In individual DNA analyses, each locus is called to a specific genotype of homozygous for allele 1, heterozygous, or homozygous for allele 2. Whereas in pooled DNA analyses the proportion of each allele at that locus are used in the analysis (Keele et al., 2015) because each SNP locus is biallelic so the peak intensity of one allele can be calculated by taking the peak intensity of that allele divided by the sum of peak intensities in the sample (Yang et al., 2006). These ratios then serve as an estimate for the allele frequency for the samples in the pool and is to referred to as pooling allele frequency. As a result, a major difference between using pooled DNA compared to individual DNA information is that the genotypes are expressed as a continuous variable instead of a categorical variable.

An important consideration when using pooled DNA data is the size of the pool. Several factors are reviewed and examined when considering pool size. Given equivalent total samples, the larger the pools the fewer arrays that would be used, which would overall reduce the cost. However, previous research has suggested that most of the error in DNA pooling comes from technical error in the array rather than in pool construction (Macgregor, 2007). Macgregor (2007)

recommended that a possible strategy to control for this is to run multiple arrays per pooled sample however, this has been shown to not be necessary previously (McDaneld et al., 2014). This relationship only held when pools were carefully constructed, and many individuals were included in each pool. The goal of using DNA pooling in research is to reduce overall cost so the optimal number of pools and samples per pool must be established in order to increase and enhance the utility of this approach.

The main function of the identification of the proper construction of pooled DNA analyses is determining statistical power of the experiment. When comparing statistical power of pooled DNA versus individual genotyping, the statistical power of pooled DNA is lower than that of individual genotyping (Risch and Teng, 1998). Whereas in pooled DNA analyses, power of the experiment is determined as a function of the percentage of variance explained, allelic frequency, and disease prevalence in the population, where disease is the primary trait of interest. Prior to the use of pooled DNA analyses in human genetics, the common method to determine underlying genetic risk was through linkage analyses. In these analyses, affected siblings were compared to non-affected parents. The more often that the affected siblings shared the same allele at a particular site, the higher the likelihood that site was linked to a gene that causes the disorder or predisposition to a disease (Risch and Merikangas, 1996). One of the challenges faced with this type of analyses was that only very large SNP effects were identified. A transition over to linkage disequilibrium analyses was presented as an alternative but these types of analyses could also use pooled DNA information. Table 2.2 is adapted from Risch and Merikangas (1996) which shows the estimated numbered of samples required to detect differences at different risk ratios and allele frequencies.

Table 2.3 Comparison of Linkage and association studies. Number of families needed for identification of a disease gene adapted from Risch and Merikangas (1996).

		Linkage			Association			
					Singletons		Sub pairs	
Genotypic Risk Ratio (ϒ)	Frequency of disease allele A (p)	Probability of allele sharing (Y)	No. of families required (N)	Probability of transmitting disease allele A P(tr-A)	Proportion of heterozygous parents (Het)	(N)	(Het)	(N)
4.0	0.01	0.520	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.576	297	0.800	0.500	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296,710	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	695	0.323	264
	0.50	0.526	2498	0.667	0.500	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620,807	0.600	0.025	19,320	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2218	0.253	941
	0.50	0.510	17,997	0.600	0.500	949	0.490	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

In the above table, the power calculations were formed based on individual genotyping of cases and controls. It also displays the power of the linkage disequilibrium or association analyses over the previously used linkage approach. To generalize the table to pooled DNA approach data, the difference would compare allele frequencies between affected and unaffected pooled individuals (Risch and Teng, 1998). The authors recommended that pool construction should consider confounding if using pooled data. For example, in a case control experiment of unrelated individuals, if one ethnicity (humans) or breed (animals) are overrepresented in one of the pools then an allele frequency may emerge that was an artifact of the confounding of these effects. One of the ways to circumvent this confounding was to sample cases and controls using related individuals. This approach also helps to make sure that families/breeds are equally represented in

matched pools. The number of samples needed to be collected when using affected individuals versus unaffected siblings are presented in Table 2.3.

Table 2.4. Number of families required to detect linkage disequilibrium for siblings with r affected and s unaffected siblings, without parents, using DNA pooling at different levels of gene frequency(p), Adapted from Risch and Teng (1998).

	r =1		r =2		r = 3	r = 4
	s = 1	s = 2	s = 1	s = 2	s = 2	s = 2
Dominant						
p = 0.05	753	534	355	227	147	126
p = 0.20	489	357	376	247	248	296
p = 0.70	5,719	4,317	6,490	4,357	5,638	7,860
Recessive						
p = 0.05	79,556	59,234	22,031	14,555	4,810	1,883
p = 0.20	2,022	1,498	745	494	237	145
p = 0.70	341	271	269	196	206	255
Multiplicative						
p = 0.05	2,811	2,032	1,535	992	605	405
p = 0.20	891	655	547	361	258	209
p = 0.70	831	642	689	478	471	515
Additive						
p = 0.05	1,690	1,213	885	569	351	253
p = 0.20	717	526	483	318	258	239
p = 0.70	1,292	990	1,117	765	760	818

Table 2.3 presents four different models of inheritance for the alleles that are disease causing. For dominant models, as the gene frequency (pp) increases so does the required number of families. This was in contrast to recessive and multiplicative models as the gene frequency increases the number of families required decreases. For an additive model and intermediate gene frequency results in the lowest number of samples needed.

2.3 Analysis of Pooled DNA data

The goal of a pooled DNA analysis is to identify SNP that are linked to causal variants that cause differences in phenotype. Pooled DNA can be applied to any type of analyses to identify significant genomic markers but is commonly used for disease studies. For the study of disease, one of the most common stratification approaches is through a case control strategy.

In genome wide association analyses (GWAA) the primary statistic of interest is to estimate the proportion of alleles between pools (Macgregor et al., 2006). Traditionally these analyses are conducted using the pooling allele frequency (PAF) as the response variable. The PAF is the proportion of the A Allele for each pool, which is calculated using the following formula:

Equation 2.2
$$PAF = \frac{Intensisty\ of\ A\ Allele}{(Intensity\ of\ A\ Allele + Intensity\ of\ B\ Allele)}$$

When using genotype data, the intensity of each allele is identified using the intensity of each allele from the chip. In most cases there are two colors: red and green. Each color would correspond to one of the alleles. Calculating this proportion then acts as a proxy for each allele frequency in the pool. Once this response variable is established then a linear model can be fitted. Generally, the test statistic that is of interest is a t-test. In pooled DNA analysis, (Macgregor et al., 2006) presented the test statistic as follows:

Equation 2.3
$$T_{Simple} = \frac{(\tilde{p}_a - \tilde{p}_u)^2}{Var(\hat{p}_a - \hat{p}_u)} \simeq \frac{(\tilde{p}_a - \tilde{p}_u)^2}{Var(\tilde{p}_a - \tilde{p}_u)}$$

In equation 2.3 equation, p_a represents the population frequency in cases, and p_u is the frequency in controls. The pooled sample estimate of allele frequency is denoted $\tilde{p}_a$ and $\tilde{p}_u$ and the allele

frequency if the samples were individually genotyped without error are $\hat{p}_a$ and $\hat{p}_u$. The above test statistic can be used to calculate a p value to test for significance. One challenge to this approach is that each loci is tested resulting in a multiple testing penalty that must be applied to p values to limit false discovery rate. The common corrections used for large number of tests include Bonferroni and Benjamini and Hochberg corrections (Jafari and Ansari-Pour, 2019).

Within linear regression, one important factor is determining what effects will be included in the model. There are two different types of effects that can be fitted within these models. The first type of effect is commonly referred to as fixed effects because the classification of these effects are assumed to be known prior to the data analysis. Examples of these types of effects could include disease status of the pool (example: Sick vs. Healthy), environmental effects, or the potential information of an individual's ancestry in the pool (breed, breed composition, principal component, or matched case/control pair) The second type of effect is a random effect. Random effects are assumed to be drawn from different distributions within a larger population and inferences about random effects relate to the population. Since GWAA are linear models, both fixed and random effects can be included. When using pooled DNA, the different effects that would traditionally be included in the model are accounted for in the pool construction. One effect that is of particular interest to include in the model would be to account for the genetic relationships among the different pools. It is important to consider controlling for genetic architecture in the evaluation in the context of pooled DNA because it can reduce confounding. Traditionally, the way that these genetic relationships can be calculated is by estimating the genetic variance and covariance matrix among the pools (McDaneld et al., 2014, Keele et al., 2015, Keele et al., 2016).

An alternative to linear regression for analysis of genomic data are machine learning algorithms. These types of models are well suited for problems encountered with genotype

analyses, namely where predictors are greater than the number of samples. One of the methodologies that has previously been applied to genomic data is the Random Forrest (RF) algorithm. Chen and Ishwaran (2012) described the RF algorithm as a tree-based ensemble machine learning tool that is highly data adaptive, applies to "large p, small n", and can accommodate interaction and correlations among features. These models have several useful applications when applied to genomic data. They can be used for prediction by using genotype information to predict phenotype or susceptibility to a disease. Variable importance can also be accommodated, and when applied to genomic data can help to identify important SNP. This type of model can also be applied to GWAA. The basic unit of RF is the known as a tree, that split the data into binary groups. The basic steps of the model can be described as:

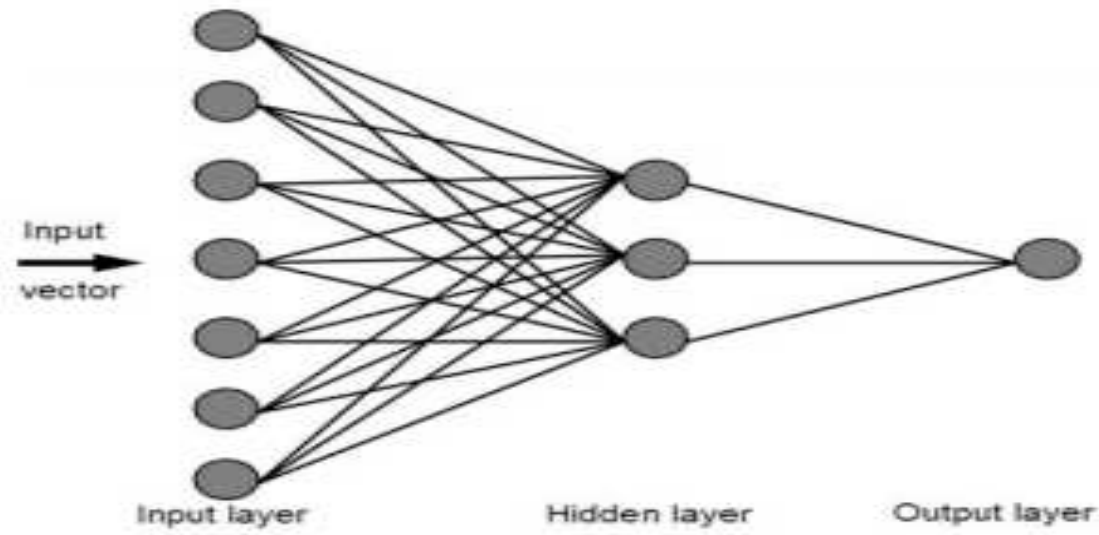
1. Draw n tree bootstrap samples
2. Grow a tree from each bootstrap data set. At each node of the tree randomly select a subset of variables for splitting.
3. Aggerate information from the number of trees for new data
4. Compute out of the bag error rate based on classification of data not in the bootstrap sample (Chen and Ishwaran, 2012).

An advantage of the RF model is that it could be applied with minimal effort. A significant contributory step would be running a tuning algorithm that identified the ideal parametrizations for different facets of the model. Some of these tuning parameters could include the number of trees or the number of input SNP to test at each node. Tuning is achieved by running what is known as a "grid search." This test examines different combinations of these variables that lead to the lowest error rate. For each model there is an ideal combination of these parameters that will lead to the most accurate model. A drawback to this approach is that there is not a test statistic of

significance for these models, resulting in situations where a set of variables with moderate predictive power individually have significance when used in combination with the other SNP. A proposed suggestion for these situations is to impose restrictions on the SNP used within each forest to minimize the LD between SNP used for testing.

Neural networks are another method that has also been investigated for GWAA analyses. The framework for neural networks is to mimic the human brain and the way it processes information. This framework occurs through the minor adjustments of "neurons" or interconnected processing elements in the model. A visual representation of this process is presented in Figure 2.1 from (Lancashire et al., 2009).

Figure 2.2. Architecture of a multi-layered artificial neural network Lancashire et al. (2009).



Within the neural network, the number of hidden layers and the number of neurons is determined based on the complexity of the problem and the interactions among prediction variables. Much like the RF, the parameters that determine the model need to be tuned or identified to find the best

model for the data. Inside the neural network some of the most common tunable parameters include the learning rate, the batch size, the momentum, and weight decay (Koumakis, 2020). The learning rate of the model is used to update the weights during training, optimization of this parameter can help with the efficiency of the training and prevents overfitting of the model. The batch size of the model represents the number of samples that pass through the network before the weights in the nodes are updated. The momentum of the model is a parameter that iteratively determines how to update model parameters that minimize the loss function of the model. Weight decay of the model is a technique that regularized the weights in a neural network.

The most common architecture used for genomic data is known as the convolutional neural network. This architecture was first introduced for classification of handwritten characters. It has layers of convolution that receive units from the previous layer and produce a proximity (Koumakis, 2020). In a convolutional neural network, there are multiple layers included an input layer, hidden layers, and an output layer. Generally, in multi-layer models the intermediate layers are known as the hidden layers. The hidden layers are termed this because the inputs and outputs of these layers are hidden by the activation function that is used. A common activation function used in these types of networks are a rectifier (ReLU). This is the main function of the neuron which takes the input and then defines the output of the neuron. The goal of this approach is to combine feature maps from one layer to the next until the desired output is achieved. Within convolutional neural networks there are two main methods that can be used to train the model. The most straightforward method is known as a feed forward architecture. This method functions where each neuron layer is connected only to neurons in layer i+1 and all of the edges can have different weights (Zou et al., 2019). This model is effective for prediction when no special relationships are present among the input data. A different architecture that can be used in neural

networks is a recurrent neural network. The functionality of this architecture is similar to that of the feed forward architecture (Koumakis, 2020). However, the main difference between the two is that when recurrent architecture is used the information can travel between the hidden layers in both a forward and backward direction. The challenge is that the recurrent approach causes additional complexity to the neural network compared the feed forward architecture.

2.4 Genetic Selection for Improved Health Traits in Animals

Traditional methods for genetic improvement have focused on data collection and submission of that information to breed associations. The associations then deliver genetic predictions for these traits to their members and customers. These tools are utilized for genetic selection predominately in the seedstock sector to provide breeding stock for the commercial industries. However, one challenge with selection for disease resistance is that these traits are more commonly recorded on the commercial level. Traditional methodology for genetic improvement of traits associated with disease resistance must rely on a different approach for genetic improvement. Despite this limitation there are still many examples across multiple species where genetic improvement has been able to be achieved.

One of the industries that has had the most success in development and implementation of genetic predictions for health traits is the dairy industry. Currently the Holstein Association USA publishes a health trait index that looks to improve selection for milk fever, displaced abomasum, ketosis, mastitis, metritis, and retained placenta (https://www.holsteinusa.com/genetic_evaluations/ss_tpi_formula.html). While this information is available independently it is also combined into the Holstein Association's TPI index which includes all economically relevant traits. More recently, the American Jersey Cattle Association

have also implemented selection for health traits in their population using the same set of traits as the Holstein Association (https://www.uscdcb.com/wp-content/uploads/2020/01/CDCB-Jersey-Health-Traits.pdf).

Of the traits that are included in both indexes, the one that has been the most studied is mastitis. One of the most common approaches currently applied is marker assisted selection. Under this approach, samples can be collected and tested for previously identified QTL for the traits of interest (Boichard et al., 2006) this approach allows for selection of animals prior to data collection or observing their susceptibility to the disease. Genetic information is also being implemented for the treatment of this disease and is generally referred to as Recombinant Protein and Somatic Gene Therapy (Kerr and Wellnitz, 2003, Zhang et al., 2007). This type of therapy is looking to harness the host's immune system response to clear infection and prevent antimicrobial resistance (Saleem et al., 2024).

Small ruminants, such as sheep and goats, also use genetic prediction to improve a health-related traits. One of the first genetic selection tools that was implemented in the sheep and goat industry was the identification of the genomic region responsible for scrapie susceptibility. Scrapie is a neurogenerative disease that affects the central nervous system of sheep and goats. In sheep, several genetic variants in the PrP gene have been identified that have been shown to be significantly associated with susceptibility to the disease (Bossers et al., 1996). In addition, genetic predictions for increased parasite resistance are available through genetic evaluations conducted as part of the National Sheep Improvement Program (NSIP) in the United States (Burke and Miller, 2020). The genetic predictions for this trait are based on data obtained from collection of fecal egg counts collected on animals at weaning (60-90 days of age) and during the post-weaning period

up to a year of age with the goal to reduce fecal egg counts for the animals and in turn reduce the reliance on de-worming medications.

Another industry placing emphasis on selection for improved health traits is the swine industry. While this industry has traditionally put a large emphasis on biosecurity, there have still been several pandemics that have threatened commercial swine operations. Much of the work in swine has focused on identification of genetic variants that influence resistance to a specific pathogen. One successful application of this approach was the identification of a variant in the SSC4 gene located on Chr4 that had a significant impact on viral load against Porcine Reproductive and Respiratory Syndrome (PRRS) (Rowland et al., 2012). The discovery of this variant did not make the animals with the favorable genotype more resistant to the disease, but the genotype had more favorable growth rates and were less influenced by the PRRS infection. While this breakthrough did allow for a reduction in the effect of this virus on pig production, it did not eliminate susceptibility to the disease. A further advancement came when it was discovered that the protein called CD163 was responsible for modifying the virus to allow for infection (Whitworth et al., 2016). This discovery led to the production of pigs that were completely resistant to the PRRS virus by knocking out the DNA that produced the CD163 protein. Currently this technology has only been used on research animals but is planned to be applied to the commercial industry through introduction into sow lines and could eliminate the impact of this disease on the swine industry (Prather et al., 2017). The above application of genome editing in animals to address disease susceptibility offers an exciting opportunity for all livestock species to combat disease.

2.5 Conclusions

Given the large economic impactions that BRDC has on the beef industry, research into methods that can reduce the incidence of the disease are warranted. One of the best methods to create generational improvement and reduce treatment rates is through the use of genetic selection. Previous studies in beef cattle have looked to identify genomic regions. However, this has not been done on many beef breed type animals using pooled DNA analyses. If genomic regions are identified these regions could be used to select for reduced incidence in beef cattle populations. Due to the unique nature of disease traits new and different analyses strategies for this type of data may prove to be more effective. Similar to other species, once identification and selection against regions that lead to increased susceptibility to the disease is undertaken then the overall economic cost to the industry can also be reduced.

LITERATURE CITED

Allen, J. W., L. Viel, K. G. Bateman and S. Rosendal (1992). Changes in the bacterial flora of the upper and lower respiratory tracts and bronchoalveolar lavage differential cell counts in feedlot calves treated for respiratory diseases. Can J Vet Res **56**(3): 177-183.

APHIS, U. (2013). National Animal Health Monitoring System Beef Feedlot Study 2011. T*ypes and Costs of Respiratory Disease Treatments in U.S. Feedlots* . Info Sheet (2013)

Arnheim, N., C. Strange and H. Erlich (1985). Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. Proc Natl Acad Sci U S A **82**(20): 6970-6974.

Baller, J. L., S. D. Kachman, L. A. Kuehn and M. L. Spangler (2020). Genomic prediction using pooled data in a single-step genomic best linear unbiased prediction framework. J Anim Sci **98**(6).

Barcellos, L. F., W. Klitz, L. L. Field, R. Tobias, A. M. Bowcock, R. Wilson, M. P. Nelson, J. Nagatomi and G. Thomson (1997). Association mapping of disease loci, by use of a pooled DNA genomic screen. Am J Hum Genet **61**(3): 734-747.

Bell, A. M., J. M. Henshall, L. R. Porto-Neto, S. Dominik, R. McCulloch, J. Kijas and S. A. Lehnert (2017). Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. Genet Sel Evol **49**(1): 28.

Bishop, S. C. and C. A. Morris (2007). Genetics of disease resistance in sheep and goats. Small ruminant research **70**(1): 48-59.

Boichard, D., S. Fritz, M. N. Rossignol, F. Guillaume, J. J. Colleau and T. Druet (2006).

IMPLEMENTATION OF MARKER-ASSISTED SELECTION: PRACTICAL LESSONS FROM DAIRY CATTLE. 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, MG, Brasil.

Bossers, A., B. E. Schreuder, I. H. Muileman, P. B. Belt and M. A. Smits (1996). PrP genotype contributes to determining survival times of sheep with natural scrapie. J Gen Virol **77 (Pt 10)**: 2669-2673.

Bourdon, R. M. (1997). Understanding Animal Breeding, Prentice-Hall.

Burke, J. M. and J. E. Miller (2020). Sustainable Approaches to Parasite Control in Ruminant Livestock. Vet Clin North Am Food Anim Pract **36**(1): 89-107.

Callan, R. J. and F. B. Garry (2002). Biosecurity and bovine respiratory disease. Vet Clin North Am Food Anim Pract **18**(1): 57-77.

Cardoso, F. F., C. C. G. Gomes, B. P. Sollero, M. M. Oliveira, V. M. Roso, M. L. Piccoli, R. H. Higa, M. J. Yokoo, A. R. Caetano and I. Aguilar (2015). Genomic prediction for tick resistance in Braford and Hereford cattle. Journal of animal science **93**(6): 2693-2705.

Chen, X. and H. Ishwaran (2012). Random forests for genomic data analysis. Genomics **99**(6): 323-329.

Cockrum, R. R., S. E. Speidel, J. L. Salak-Johnson, C. C. Chase, R. K. Peel, R. L. Weaber, G. H. Loneagan, J. J. Wagner, P. Boddhireddy, M. G. Thomas, K. Prayaga, S. DeNise and R. M. Enns (2016). Genetic parameters estimated at receiving for circulating cortisol, immunoglobulin G, interleukin 8, and incidence of bovine respiratory disease in feedlot beef steers. J Anim Sci **94**(7): 2770-2778.

Czuprynski, C. J., F. Leite, M. Sylte, C. Kuckleburg, R. Schultz, T. Inzana, E. Behling-Kelly and L. Corbeil (2004). Complexities of the pathogenesis of Mannheimia haemolytica and Haemophilus somnus infections: challenges and potential opportunities for prevention? Anim Health Res Rev **5**(2): 277-282.

Daniels, J., P. Holmans, N. Williams, D. Turic, P. McGuffin, R. Plomin and M. J. Owen (1998). A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. Am J Hum Genet **62**(5): 1189-1197.

Duff, G. C. and M. L. Galyean (2007). Board-invited review: recent advances in management of highly stressed, newly received feedlot cattle. J Anim Sci **85**(3): 823-840.

Fulton, R. W., B. J. Cook, D. L. Step, A. W. Confer, J. T. Saliki, M. E. Payton, L. J. Burge, R. D. Welsh and K. S. Blood (2002). Evaluation of health status of calves and the impact on feedlot performance: assessment of a retained ownership program for postweaning calves. Can J Vet Res **66**(3): 173-180.

Fulton, R. W., J. M. d'Offay, C. Landis, D. G. Miles, R. A. Smith, J. T. Saliki, J. F. Ridpath, A. W. Confer, J. D. Neill, R. Eberle, T. J. Clement, C. C. Chase, L. J. Burge and M. E. Payton (2016). Detection and characterization of viruses as field and vaccine strains in feedlot cattle with bovine respiratory disease. Vaccine **34**(30): 3478-3492.

Fulton RW, Step DL, Wahrmund J, Burge LJ, Payton ME, Cook BJ, Burken D, Richards CJ, Confer AW. Bovine coronavirus (BCV) infections in transported commingled beef cattle and sole-source ranch calves. Can J Vet Res. 2011 Jul;75(3):191-9. PMID: 22210995; PMCID: PMC3122965.

Gonda, M. G., J. A. Arias, G. E. Shook and B. W. Kirkpatrick (2004). Identification of an

ovulation rate QTL in cattle on BTA14 using selective DNA pooling and interval mapping. Anim Genet **35**(4): 298-304.

Griffin, D. (2014). The monster we don't see: subclinical BRD in beef cattle. Anim Health Res Rev **15**(2): 138-141.

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

Heringstad, B., Y. M. Chang, D. Gianola and O. Østerås (2008). Short Communication: Genetic Analysis of Respiratory Disease in Norwegian Red Calves. Journal of dairy science **91**(1): 367-370.

Huang, W., B. Kirkpatrick, G. Rosa and H. Khatib (2010). A genome-wide association study using selective DNA pooling identifies candidate markers for fertility in Holstein cattle. Animal genetics **41**(6): 570-578.

Irwin, M. R., S. McConnell, J. D. Coleman and G. E. Wilcox (1979). Bovine respiratory disease complex: a comparison of potential predisposing and etiologic factors in Australia and the United States. J Am Vet Med Assoc **175**(10): 1095-1099.

Jafari, M. and N. Ansari-Pour (2019). Why, When and How to Adjust Your P Values? Cell J **20**(4): 604-607.

Keele, J. W., L. A. Kuehn, T. G. McDaneld, R. G. Tait, S. A. Jones, B. N. Keel and W. M. Snelling (2016). Genomewide association study of liver abscess in beef cattle. J Anim Sci **94**(2): 490-499.

Keele, J. W., L. A. Kuehn, T. G. McDaneld, R. G. Tait, S. A. Jones, T. P. Smith, S. D.

Shackelford, D. A. King, T. L. Wheeler, A. K. Lindholm-Perry and A. K. McNeel (2015). Genomewide association study of lung lesions in cattle using sample pooling. J Anim Sci **93**(3): 956-964.

Kerr, D. E. and O. Wellnitz (2003). Mammary expression of new genes to combat mastitis. J Anim Sci **81 Suppl 3**: 38-47.

Koumakis, L. (2020). Deep learning models in genomics; are we there yet? Comput Struct Biotechnol J **18**: 1466-1473.

Krumbiegel, M., F. Pasutto, U. Schlötzer-Schrehardt, S. Uebe, M. Zenkel, C. Y. Mardin, N. Weisschuh, D. Paoli, E. Gramer, C. Becker, A. B. Ekici, B. H. Weber, P. Nürnberg, F. E. Kruse and A. Reis (2011). Genome-wide association study with DNA pooling identifies variants at CNTNAP2 associated with pseudoexfoliation syndrome. Eur J Hum Genet **19**(2): 186-193.

Kuhnlein, U., L. Ni, D. Zadworny and W. Fairfull (1997). DNA polymorphisms in the chicken growth hormone gene: response to selection for disease resistance and association with egg production. Animal genetics **28**(2): 116-123.

Lancashire, L. J., C. Lemetre and G. R. Ball (2009). An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies. Brief Bioinform **10**(3): 315-329.

Macgregor, S. (2007). Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. Eur J Hum Genet **15**(4): 501-504.

Macgregor, S., P. M. Visscher and G. Montgomery (2006). Analysis of pooled DNA samples on

high density arrays without prior knowledge of differential hybridization rates. Nucleic Acids Res **34**(7): e55.

Martin, P., H. W. Barkema, L. F. Brito, S. G. Narayana and F. Miglior (2018). Symposium review: Novel strategies to genetically improve mastitis resistance in dairy cattle. Journal of dairy science **101**(3): 2724-2736.

Martin, S. W. and J. G. Bohac (1986). The association between serological titers in infectious bovine rhinotracheitis virus, bovine virus diarrhea virus, parainfluenza-3 virus, respiratory syncytial virus and treatment for respiratory disease in Ontario feedlot calves. Can J Vet Res **50**(3): 351-358.

McDaneld, T. G., L. A. Kuehn, M. G. Thomas, W. M. Snelling, T. P. Smith, E. J. Pollak, J. B. Cole and J. W. Keele (2014). Genomewide association study of reproductive efficiency in female cattle. J Anim Sci **92**(5): 1945-1957.

McGuirk, S. M. (2008). Disease management of dairy calves and heifers. Vet Clin North Am Food Anim Pract **24**(1): 139-153.

Muggli-Cockett, N. E., L. V. Cundiff and K. E. Gregory (1992). Genetic analysis of bovine respiratory disease in beef calves during the first year of life. J Anim Sci **70**(7): 2013-2019.

NAHMS (2011). Feedlot 2011 Part IV: Health and health management on U.S. feedlots with a capacity of 1,000 or more head. https://www.aphis.usda.gov/animal_health/nahms/feedlot/downloads/feedlot2011/Feed11_dr_PartI.pdf.

Neibergs, H. L., C. M. Seabury, A. J. Wojtowicz, Z. Wang, E. Scraggs, J. N. Kiser, M. Neupane,

J. E. Womack, A. Van Eenennaam, G. R. Hagevoort, T. W. Lehenbauer, S. Aly, J. Davis, J. F. Taylor and B. R. D. C. C. A. P. R. Team (2014). Susceptibility loci revealed for bovine respiratory disease complex in pre-weaned holstein calves. BMC Genomics **15**: 1164.

Pacek, P., A. Sajantila and A. C. Syvänen (1993). Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. PCR Methods Appl **2**(4): 313-317.

Panciera, R. J. and A. W. Confer (2010). Pathogenesis and pathology of bovine pneumonia. Veterinary Clinics: Food Animal Practice **26**(2): 191-214.

Popescu, L., N. N. Gaudreault, K. M. Whitworth, M. V. Murgia, J. C. Nietfeld, A. Mileham, M. Samuel, K. D. Wells, R. S. Prather and R. R. R. Rowland (2016). Genetically edited pigs lacking CD163 show no resistance following infection with the African swine fever virus isolate, Georgia 2007/1. Virology (New York, N.Y.) **501**: 102-106.

Prather, R. S., K. D. Wells, K. M. Whitworth, M. A. Kerrigan, M. S. Samuel, A. Mileham, L. N. Popescu and R. R. R. Rowland (2017). Knockout of maternal CD163 protects fetuses from infection with porcine reproductive and respiratory syndrome virus (PRRSV). Scientific reports **7**(1): 13371-13375.

Reinhardt, C. D., W. D. Busby and L. R. Corah (2009). Relationship of various incoming cattle traits with feedlot performance and carcass traits. J Anim Sci **87**(9): 3030-3042.

Risch, N. and K. Merikangas (1996). The future of genetic studies of complex human diseases. Science **273**(5281): 1516-1517.

Risch, N. and J. Teng (1998). The relative power of family-based and case-control designs for

linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Res **8**(12): 1273-1288.

Rivas, M. A., M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, T. Fennell, A. Kirby, A. Latiano, P. Goyette, T. Green, J. Halfvarson, T. Haritunians, J. M. Korn, F. Kuruvilla, C. Lagacé, B. Neale, K. S. Lo, P. Schumm, L. Törkvist, M. C. Dubinsky, S. R. Brant, M. S. Silverberg, R. H. Duerr, D. Altshuler, S. Gabriel, G. Lettre, A. Franke, M. D'Amato, D. P. McGovern, J. H. Cho, J. D. Rioux, R. J. Xavier, M. J. Daly, N. I. o. D. a. D. K. D. I. B. D. G. C. N. IBDGC), U. K. I. B. D. G. Consortium and I. I. B. D. G. Consortium (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet **43**(11): 1066-1073.

Rowland, R. R., J. Lunney and J. Dekkers (2012). Control of porcine reproductive and respiratory syndrome (PRRS) through genetic improvements in disease resistance and tolerance. Front Genet **3**: 260.

Schneider, M. J., R. G. Tait, W. D. Busby and J. M. Reecy (2009). An evaluation of bovine respiratory disease complex in feedlot cattle: Impact on performance and carcass traits using treatment records and lung lesion scores. J Anim Sci **87**(5): 1821-1827.

Schneider, M. J., R. G. Tait, M. V. Ruble, W. D. Busby and J. M. Reecy (2010). Evaluation of fixed sources of variation and estimation of genetic parameters for incidence of bovine respiratory disease in preweaned calves and feedlot cattle. J Anim Sci **88**(4): 1220-1228.

Seabury, C. M., H. L. Neibergs, J. F. Taylor and J. E. Womack (2016). 0285 Genome-wide association study of bovine respiratory disease complex in U.S. feedlot cattle. Journal of Animal Science **94**(suppl_5): 135-135.

Sham, P., J. S. Bader, I. Craig, M. O'Donovan and M. Owen (2002). DNA Pooling: a tool for large-scale association studies. Nat Rev Genet **3**(11): 862-871.

Shaw, S. H., M. M. Carrasquillo, C. Kashuk, E. G. Puffenberger and A. Chakravarti (1998). Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome Res **8**(2): 111-123.

Snowder, G. D., L. D. Van Vleck, L. V. Cundiff and G. L. Bennett (2005). Influence of breed, heterozygosity, and disease incidence on estimates of variance components of respiratory disease in preweaned beef calves. Journal of animal science **83**(6): 1247-1261.

Snowder, G. D., L. D. Van Vleck, L. V. Cundiff and G. L. Bennett (2006). Bovine respiratory disease in feedlot cattle: environmental, genetic, and economic factors. J Anim Sci **84**(8): 1999-2008.

Snowder, G. D., L. D. Van Vleck, L. V. Cundiff, G. L. Bennett, M. Koohmaraie and M. E. Dikeman (2007). Bovine respiratory disease in feedlot cattle: phenotypic, environmental, and genetic correlations with growth, carcass, and longissimus muscle palatability traits. J Anim Sci **85**(8): 1885-1892.

Taylor, B. A. and S. J. Phillips (1996). Detection of obesity QTLs on mouse chromosomes 1 and 7 by selective DNA pooling. Genomics **34**(3): 389-398.

Taylor, J. D., R. W. Fulton, T. W. Lehenbauer, D. L. Step and A. W. Confer (2010). The epidemiology of bovine respiratory disease: What is the evidence for predisposing factors? Can Vet J **51**(10): 1095-1102.

Whitworth, K. M., R. R. Rowland, C. L. Ewen, B. R. Trible, M. A. Kerrigan, A. G. Cino-Ozuna, M. S. Samuel, J. E. Lightner, D. G. McLaren, A. J. Mileham, K. D. Wells and R. S. Prather (2016). Gene-edited pigs are protected from porcine reproductive and respiratory syndrome virus. Nat Biotechnol **34**(1): 20-22.

Yang, H. C., C. H. Lin, S. I. Hung and C. S. J. Fann (2006). A Comparison of Individual Genotyping and Pooled DNA Analysis for Polymorphism Validation Prior to Large-Scale Genetic Studies. Annals of human genetics **70**(3): 350-359.

Zhang, J. X., S. F. Zhang, T. D. Wang, X. J. Guo and R. L. Hu (2007). Mammary gland expression of antibacterial peptide genes to inhibit bacterial pathogens causing mastitis. J Dairy Sci **90**(11): 5218-5225.

Zhu, Dong, Madeline Giles, Tim Daniell, Roy Neilson, and Xiao-ru Yang. Does reduced usage of antibiotics in livestock production mitigate the spread of antibiotic resistance in soil, earthworm guts, and the phyllosphere?. Environment international 136 (2020): 105359.

Zou, J., M. Huss, A. Abid, P. Mohammadi, A. Torkamani, A. Telenti (2019). A primer on deep learning in genomics. Nat Genet **51**(1): 12-18.

CHAPTER III

Comparison of Genomic Relationship Matrices Using Differing Number of SNP in Pooled DNA Analyses

3.1 Introduction

In the beef cattle industry, genetic evaluations have traditionally used several important pieces of information for estimation of breeding values or expected progeny differences including pedigree, individual performance, contemporary group, and most recently, genomic information (Guidelines for Uniform Beef Improvement Programs. BIF Guidelines Wiki, 2023). However, this data requirements are not the most accommodating to alternative/nontraditional data sources like that from commercial beef industry production systems where data may be less precise and/or complete. For example, within this sector of the industry, individual pedigree or performance information is often unavailable. These factors limit the use of the massive amounts of commercial data that is collected but not currently incorporated into a traditional genetic evaluation. There are phenotypes such as carcass measurements, feed intake, and health that are commonly recorded in this sector, but the data is not able to be utilized for genetic evaluation because of missing data and data connections (pedigree, etc.).

One possible strategy to overcome this limitation is through the use of DNA pooling before genotyping as a means to link phenotypic information to genetic information. The premise of DNA pooling is to combine multiple DNA samples and perform a single genotype chip on the pool (Sham et al., 2002). Previous research has investigated this in both a simulation context (Alexandre et al., 2019; Baller et al., 2020; Vargas Jurado et al., 2021), as well as, with real-world actual commercial beef cattle or sheep populations (Reverter et al., 2016; Bell et al., 2017). Most of the

previous reports have evaluated the optimization of pool construction in terms of individuals in the pool and pooling strategies for individuals. These studies have focused on higher density genotypes. However, previous reports have not explored if costs could be reduced by using lower density panels on the pools. Our goal was to evaluate the influence of number of SNP sampled genome wide on the genomic relationship matrix among pools of commercial cattle. Our hypothesis is that a small proportion of the high-density SNP will capture most of the information needed.

3.2 Materials and Methods

3.2.1 Sample Preparation and Pooling

Samples were collected from three separate beef processing plants in Nebraska and Colorado. Initial sample collection was done by collecting ears that were removed as part of the harvesting procedure. Animal samples were selected using treatment records (cases and controls for bovine respiratory disease) from cooperating feedlots. Samples were then frozen (-20℃) within 10 h after collection and stored frozen until processing for DNA extraction and pool construction could occur as part of another project goal. From the partially thawed ears, a 0.95 cm diameter sample was collected from each. This sample was then dissected into two samples where approximately one-third of the sample was used for the DNA extraction (Qiagen BioSprint 96 DNA Blood Kit; Qiagen, Germantown, MD) and pool construction. The remaining two-thirds of the sample was retained for future use in other studies. Equal amounts of DNA from groups of 96 animals were then combined into a single pool based on the phenotype of treatment or no treatment for respiratory disease during the feedlot period. Animals were paired in pools based on feedlot lot

allocation so equal pen or lot representation in both case and control pools were achieved for each matching pair. Fifty-three pools were constructed for each phenotype which resulted in 106 pools total. DNA quantity and quality of pools were determined by DNA spectrophotometer (DeNovix DS-11 FX Series; Wilmington, DE) and gel electrophoresis, respectively. DNA pools were then sent to Neogen Corporation (Lincoln, NE) for analysis with the Ilumina BovineHD Bead Array (777,962 SNP; Illumina Inc., San Diego, CA).

3.2.2 Statistical Analysis

For each SNP within pool, a pooling allele frequency (**PAF**; (Yang, Lin et al. 2006) was computed using the following formula:

$$PAF = \frac{Intensisty\ of\ A\ Allele}{(Intensity\ of\ A\ Allele + Intensity\ of\ B\ Allele)}.$$

Individual SNP that did not meet a minor allele frequency (MAF) of 1% were removed from the analysis. If a pool was missing an observation for PAF at an individual SNP, the overall average PAF across all polls with known values was used . Population stratification and outliers were visualized using a neighbor-joining tree based in Euclidian distances among pools using the dist() and ape functions in R (R Core Team, 2023). Once the PAF was computed for each of the pools, relationships amongst the pools were computed by calculating a variance-covariance matrix (**A**) among the pools. For this analysis, the A matrix was computed using the deviation of individual SNP from the SNP average frequency which can be written as $A = [(Y - \bar{\mu}1_n)'(Y - \bar{\mu}1_n)]/\bar{\mu}'(1 - \bar{\mu})$, where Y was a m x n matrix of PAF, where m was the number of SNP, and n was the number of pools, 1_n was a n x 1 vector containing where all elements are 1, and $\bar{\mu} = Y1_n(1_n{}'1_n)^{-1}$ (Keele et al., 2015b). Since the mean in the above equation is calculated based on the data, the A

matrix contained dependencies; thus the A matrix was made positive-definite in the same manner as traditional genomic relationship matrices which are? (VanRaden, 2008).

For the analysis, a full relationship coefficient matrix (A_F) was constructed using all 776,749 SNP after removing SNP for low MAF. Then SNP were randomly sampled from 500 to 770,000 SNP in 500 SNP increments and used to form reduced relationship matrices (A_R) with 50 replications at each level SNP count. To test the equivalence of the matrices, A_R was standardized using Eigenvalue decomposition of A_F with the following equation: $A_{R\ Standardized} = K' A_R K$, where K was an n x n matrix calculated by multiplying the Eigenvectors by diag ($1/\sqrt{Eigen\ Values}$) from A_F. After standardization of the matrix the variation of the Eigenvalues of $A_{R\ Standardized}$ were computed. If the matrices are proportional, then this variation should be equal to 0. To test which level of SNP created an inflection point of a scree plot the average variation across the 50 replications of each SNP level were plotted.

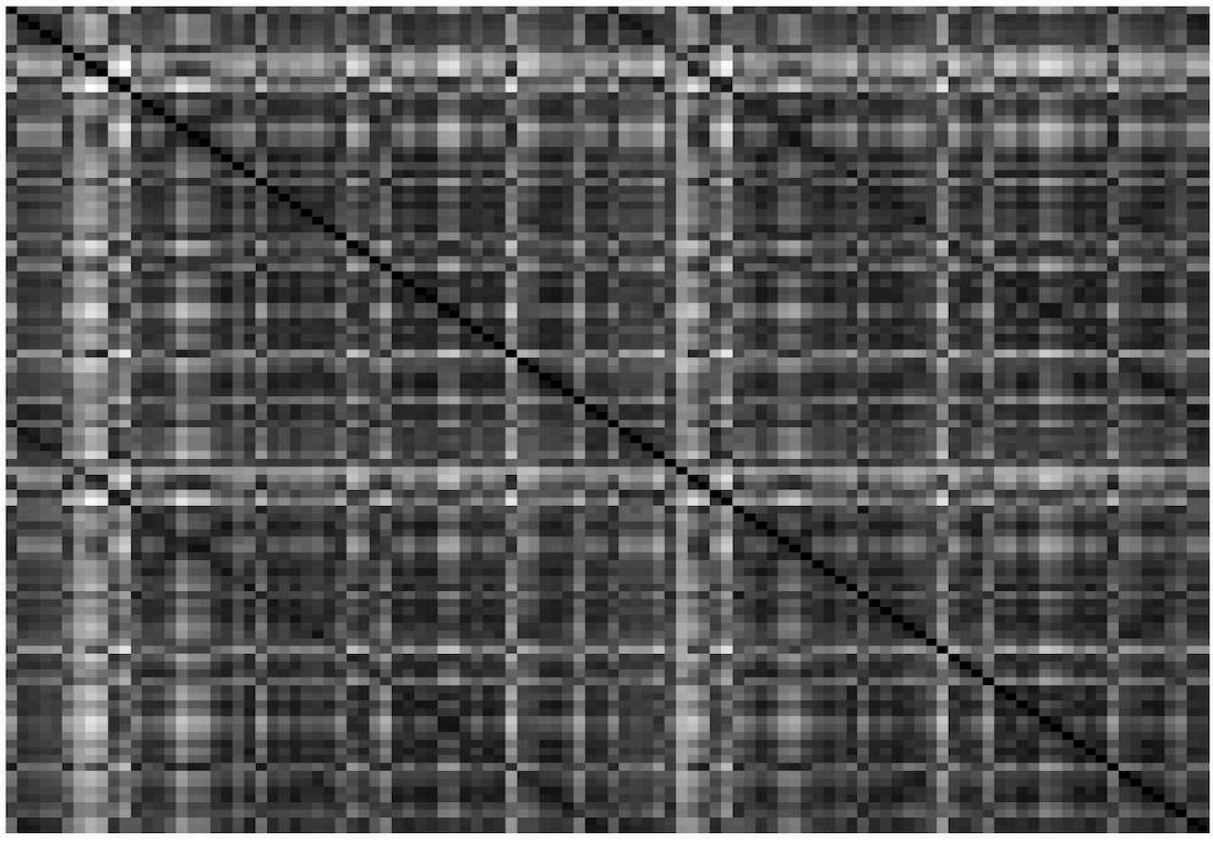
3.3 Results and Discussion

To form PAF, SNP that contained MAF less than 1% were removed from the analysis. For this group of samples, this requirement resulted in the removal of 1,210 SNP. The distribution of the removed SNP was close to equivalent across the chromosomes. The largest number of SNP removed were located on the X chromosome with 85, and the lowest number were Mitochondrial SNP where 6 SNP were removed. After the removal of these SNP there were 776,749 SNP available for the analysis.

A heat map based on Euclidian distances within the relationships among the pools was used to identify outlying pools (Figure 3.1). Since in this experiment each of the pools were matched cases and controls from the same lot, it would be expected that there should be

relationships among the pools. The heatmap of the relationship among pools that had shorter Euclidian distances show up as black in the heat map and as distance increases the color fades to white. To check for outliers using this approach it should be examined to see if there are any pools that consistently have high distances (white color in the heat map) to other pools in the evaluation. Based on the results of the heat map it does not appear that there were any outlier pools that had abnormally high distances from other pools in the analysis. Another visualization technique that can be applied to check for outliers is to examine a neighbor-joining tree.

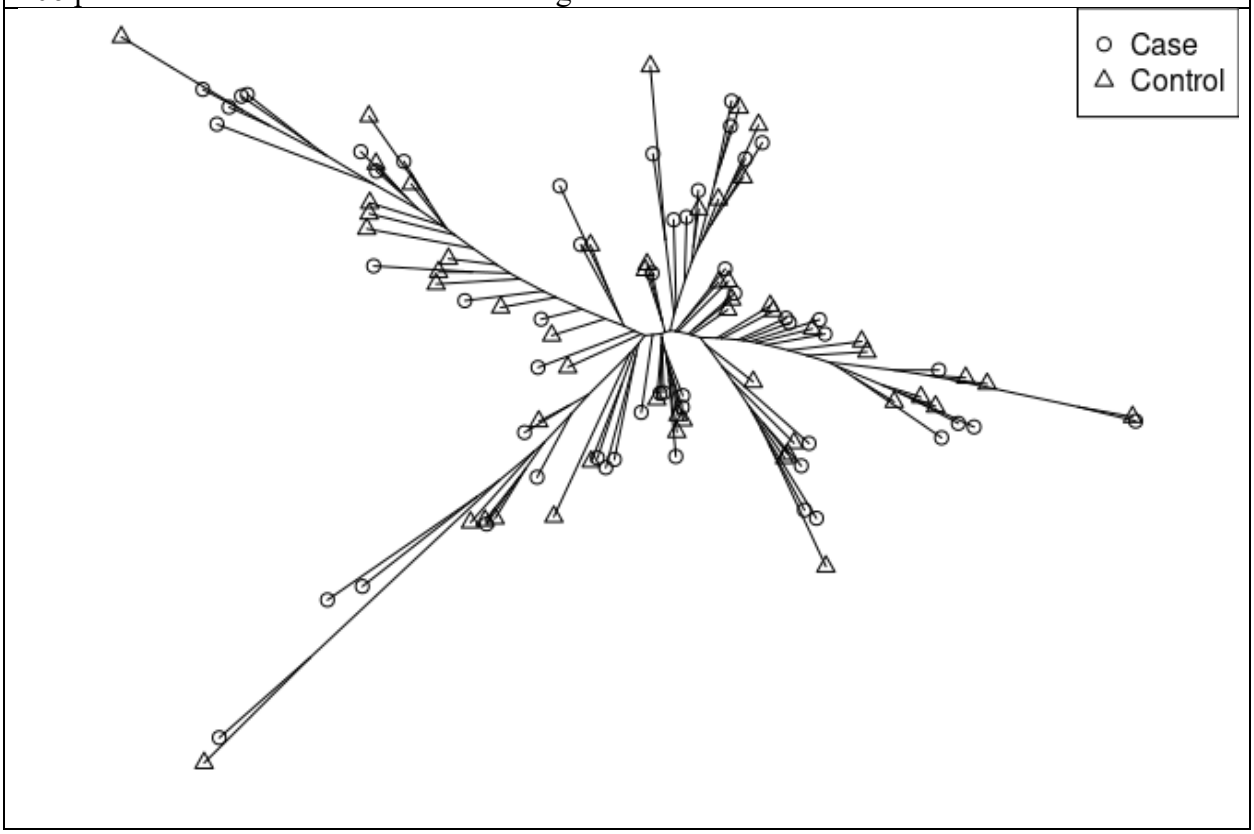
Figure 3.1. Heat map of Euclidian distances among 106 pools. All pools are represented going from left to right and top to bottom the 53 cases are listed first followed by the corresponding control pool.



This plot visualizes the distances between the different pools and is shown in Figure 3.2. Within this plot there are several case-control pairs that appear to diverge from some of the other samples

in the tree, however, these do not diverge far enough to be considered outliers and usually have a matched companion (matched case-control pair).

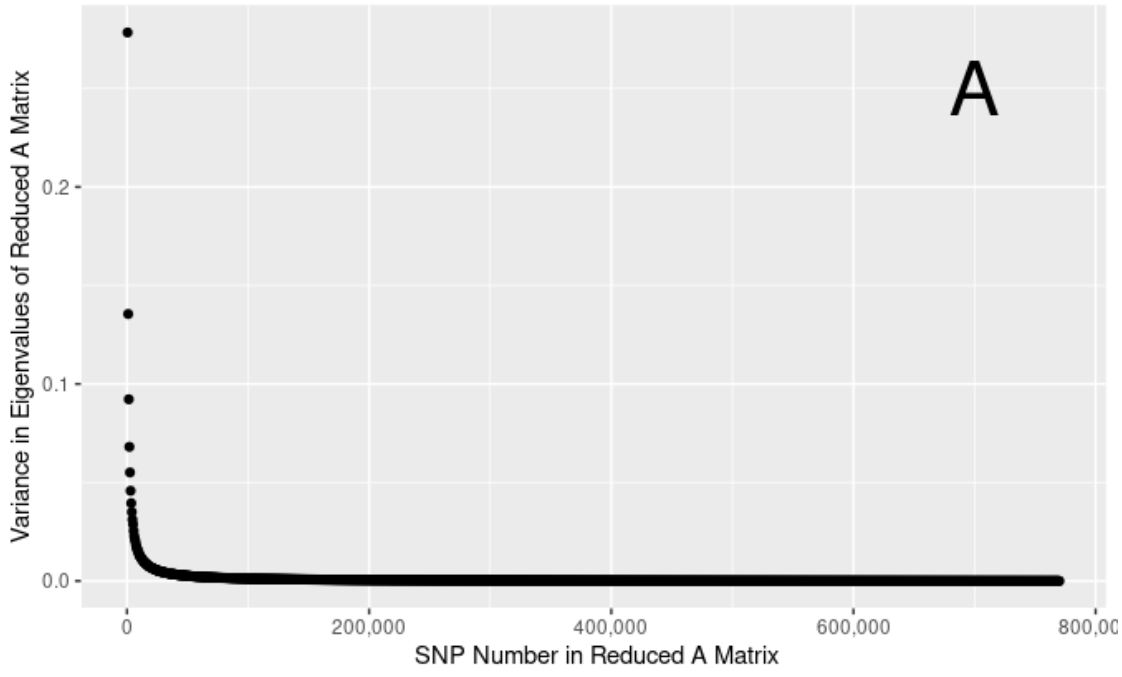
Figure 3.2. Unrooted neighbor-joining tree constructed from the Euclidian distances among the 106 pools. Circles indicate cases and triangles indicated controls.

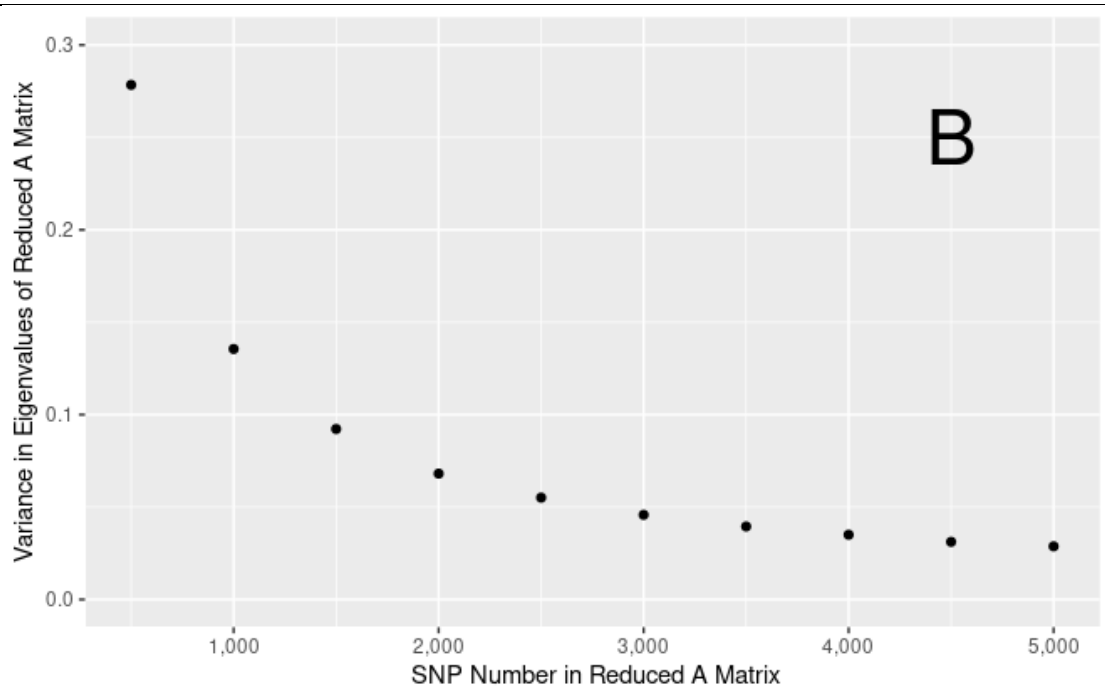


To examine the equivalence of relationship matrices using differing number of SNP, a subset of SNP was sampled randomly across the available SNP ranging from 500 to 770,000 SNP sampled in 500 SNP increments. To better account for differences in sampling 50 replicates for each SNP level were resampled. Under this approach, if the relationship matrices A_F and $A_{R\ Standardized}$ are proportional to one another, the variation in the Eigenvalues should be near 0 and the individual Eigenvalues should be close to 1. To evaluate the different levels the average variation across the 50 replicates of SNP number were plotted. This plot is represented in Figure 3. Across all SNP levels the variation in the Eigenvalues quickly approached 0 which is illustrated

in plot A (Figure 3.3A). To better see how the Eigenvalues decrease, plot B shows the change in variation in Eigenvalues from 500 to 5,000 selected SNP (Figure 3.3B). As the number of SNP increases over 1,600 the amount of variability in the Eigenvalues starts to decrease in a slower rate than at lower levels of SNP sampled. Based on these results using a less dense panel for use in pooling can lead to very similar results when incorporated as a G matrix into genetic evaluations as compared to using higher-density chips. This result can lead to additional cost savings compared to using higher density chips to further incentivize the use of DNA pooling in commercial populations.

Figure 3.3. Plot representing variance in Eigenvalues of relationship matrices constructed from a random sample of reduced SNP ranging from 500 to 770,000 are contained in figure A. B contains a plot of reduced SNP ranging from 500 to 4,600 SNP. For each level of SNP 50 replicates were taken and results were averaged over the 50 replicates.





Rolf et al. (2010) examined the relationship between using a reduced subset of SNP on the genetic prediction of animals for feed intake. It was found that on an individual animal level a subset of 10,000 randomly selected SNP was sufficient for the estimation of a genomic relationship matrix for feed intake. Similar to this analysis, the goal of this study presented herein was to examine if the use of a reduced subset of SNP could be used at a reduced cost to generate more genotyped samples. One of the biggest differences between the Rolf et al. study and this one is the difference between pooled DNA samples and individuals. For effective use in a genetic evaluation, pooled samples must be able to be related to individually genotyped animals, however, in either case, the number of SNP that needs to be included in marker panels could be reduced versus what is currently commercially available.

3.4 Conclusions

The purpose of this analysis was to examine if a reduced subset of SNP could be used to estimate proportional genomic relationship matrices in pooled DNA samples. Results based on Eigenvalue decomposition of the genomic relationship matrix with all available SNP showed that a reduced panel that across 50 replicate samples showed the minimum number of SNP that could be included in a reduced panel. These results offer the opportunity to continue to reduce the cost of DNA pooling and allow for more pools to be constructed across a larger population of animals.

LITERATURE CITED

Alexandre, P. A., L. R. Porto-Neto, E. Karaman, S. A. Lehnert, and A. Reverter. 2019. Pooled genotyping strategies for the rapid construction of genomic reference populations. J. Anim. Sci. 97:4761–4769.

Baller, J. L., S. D. Kachman, L. A. Kuehn, and M. L. Spangler. 2020. Genomic prediction using pooled data in a single-step genomic best linear unbiased prediction framework. J. Anim. Sci. 98:skaa184.

Barcellos, L. F., W. Klitz, L. L. Field, R. Tobias, A. M. Bowcock, R. Wilson, M. P. Nelson, J. Nagatomi, and G. Thomson. 1997. Association mapping of disease loci, by use of a pooled DNA genomic screen. Am. J. Hum. Genet. 61:734–747.

Bell, A. M., J. M. Henshall, L. R. Porto-Neto, S. Dominik, R. McCulloch, J. Kijas, and S. A. Lehnert. 2017. Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. Genet. Sel. Evol. 49:1–7.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57:289–300.

Casas, E., M. Garcia, J. Wells, and T. Smith. 2011. Association of single nucleotide polymorphisms in the ANKRA2 and CD180 genes with bovine respiratory disease and presence of Mycobacterium avium subsp. paratuberculosis 1. Anim. Genet. 42:571–577.

Daniels, J., P. Holmans, N. Williams, D. Turic, P. McGuffin, R. Plomin, and M. J. Owen. 1998. A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. Am. J. Hum. Genet. 62:1189–1197.

Ehret, G. B. 2010. Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. Curr. Hypertens. Rep. 12:17–25.

Griffin, D. 2014. The monster we don't see: subclinical BRD in beef cattle. Anim. Health Res. Rev. 15:138–141.

Guidelines for Uniform Beef Improvement Programs [Internet]. BIF Guidelines Wiki,. 2023. Available from: http://guidelines.beefimprovement.org/index.php?title=Guidelines_for_Uniform_Beef_Improvement_Programs&oldid=2679

Keele, J., L. Kuehn, T. McDaneld, R. Tait Jr, S. Jones, T. Smith, S. Shackelford, D. King, T. Wheeler, and A. Lindholm-Perry. 2015a. Genomewide association study of lung lesions in cattle using sample pooling. J. Anim. Sci. 93:956–964.

Keele, J., L. Kuehn, T. McDaneld, R. Tait Jr, S. Jones, T. Smith, S. Shackelford, D. King, T. Wheeler, A. Lindholm-Perry, and others. 2015b. Genomewide association study of lung lesions in cattle using sample pooling. J. Anim. Sci. 93:956–964.

Keele, J., L. Kuehn, T. McDaneld, R. Tait, S. Jones, B. Keel, and W. Snelling. 2016. Genomewide association study of liver abscess in beef cattle. J. Anim. Sci. 94:490–499.

Krumbiegel, M., F. Pasutto, U. Schlötzer-Schrehardt, S. Uebe, M. Zenkel, C. Y. Mardin, N. Weisschuh, D. Paoli, E. Gramer, and C. Becker. 2011. Genome-wide association study with DNA pooling identifies variants at CNTNAP2 associated with pseudoexfoliation syndrome. Eur. J. Hum. Genet. 19:186–193.

Li, C., J. Basarab, W. M. Snelling, B. Benkel, J. Kneeland, B. Murdoch, C. Hansen, and S. S. Moore. 2004. Identification and fine mapping of quantitative trait loci for backfat on bovine chromosomes 2, 5, 6, 19, 21, and 23 in a commercial line of Bos taurus. J. Anim. Sci. 82:967–972.

McDaneld, T., L. Kuehn, M. Thomas, W. Snelling, T. Smith, E. Pollak, J. Cole, and J. Keele. 2014. Genomewide association study of reproductive efficiency in female cattle. J. Anim. Sci. 92:1945–1957.

McGuirk, S. M. 2008. Disease management of dairy calves and heifers. Vet. Clin. North Am. Food Anim. Pract. 24:139–153.

Miller, S. L., S. Mizell, R. Walker, T. Page, and M. D. Garcia. 2016. Identification of SNPs located on BTA 6 and BTA 20 significantly associated with bovine respiratory disease in crossbred cattle. Genet. Mol. Res. 15.

Neibergs, H. L., C. M. Seabury, A. J. Wojtowicz, Z. Wang, E. Scraggs, J. N. Kiser, M. Neupane, J. E. Womack, A. V. Eenennaam, and G. R. Hagevoort. 2014. Susceptibility loci revealed for bovine respiratory disease complex in pre-weaned holstein calves. BMC Genomics. 15:1–19.

Pacek, P., A. Sajantila, and A.-C. Syvänen. 1993. Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. Genome Res. 2:313–317.

Pardon, B., and S. Buczinski. 2020. Bovine respiratory disease diagnosis: what progress has been made in infectious diagnosis? Vet. Clin. Food Anim. Pract. 36:425–444.

Peiris, B., J. Ralph, S. Lamont, and J. Dekkers. 2011. Predicting allele frequencies in DNA pools using high density SNP genotyping data. Anim. Genet. 42:113–116.

Quick, A. E., T. L. Ollivett, B. W. Kirkpatrick, and K. A. Weigel. 2020. Genomic analysis of bovine respiratory disease and lung consolidation in preweaned Holstein calves using clinical scoring and lung ultrasound. J. Dairy Sci. 103:1632–1641.

Reverter, A., L. Porto-Neto, M. Fortes, R. McCulloch, R. Lyons, S. Moore, D. Nicol, J. Henshall, and S. Lehnert. 2016. Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree. J. Anim. Sci. 94:4096–4108.

Rivas, M. A., M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, and N. Burtt. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. 43:1066–1073.

Rolf, M. M., J. F. Taylor, R. D. Schnabel, S. D. McKay, M. C. McClure, S. L. Northcutt, M. S. Kerley, and R. L. Weaber. 2010. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genet. 11:1–10.

Saleem, A., S. Saleem Bhat, F. A. Omonijo, N. A Ganai, E. M. Ibeagha-Awemu, and S. Mudasir Ahmad. 2024. Immunotherapy in mastitis: state of knowledge, research gaps and way forward. Vet. Q. 44:1–23. doi:10.1080/01652176.2024.2363626.

Sham, P., J. S. Bader, I. Craig, M. O'Donovan, and M. Owen. 2002. DNA pooling: a tool for large-scale association studies. Nat. Rev. Genet. 3:862–871.

Shaw, S. H., M. M. Carrasquillo, C. Kashuk, E. G. Puffenberger, and A. Chakravarti. 1998. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome Res. 8:111–123.

Taylor, B. A., and S. J. Phillips. 1996. Detection of obesity QTLs on mouse chromosomes 1 and 7 by selective DNA pooling. Genomics. 34:389–398.

Timsit, E., N. Dendukuri, I. Schiller, and S. Buczinski. 2016. Diagnostic accuracy of clinical illness for bovine respiratory disease (BRD) diagnosis in beef cattle placed in feedlots: A systematic literature review and hierarchical Bayesian latent-class meta-analysis. Prev. Vet. Med. 135:67–73. doi:10.1016/j.prevetmed.2016.11.006.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

Vargas Jurado, N., L. A. Kuehn, J. W. Keele, and R. M. Lewis. 2021. Accuracy of GEBV of sires based on pooled allele frequency of their progeny. G3. 11:jkab231.

CHAPTER IV

Different Statistical Approaches for Evaluating Pooled DNA Data

4.1 Introduction

When looking at conducting research using genomic data one of the largest costs is obtaining the genomic information on individuals. One way to reduce the cost of obtaining genomic information is using an approach known as DNA pooling, (Sham et al. 2002). Under this approach, DNA samples from multiple individuals are mixed and a single genotyping chip is run on the admixed sample. Instead of returning results with individual loci calls, results are obtained by looking at the proportion of red and green fluorescents to serve as proxies for allele frequency in the pool (Yang et al. 2006). These data can then be analyzed using pools constructed in a case/control manner to compare the allele frequencies at each locus for the pools.

Traditionally when using pooled DNA a single SNP GWAS approach which utilizes F-tests to determine SNP effects and p-values for significance has been used to identify causative variants. Individual Under this approach SNP must reach a high level of significance (low p-value) due to the nature of the multiple testing applied to many loci. Therefore, the multiple testing may exclude SNP that are informative due to the high threshold that needs to be achieved to p-values that would qualify as significant. Despite some of these challenges, this approach has previously been applied successfully to identify SNP that were significant for traits in beef cattle (McDaneld et al. 2014, Keele et al. 2015, Keele et al. 2016). However, additional approaches are available that have not widely been applied to pooled DNA data in beef cattle. Some of these approaches include

machine learning algorithms such as Random Forest (Breiman, 2001) that can be applied to these data sets. Previous research has not directly compared these approaches to see if these alternative approaches may provide improvement in the identification of SNP that are significant in pooled DNA experiments. The purpose of this study is to examine the alternative approach of Random Forest to a dataset that mimics pooled DNA.

4.2 Materials and Methods

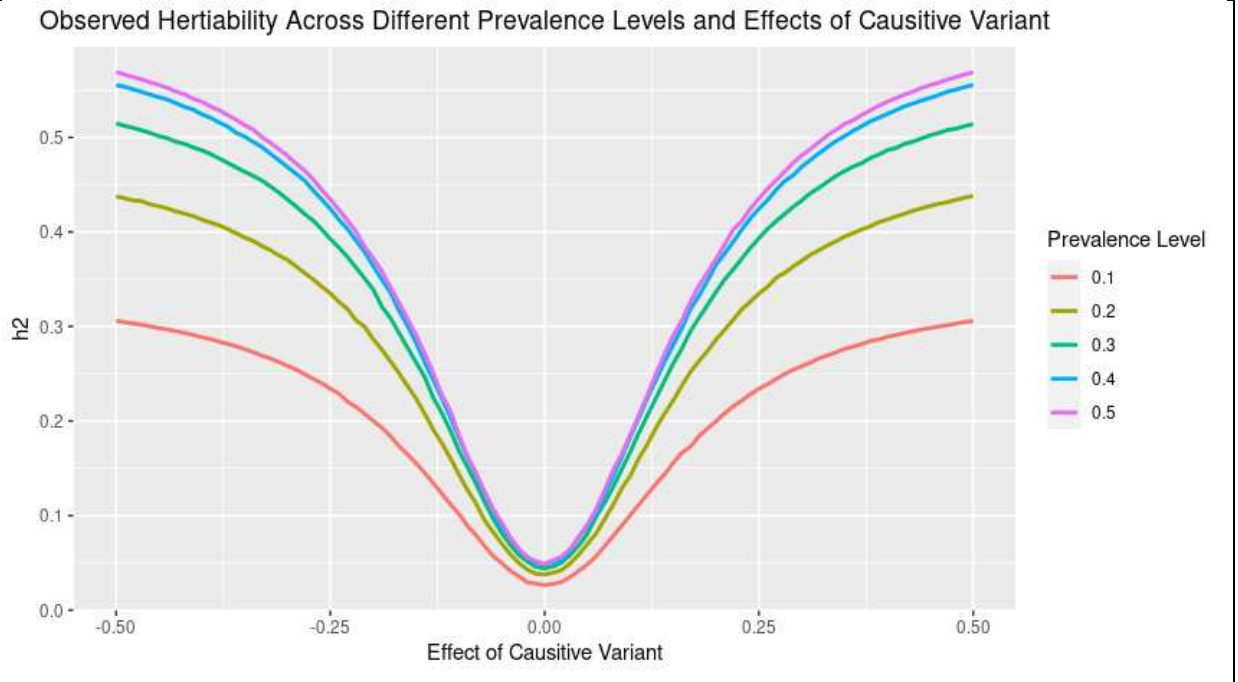
For this study, the interest was the evaluation of different statistical approaches' ability to correctly identify 100 random SNP that were simulated to be significant. To mimic data used in pooled DNA evaluation, the number of SNP loci were equivalent to those available on the Illumina Bovine HD Array (5200 Illumina Way, San Diego, CA 92122) which represents approximately 770,000 SNP across the bovine genome. The data for this evaluation were simulated using R statistical software (R Core Team 2020). All pooling allele frequencies (PAF) were simulated using a beta distribution. Under this approach the assumption was that there was no linkage disequilibrium among the SNP. Then a list of 100 randomly chosen SNP were selected to be the causative variants for the binary phenotype. The simulated frequency of these loci was then modified for the assumed binary phenotypic category based on levels of prevalence and the effect of the allele. The level of disease prevalence simulated ranged from 0.10 to 0.50 in increments of 0.10. The effect that each SNP had on susceptibility to the disease ranged from -0.50 to 0.50 in increments of 0.01. This resulted in 505 unique combinations of prevalence and effect. On the liability scale 100 replications were run to estimate heritability. For the simulation of heritability on the underlying scale the assumption for the calculation of genetic variance is 2p(1-p) a^2

(Robinson et al., 2014) and using a probit model and the assumed residual variance of 1 (Williams, 2009). In the above equation p represents the frequency of the causal variant and a is the effect size. For simulation the distribution for p~β(1,1) and a ~ N(μ_a, σ_a^2). For each level of a mean heritability was calculated on the liability scale. This was then converted to the observed scale was calculated using the following equation (Visscher et al., 2008):

$$h_O^2 = h^2 z^2 / [K(1-K)]$$

Where h_O^2 is heritability on the observed scale, h^2 is heritability on the liability scale, z is the height of the normal curve that truncates the proportion K. Visual representation of resulting observed heritability is included in Figure 4.1.

Figure 4.1. Heritability estimates of binary traits on the observed scale given different levels of effect of the disease causing allele and prevalence of the trait resulting from 100 simulations of heritability on the liability scale.



For relevance, we determine the range of effect and prevalence values corresponding to observed heritability of 30 % or less. Considering observed heritability as 30 % or less resulted in a range of causative variant effects of -0.45 to 0.46, -0.21 to 0.21, -0.17 to 0.17, -0.15 to 0.16, and -0.15 to 0.15 for prevalences of 0.10, 0.20, 0.30, 0.40 and 0.50, respectively. The range of relevant effect sizes decrease with increasing prevalence. (figure here maybe, effect ranges on y and prevalence on x) After establishing thresholds for prevalence and effect of the causative variants 100 SNP were randomly selected to represent the causative variants.

We start with a list of 100 causative variants at different chromosomal loci (or positions) each locus having the same effect on disease liability. Allele frequency (p) in the overall population (including both cases and controls) varies by locus but we consider it to be fixed and not stochastic even though we sample these frequencies from a beta distribution for convenience but only once per replicate of the population. Next, we envision fixed allele frequencies for cases and control subpopulations which depend on p, prevalence and genetic effects. To make case and control allele frequencies fixed and non-stochastic, we compute them from half of the average genotype (coded as 0, 1 and 2 copies of B allele) for cases (p_{case}) and controls ($p_{control}$) sampled from a large population of 1 million animals. Finally, we simulate stochastic pooling allele frequences (paf_{case} and $paf_{control}$) for pools of 96 animals in the pool using the beta distribution with shape1 = $2 * p_{case} * 96$ and shape2 = $(1 - p_{case}) * 96$ for cases; and shape1 = $2 * p_{control} * 96$ and shape2 = $2 * (1 - p_{control}) * 96$ for controls. In this simulation, stochastic pooling allele frequencies depend on fixed effects of overall population allele frequency, affects on liability and prevalence and random effects of sampling animals to go into pools. Genotypes (coded as 0, 1 or 2 copies of B allele) for each animal and causative variant were drawn randomly from a binomial distribution with the number of trials set to 2 and the probability of receiving a B allele equal to the overall

population allele frequency (p). The breeding value of an individual was the product of the genotype (centered at 0) and the effect summed over all causative variants. Phenotype on the underlying scale disease liability scale was simulated using a normal distribution with the mean equal to the breeding value and a standard deviation of 1. Animals were a case if their liability was ≥ 1 – prevalence quantile of the liability distribution and a control otherwise.

For each generated data set, two models were applied to determine their ability to correctly identify significant SNP. The first type of analysis was a Genome Wide Association Study (**GWAS**) that compared differences in PAF for each SNP loci using an F test. Methods for this analysis were similar to (McDaneld et al. 2014). The dependent variable was the PAF and a binary phenotype was the independent variable. To account for differences in the simulated SNP, the average variance-covariance matrix (A) across SNP over the autosomal genome among pools was estimated using the cov() function in R. The F test with a numerator df of 1 and denominator df of 104 (106 pools – 1 df for the difference between pools – 1 df for the mean) was used to calculate p-values using the pf() function in R. Results were then ordered from most significant (smallest p-value) to least significant (largest p-value) and the 100 most significant SNP were extracted. This extracted list was then compared to the list of 100 pre-identified SNP to see how many were correctly identified as significant.

The second analysis type applied to the data was a Random Forest. A Random Forest analysis is a tree-based machine learning tool that has previously been used to identify SNP that are significant for disease (Meng et al., 2009; Schwarz et al., 2010; Pudjihartono et al., 2022). For this analysis a classification Random Forest was used. The list of 100 or 106 PAF for each SNP was used as the independent variable and disease status was used as the dependent variable. To determine the hyperparameters used in all the Random Forest analyses, the dataset that contained

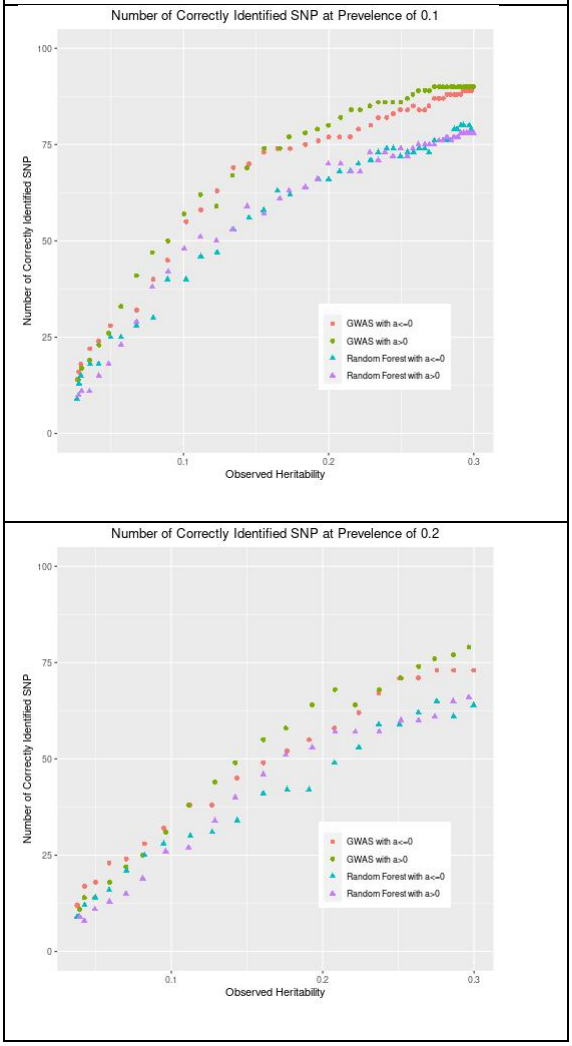
a prevalence of 0.20 and an effect of 0.05 was used to perform a grid search on the optimal number of explanatory variables to test for each tree (10,000 to 100,000 by 10,000), the minimum terminal node size (1 to 10 by 2), and the number of trees to include in the Random Forest (1,000 to 10,000 by 1,000). For each of the unique combinations of these variables, an out-of-the-box root mean square error was computed and the combination that minimized this was used as the hyper-parameters for all the analyses including different prevalence and gene effects as the GWAS analyses. To identify which SNP were most influential in the Random Forest analyses, Variable Importance (VIMP) was calculated for each analysis. The VIMP for each analysis was identified using a Gini-Index. For each repetition of the Random Forest model, the 100 SNP identified based on VIMP were extracted and compared to the pre-determined list of causative SNP.

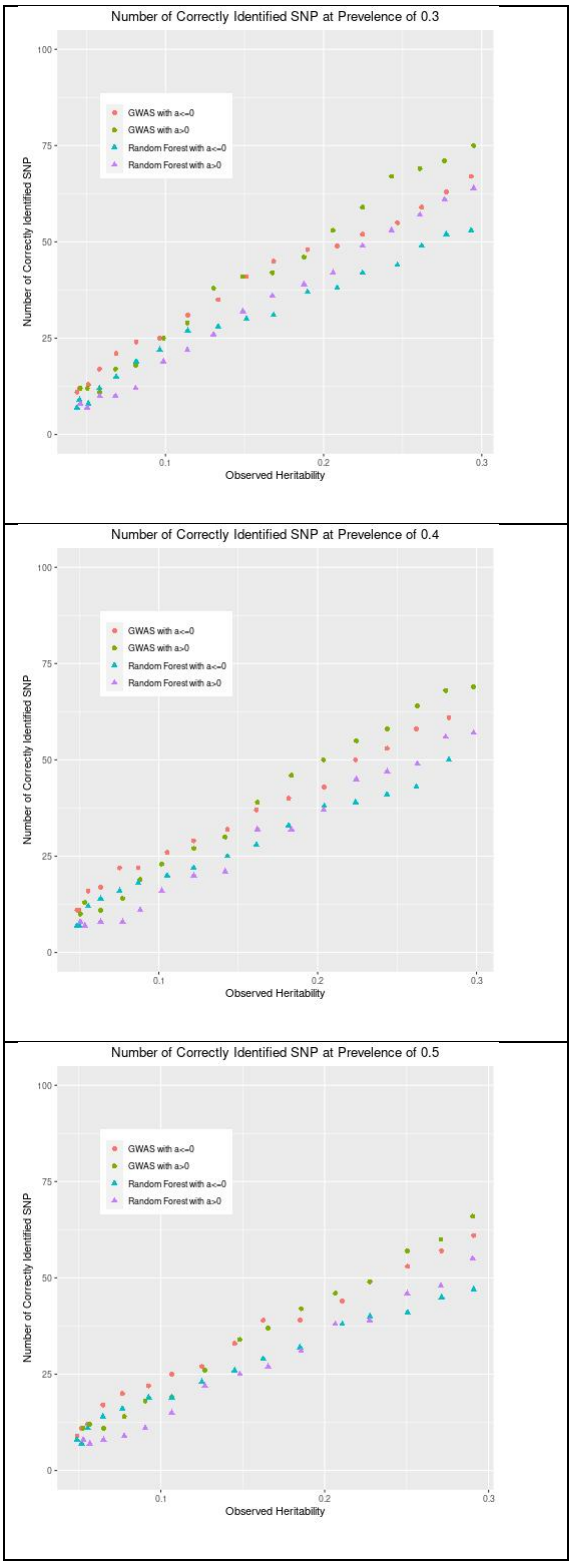
4.3 Results and Discussion

Across levels of observed heritability, both analysis types were able to identify a subset of the functional variant. Figure 2 illustrates the results of the number of correctly identified functional variants in the lowest 100 p-values for traditional GWAS and top 100 based on highest VIMP. For all levels of prevalence at lower levels of observed heritability (<=0.10) both models were able to identify a similar number of SNP. As the observed heritability of the trait increased, traditional GWAS procedures using an F-test were able to identify a greater proportion of the functional variants compared to the number identified by Random Forrest at the same level of observed heritability. For both analyses the number of identified variants was inverse of the level of prevalence. This was due to the increased effect that was simulated to achieve the same level of observed heritability. The approach that was able to most correctly identify the largest number of SNP was the traditional GWAS model at a prevalence of 0.10 with 90 correctly identified

functional variants. This was achieved when the level of effect for the function variant exceeded 0.34 when the effect was positive.

Figure 4.2. Plots of number of correctly identified SNP using a Traditional GWAS and Random Forest analysis at varying levels of prevalence and effect of the causative variant.




Number of Correctly Identified SNP at Prevelence of 0.3
Number of Correctly Identified SNP
Observed Heritability
GWAS with a<=0
GWAS with a>0
Random Forest with a<=0
Random Forest with a>0
Number of Correctly Identified SNP at Prevelence of 0.4
Number of Correctly Identified SNP
Observed Heritability
GWAS with a<=0
GWAS with a>0
Random Forest with a<=0
Random Forest with a>0
Number of Correctly Identified SNP at Prevelence of 0.5
Number of Correctly Identified SNP
Observed Heritability
GWAS with a<=0
GWAS with a>0
Random Forest with a<=0
Random Forest with a>0

Roshan et al., (2011) examined the difference between 1 df chi-squared, support vector machines, and random forests to identify causal variants in simulated and real genotype data using a case and control design for disease. Their results were that the support vector machine and random forest performed superior when looking at ranking the causal variants on their effect of the disease. While there are similarities to the current study there are several key differences. While the current study also applies a single SNP approach as well as, random forest, the SNP rankings in the cited study applied to a reduced number of SNP that were a multiple of two to ten times the number of SNP that were identified after a Bonferroni correction based on all SNP (Hochberg, 1988). When lower number of SNP were provided to the two multivariate analyses their performance was superior to the 1 df chi squared approach. However, when the number of SNP supplied to the model was 10 times the number of significant SNP after multiple testing all of the analyses performed more similarly. These results may help to explain some of the results that we see in the current study. Since in the current study all SNP were applied to both models, the multi-variate model may have not performed as well as, if a smaller subset was used.

One advantage of using a machine learning approach over the traditional f test is that the machine learning approach can also better account for relationships that occur among SNP such as linkage (Touw et al., 2013). Within the parameterization of the model is to identify the number of predictor variables that should be considered to identify and make decision trees (Chen and Ishwaran 2012). In the case of pooled genotype data, this would be each of the different SNP. In this experiment, the same number of SNP were considered at each decision tree. This represented roughly one-third of the total available SNP. One challenge with this as well as other experiments is that although the number of SNP being considered is large, there is a larger probability that all non-significant SNP could be randomly selected. However, if another SNP is in LD with a

significant SNP, then these decision trees may still be able to glean information. In this experiment, there was no LD among the SNP included. All non-significant SNP were distributed at random to help to see if the model could identify only the significant SNP. As a counter argument, in the random forest a random subset of predictor variables which in the case of GWAS would be individual SNP and that individual classifiers are weak predictors and the culmination of the entire set of classifiers is what makes the random forest a powerful prediction algorithm (Touw et al., 2013). However, in the instance that we have simulated in the study it appears that inclusion of the causal variants was important to identification of the variables.

Traditionally GWAS studies have been shown as be useful in identification of common variants (minor allele frequency >=5%) which can be shown to contribute to inherited components of a binary phenotype that could represent disease (McCarthy et al., 2008). In the simulated data the assumption would be that variants were common across the population. When the effect of these alleles was high (higher observed heritability) and common this situation would be highly unusual for many situations pertaining to diseases (McCarthy et al., 2008). The alternative to this would be rare variants in the population that tend to have higher effect on the outcome compared to the common variants identified via GWAS. However, in the situations where the effect of the common causative variant is low illustrated in this study by low observed heritability then detectability by GWAS analysis is difficult (Zeggini et al., 2005). In the current study this is illustrated at every level of assumed prevalence where less than 25% of causative variants were identified at where the effect was assumed to be less than the absolute value of 0.07. This would correspond to observed heritability that ranges from 0.03 to 0.10. Given this result it would be difficult for either tested model to differentiate many small effect causative variants that would contribute to the binary outcome. In this example, the study of rare variants rather than common

ones may be a more appropriate method to be able to accurately identify variants that differentiate individuals for a binary phenotype (Visscher et al., 2012).

4.4 Conclusions

In this experiment, neither the Random Forest nor traditional GWAS model were able to identify all 100 of the significant SNP at any combination of prevalence or effect of a disease-causing allele. The highest number of correctly identified SNP in both models were identified when the effect of the causative variant was large. At lower levels of effect of the causative variants both the models performed similarly. Overall, the traditional GWAS model appears to be superior for identification of causal variants.

LITERATURE CITED

Breiman, L. 2001. Random forests. Mach. Learn. 45:5–32.

Chen, X. and H. Ishwaran.2012. "Random forests for genomic data analysis." Genomics **99**(6): 323-329.

Hochberg, Y. 1988. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 75:800–802.

Keele, J. W., L. A. Kuehn, T. G. McDaneld, R. G. Tait, S. A. Jones, B. N. Keel and W. M. Snelling. 2016. "Genomewide association study of liver abscess in beef cattle." J Anim Sci **94**(2): 490-499.

Keele, J. W., L. A. Kuehn, T. G. McDaneld, R. G. Tait, S. A. Jones, T. P. Smith, S. D. Shackelford, D. A. King, T. L. Wheeler, A. K. Lindholm-Perry and A. K. McNeel. 2015. "Genomewide association study of lung lesions in cattle using sample pooling." J Anim Sci **93**(3): 956-964.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. 9:356–369.

McDaneld, T. G., L. A. Kuehn, M. G. Thomas, W. M. Snelling, T. P. Smith, E. J. Pollak, J. B. Cole and J. W. Keele. 2014. "Genomewide association study of reproductive efficiency in female cattle." J Anim Sci **92**(5): 1945-1957.

Meng, Y. A., Y. Yu, L. A. Cupples, L. A. Farrer, and K. L. Lunetta. 2009. Performance of random forest when SNPs are in linkage disequilibrium. BMC Bioinformatics. 10:1–17.

Pudjihartono, N., T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan. 2022. A review of feature selection methods for machine learning-based disease risk prediction. Front. Bioinforma. 2:927312.

Robinson, M. R., N. R. Wray, and P. M. Visscher. 2014. Explaining additional genetic variation in complex traits. Trends Genet. 30:124–132.

Roshan, U., S. Chikkagoudar, Z. Wei, K. Wang, and H. Hakonarson. 2011. Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. Nucleic Acids Res. 39:e62–e62.

Schwarz, D. F., I. R. König, and A. Ziegler. 2010. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. Bioinformatics. 26:1752–1758.

Sham, P., J. S. Bader, I. Craig, M. O'Donovan and M. Owen. 2002. "DNA Pooling: a tool for large-scale association studies." Nat Rev Genet **3**(11): 862-871.

Team, R. C. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Touw, W. G., J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. van Hijum. 2013. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief. Bioinform. 14:315–326.

Visscher, P. M., M. A. Brown, M. I. McCarthy, and J. Yang. 2012. Five years of GWAS discovery. Am. J. Hum. Genet. 90:7–24.

Visscher, P. M., W. G. Hill, and N. R. Wray. 2008. Heritability in the genomics era—concepts and misconceptions. Nat. Rev. Genet. 9:255–266.

Williams, R. 2009. Using heterogeneous choice models to compare logit and probit coefficients across groups. Sociol. Methods Res. 37:531–559.

Yang, H. C., C. H. Lin, S. I. Hung and C. S. J. Fann. 2006. "A Comparison of Individual Genotyping and Pooled DNA Analysis for Polymorphism Validation Prior to Large-Scale Genetic Studies." Annals of human genetics **70**(3): 350-359.

Zeggini, E., W. Rayner, A. P. Morris, A. T. Hattersley, M. Walker, G. A. Hitman, P. Deloukas, L. R. Cardon, and M. I. McCarthy. 2005. An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. Nat. Genet. 37:1320–1322.

CHAPTER V

Genomewide Association Study for Animals Treated for Bovine Respiratory Disease During the Finishing Period Using Pooled DNA

5.1 Introduction

Bovine Respiratory Disease Complex (BRDC) is one of the costliest diseases that currently affect beef cattle production. It is estimated that this disease complex costs the industry over one billion dollars annually. This cost is not only associated with the cost of treating the disease but also with lost production efficiency of animals suffering from symptoms. While BRDC can affect all classes and ages of cattle, the most common manifestation of the disease occurs during the feedlot stage of production. Estimates state that in feedlots with a capacity of 1,000 head or more, 16.2% of animals developed BRDC symptoms (NAHMS 2011). The cost of treating these affected animals averaged $23.60 per case (APHIS 2013). While strict protocols have been developed for the prevention and treatment of BRDC, incidence of the disease has remained constant.

A potential approach to reduce the cost of the disease on the industry that could be used is through the genetic selection of animals that are less likely to be affected by BRDC. One of the challenges with this approach is the tracking of data and information from the commercial level to seedstock animals. An alternative approach to allow for selection against this trait is to identify genomic markers that are effective in predicting resistance to the development of BRDC symptoms in a commercial setting.

Pooling DNA has previously been used to conduct Genome-wide Association Studies (GWAS) and can be used to reduce the overall cost of genotyping for this type of research (Sham

et al., 2002). Using this approach, animals are selectively grouped, and the composite sample is used to generate genomic information for analysis. This approach has previously been applied to study disease in humans (Pacek et al., 1993; Barcellos et al., 1997; Daniels et al., 1998; Shaw et al., 1998; Krumbiegel et al., 2011; Rivas et al., 2011), as well as, in animals (Taylor and Phillips, 1996; Keele et al., 2015a; Keele et al., 2016). Each of the above examples were applied in a case-control strategy where DNA chip intensity from the pools of each class can be compared. These different florescent intensities can serve as a proxy for allele frequency in the pool. Significant differences among these allele frequency proxies allow for the identification of significant genomic regions. The objective of this paper is to identify causative variants for animals that were treated for BRDC during the feeding period.

5.2 Materials and Methods

5.2.1 Sample Collection

Animal care and use committee approval was not required for this study as DNA sampling procedures were collected on commercial carcasses after harvest and no live animals were used.

Ears were sampled from three large commercial beef processing plants in Colorado and Nebraska. Each of the animals sampled were pre-identified from feedlots having a minimum treatment rate within the pen. When a pen was identified to be sampled, the entire group was sampled when the animals were harvested. A sample of the ear was collected that included individual identification information for each animal. After ear samples were collected at the plant they were transported and frozen (-20°C) until DNA samples for pools could be extracted.

Animals sampled for the study were collected over a four-year period. A majority of the samples were collected during the spring of the year, but samples were also collected periodically at other times of the year as well. No requirements or background information were provided on the animals. So, animals were a representative sample of commercial cattle. For each animal that was collected, treatment history from the feedlot was provided. Any samples that were collected that were where the entire pen was treated were discarded and not included in a pool. Animals were identified to be cases if they were either diagnosed with respiratory disease or were administered antibiotics used to treat BRDC. Matching controls were sampled from the same animals that were identified as being from the same lot/pen by the feedlot. For each ear that was used for the pooling, two ¼" punches were extracted from each ear. Then one of the punches was dissected to have 1/3rd of the punch available for DNA extraction and pooling, and the remaining 2/3rd of the first punch and the second punch were retained as secondary samples. These samples were the first 96 animals that were treated and were considered the first case pool, the second 96 were the second case pool, etc. The exact same process was followed for control pools where the number from each lot in the matching control pool was the same. In total there were 106 pools with equal number of case and controls represented. Pools were formed by first extracting DNA on each individual and then combining the samples based on equal concentrations of each individual in the pool. Sample pools were sent to Neogen (Lincoln, NE) for analysis with the Illumina Bovine HD genotyping array which contains makers approximately 770,000 SNP.

5.2.2 Single SNP GWAS Statistical Procedure

The statistical procedures used in this analysis are similar to previously described methods in (McDaneld, Kuehn et al. 2014, Keele, Kuehn et al. 2015, Keele, Kuehn et al. 2016). For each pool in the analysis, a pooling allele frequency (PAF; (Peiris et al., 2011)), which is a ratio of the intensity of the normalized red fluorescence divided by the intensity of the normalized red and green fluorescence, was calculated. The results of the PAF calculation served as an estimate of the A allele frequency in the pool. Within the PAF calculations, SNP were removed if they had a minor PAF of less than 1% or if observations were missing from all pools in the analysis. For pools that had SNP loci that were missing or not called, the average PAF across all pools were used. To evaluate the pool stratification, pools were visualized using a neighbor-joining tree. This was constructed from Euclidean distances among pools using all SNP in the ape and dist functions in R (R Core Team, 2023).

To account for population stratification and technical errors common to all SNP for a specific pool, a covariance matrix (A) among the pools were computed using all SNP. This approach is modeled after the approach presented by VanRaden (2008) for calculating genomic relationship matrices using SNP data. The covariance was estimated using the PAF deviations from the SNP average. The formula for calculating A was $A= [(Y-\bar{\mu}1_n)' (Y-\bar{\mu}1_n)]/ \bar{\mu}'(1-\bar{\mu})$, where Y was a m x n matrix of PAF, m was the number of SNP, n was the number of pools, 1_n was a n x 1 vector with each element equal to 1, and $\bar{\mu}= Y1_n(1_n' 1_n)^{-1}$. With this approach, there is a dependency that is part of A caused by $\bar{\mu}$ from the data which causes A to not be positive definite. To overcome this, the diagonal elements of the matrix were multiplied by 1.01 to make the matrix positive definite (VanRaden 2008).

Mixed model methodology was used to solve for effects in the model. Y was assumed to be distributed as multivariate normal with a mean for if the pool was identified as either a case or control with a variance proportional to A. There for y~MVN($X\beta, A\sigma^2$), where MVN is the multivariate normal distribution and σ^2 was a SNP specific multiplication factor. X was a n x 2 matrix with the first column identifying case pools with a value of 1 and controls as a value of 0, the second column is the opposite where controls are identified with a value of 1 and cases are identified with a value of 0. The mean PAF results for both cases and controls are in β. To test for SNP significance values in β were computed between cases and controls PAF. This resulted in performing an F-test and the resulting P-values with 1 numerator degree of freedom and n-2 denominator degrees of freedom. To control for multiple testing, the Bejamin-Hockburg procedure was applied to all resulting p-values to test for significance (Benjamini and Hochberg, 1995).

5.2.3 SNP Groups Statistical Procedure

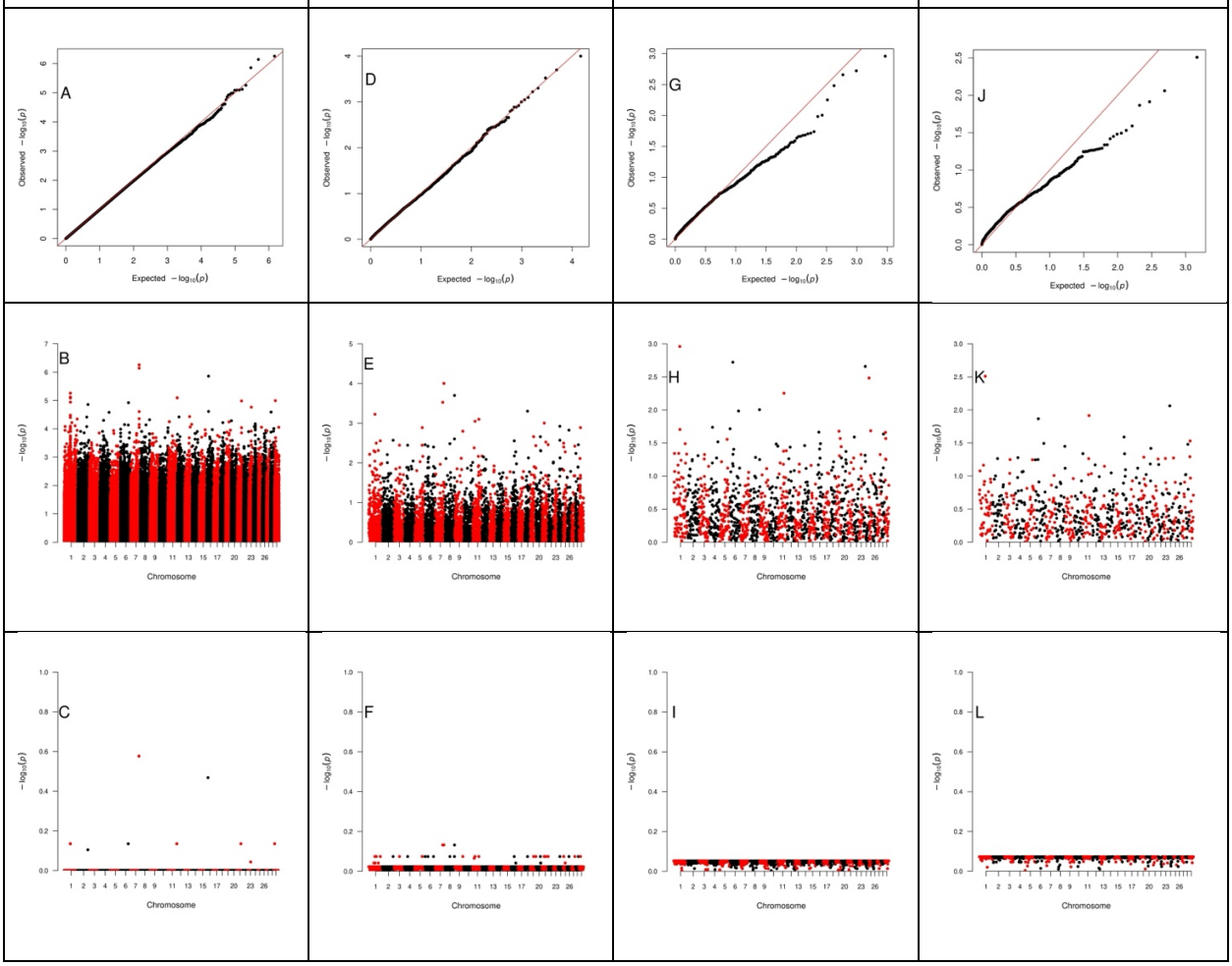
Data were also analyzed using genomic regions containing 100, 500, and 1,000 SNP regions. Each autosome was divided where the corresponding number of SNP were divided among the windows so that each window represented equal numbers of sequential SNP based on position on the chromosome. Then a distance matrix was calculated based on the A matrix of the PAF for the SNPs in each window. This was used as the response variablein an analysis of variance analysis using the adonis package in R. Fixed effects in the model included a distance matrix of the A matrix and a 2x106 binary matrix indicating if a pool was sick or healthy. As part of the analysis pseudo F ratios were used to test if there was differences in the distance matrix within the specific window between cases and controls. These ratios had corresponding p-values that could test for significance.

5.3 Results and Discussion

Figure 5.1 parts A, D, G, and J show the QQ plot for the observed p-values versus expected p-values. This plot examines the relationship between the observed and the theoretical chi-squared distribution (Ehret, 2010). Under the null hypothesis, values would follow the red line in the middle of the graph and deviations from this would signify there are differences in population stratification. Across the range of p-values the expected and observed are very similar suggesting there are no systematic effects affecting the results of the analysis. In Figure 5.1 part B of Figure 5.1, the Manhattan plots of the unadjusted p-values are presented for the single SNP analysis. In this graph there are regions that reach higher levels of significance >5-$\log_{10}$ on BTA 1, BTA 7, BTA 11, and BTA 16 when adjustments for multiple testing are not considered. For the analysis using 100 SNP windows (Figure 5.1 part E) the locations of unadjusted p-values that show the highest level of significance (>3-$\log_{10}$) are located on BTA1, BTA7, BTA8, BTA11, BTA 18, and BTA 21. For the analysis using 500 SNP windows (Figure 5.1 part H) the locations of the unadjusted p-values that show the highest level of significance (>2.5-$\log_{10}$) are located on BTA1, BTA6, and BTA24. For the analysis using 1000 SNP windows (Figure 5.1 part K) the locations of unadjusted p-values that show the highest levels of significance (>2-$\log_{10}$) are located on BTA1 and BTA24. However, in Figure 5.1 when p-values are adjusted using the Benajamin-Hockburg procedure (Figure 5.1 parts C, F, I, and L) none of the SNPs reached the level of genome wide significance when accounting for multiple testing at $\alpha < 0.05$. Previous studies have been able to identify significant associations between data from pooled DNA and other disease traits in beef cattle (Keele et al., 2015a; Keele et al., 2016) as well as other important beef cattle traits (McDaneld et al., 2014). While the current study was not successful in identification of significant SNP, the use of this approach could be very useful for other traits in the future, especially those

that are more commonly recorded on groups. Given the complex nature of BRD it is not hard to imagine that one of the potential reasons that genomic regions were not identified is based on the specific pathogen that is affecting an animal may influence the resulting important genomic regions. In the current study this information was not available since phenotypes were identified based on treatment records.

Figure 5.1. QQ plots of expected p-values and observed p-values for analyses conducted on single SNP, 100 SNP windows, 500 SNP windows, 1000 SNP windows are in the first row. Manhattan plots of unadjusted P-values for single SNP analysis, 100 SNP windows, 500 SNP windows, and 1000 SNP windows are contained in row 2. Manhattan plot of Benjamin-Hockburg adjusted p-values for single SNP, 100 SNP windows, 500 SNP windows, and 1000 SNP windows are contained in row 3.

Single SNP Analysis	100 SNP Window Analysis	500 SNP Window Analysis	1,000 SNP Window Analysis
			

Previous studies that have also individually genotyped animals and performed GWAS to identify SNPs that are influential for susceptibility to BRD. Neibergs et al. (2014) performed GWAS on groups of Holstein calves which were classified as healthy or sick based on the McGuirk Health Scoring System (McGuirk, 2008). Four different analysis types were performed, and SNPs were ranked based on significance of association and then separated into megabase regions with

the lowest ranking SNP representing the score for that region. In the combined analysis for calves from both locations were located on BTA16 (70-71), BTA14 (7-8), BTA18 (65-66), BTA12 (77-78), BTA5 (23-24), BTA13 (56-57), BTA2 (2-3), BTA13 (71-72), BTA16 (64-65), BTA28 (26-27), BTA21 (47-48), BTA25 (22-23), BTA13 (67-68), BTA10 (28-29), BTA17 (72-73), BTA14 (9-10), BTA21 (50-51), BTA19 (26-27), BTA14 (10-11), BTA8 (63-64), BTA13 (6-7), and BTA13 (53-54). Table 5.1 reports the rankings of the smallest p-values for each of the regions using the different analysis techniques. For the single SNP analysis all but BTA16 (70-71), BTA16 (70-71), BTA14 (9-10), and BTA8 (63-64) had SNP that were in the top 1% SNP. For the analysis with containing windows none of the regions are among the top 1% of significant windows. The results of the single SNP analysis suggest that there is similar signal in the results although none have reached genome wide significance. However, the window analyses are not conclusive and do not show that when additional SNP are included that the significance of the regions is not as strong.

Quick et al., (2020) using a muti-omics approach identified two SNPs that were significantly associated with BRD in 143 muti-breed beef cattle. The first SNP was located on Chr5:25858264. For the current study there was one SNP that was within a kb of the identified SNP with location Chr5:25858337. In the current study the adjusted p-value was >0.99, 0.95, 0.91, and 0.86 for analysis with single SNP, 100 SNP windows, 500 SNP windows, and 1,000 SNP windows, respectively. The second SNP that was identified was BovineHD1800016801. In the current study the adjusted p-value for this SNP was >0.99, 0.95, 0.89, and 0.86 for analysis with single SNP, 100 SNP windows, 500 SNP windows, and 1,000 SNP windows, respectively. These SNP did not appear to have differentiation of allele frequency in cases and controls in the current study.

Table 5.1 P-value rankings for one mega base regions that were identified by Neibergs et al. (2014). Four analyses types were applied to the data set either looking at single SNP or grouped SNP analeses. For each analysis ranking for the SNP/region that fell within the identified regions and was the lowest (smallest p-value) was reported.

Location CHR (MB Region)	**Ranking Single SNP**	**Ranking 100 SNP Regions**	**Ranking 500 SNP Regions**	**Ranking 1,000 SNP Regions**
BTA16 (70-71)	13,639	3684	804	305
BTA14 (7-8)	**675**	351	449	298
BTA18 (65-66)	**3,290**	1632	1024	449
BTA12 (77-78)	**942**	1848	671	382
BTA5 (23-24)	**2,013**	919	58	80
BTA13 (56-57)	**5,601**	3712	1339	640
BTA2 (2-3)	**3,123**	2486	615	197
BTA13 (71-73)	**7,458**	2603	966	468
BTA16 (64-65)	**2,471**	1234	280	300
BTA28 (26-27)	64,522	2463	624	8
BTA21 (47-48)	**11**	571	162	20
BTA25 (22-23)	16,265	3752	1126	63
BTA13 (6-7)	**1,887**	406	182	239
BTA10 (28-29)	**7,018**	1739	321	11
BTA17 (72-73)	8,300	3556	1425	671
BTA14 (9-10)	**4,425**	1195	655	298
BTA21 (50-51)	**2,613**	804	163	20

BTA19 (26-27)	**2,018**	402	397	369
BTA14 (10-11)	12,658	3550	1063	466
BTA8 (63-64)	**424**	1215	245	125
BTA13 (6-7)	**6,865**	1790	988	455
BTA13 (53-54)	**23**	104	67	29

Miller et al., (2016) examined two regions that had previously been identified on BTA 6 (Li et al., 2004), as well as, on BTA 20 (Casas et al., 2011). Whitin these regions 3 significant SNP were identified on BTA6, and 3 SNP were identified on BTA 20. The location of the significant SNP are 6:2493836, 6:3914207, 6:2998337, 20:2626346, 20:2659865, and 20:2918960 (https://useast.ensembl.org/Bos_taurus/Info/Index). The closest SNP in the current study were 6:2492113, 6:3914242, 6:2998606, 20:2626309, 20:2654086, and 20:2921576, which is a distance of 1723, 35, 269, 37, 5779, 2616 base pairs, respectively. For all locations associated with the first region the adjusted p-value was >0.99, 0.95, 0.88, and 0.85 for single SNP analysis, 100 SNP windows, 500 SNP windows, and 1,000 SNP windows, respectively. For the second region the associated adjusted p-values were >0.99, 0.95, 0.89, and 0.85 for single SNP analysis, 100 SNP windows, 500 SNP windows, and 1,000 SNP windows, respectively. The above results suggest that similar genetic differences were not seen in this study.

5.4 Conclusions

In the current study no SNP were significant when accounting for a FDR <0.05%. Although no SNP were significant in the current study some commonality to regions previously identified in previous studies were observed. Continued research into identification of genomic regions that are predictive of an animal's susceptibility to BRD is warranted.

LITERATURE CITED

Alexandre, P. A., L. R. Porto-Neto, E. Karaman, S. A. Lehnert, and A. Reverter. 2019. Pooled genotyping strategies for the rapid construction of genomic reference populations. J. Anim. Sci. 97:4761–4769.

Baller, J. L., S. D. Kachman, L. A. Kuehn, and M. L. Spangler. 2020. Genomic prediction using pooled data in a single-step genomic best linear unbiased prediction framework. J. Anim. Sci. 98:skaa184.

Barcellos, L. F., W. Klitz, L. L. Field, R. Tobias, A. M. Bowcock, R. Wilson, M. P. Nelson, J. Nagatomi, and G. Thomson. 1997. Association mapping of disease loci, by use of a pooled DNA genomic screen. Am. J. Hum. Genet. 61:734–747.

Bell, A. M., J. M. Henshall, L. R. Porto-Neto, S. Dominik, R. McCulloch, J. Kijas, and S. A. Lehnert. 2017. Estimating the genetic merit of sires by using pooled DNA from progeny of undetermined pedigree. Genet. Sel. Evol. 49:1–7.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57:289–300.

Casas, E., M. Garcia, J. Wells, and T. Smith. 2011. Association of single nucleotide polymorphisms in the ANKRA2 and CD180 genes with bovine respiratory disease and presence of Mycobacterium avium subsp. paratuberculosis 1. Anim. Genet. 42:571–577.

Daniels, J., P. Holmans, N. Williams, D. Turic, P. McGuffin, R. Plomin, and M. J. Owen. 1998. A simple method for analyzing microsatellite allele image patterns generated from DNA pools and its application to allelic association studies. Am. J. Hum. Genet. 62:1189–1197.

Ehret, G. B. 2010. Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. Curr. Hypertens. Rep. 12:17–25.

Griffin, D. 2014. The monster we don't see: subclinical BRD in beef cattle. Anim. Health Res. Rev. 15:138–141.

Guidelines for Uniform Beef Improvement Programs [Internet]. BIF Guidelines Wiki,. 2023. Available from: http://guidelines.beefimprovement.org/index.php?title=Guidelines_for_Uniform_Beef_Improvement_Programs&oldid=2679

Keele, J., L. Kuehn, T. McDaneld, R. Tait Jr, S. Jones, T. Smith, S. Shackelford, D. King, T. Wheeler, and A. Lindholm-Perry. 2015a. Genomewide association study of lung lesions in cattle using sample pooling. J. Anim. Sci. 93:956–964.

Keele, J., L. Kuehn, T. McDaneld, R. Tait Jr, S. Jones, T. Smith, S. Shackelford, D. King, T. Wheeler, A. Lindholm-Perry, and others. 2015b. Genomewide association study of lung lesions in cattle using sample pooling. J. Anim. Sci. 93:956–964.

Keele, J., L. Kuehn, T. McDaneld, R. Tait, S. Jones, B. Keel, and W. Snelling. 2016. Genomewide association study of liver abscess in beef cattle. J. Anim. Sci. 94:490–499.

Krumbiegel, M., F. Pasutto, U. Schlötzer-Schrehardt, S. Uebe, M. Zenkel, C. Y. Mardin, N. Weisschuh, D. Paoli, E. Gramer, and C. Becker. 2011. Genome-wide association study with DNA pooling identifies variants at CNTNAP2 associated with pseudoexfoliation syndrome. Eur. J. Hum. Genet. 19:186–193.

Li, C., J. Basarab, W. M. Snelling, B. Benkel, J. Kneeland, B. Murdoch, C. Hansen, and S. S. Moore. 2004. Identification and fine mapping of quantitative trait loci for backfat on bovine chromosomes 2, 5, 6, 19, 21, and 23 in a commercial line of Bos taurus. J. Anim. Sci. 82:967–972.

McDaneld, T., L. Kuehn, M. Thomas, W. Snelling, T. Smith, E. Pollak, J. Cole, and J. Keele. 2014. Genomewide association study of reproductive efficiency in female cattle. J. Anim. Sci. 92:1945–1957.

McGuirk, S. M. 2008. Disease management of dairy calves and heifers. Vet. Clin. North Am. Food Anim. Pract. 24:139–153.

Miller, S. L., S. Mizell, R. Walker, T. Page, and M. D. Garcia. 2016. Identification of SNPs located on BTA 6 and BTA 20 significantly associated with bovine respiratory disease in crossbred cattle. Genet. Mol. Res. 15.

Neibergs, H. L., C. M. Seabury, A. J. Wojtowicz, Z. Wang, E. Scraggs, J. N. Kiser, M. Neupane, J. E. Womack, A. V. Eenennaam, and G. R. Hagevoort. 2014. Susceptibility loci revealed for bovine respiratory disease complex in pre-weaned holstein calves. BMC Genomics. 15:1–19.

Pacek, P., A. Sajantila, and A.-C. Syvänen. 1993. Determination of allele frequencies at loci with length polymorphism by quantitative analysis of DNA amplified from pooled samples. Genome Res. 2:313–317.

Pardon, B., and S. Buczinski. 2020. Bovine respiratory disease diagnosis: what progress has been made in infectious diagnosis? Vet. Clin. Food Anim. Pract. 36:425–444.

Peiris, B., J. Ralph, S. Lamont, and J. Dekkers. 2011. Predicting allele frequencies in DNA pools using high density SNP genotyping data. Anim. Genet. 42:113–116.

Quick, A. E., T. L. Ollivett, B. W. Kirkpatrick, and K. A. Weigel. 2020. Genomic analysis of bovine respiratory disease and lung consolidation in preweaned Holstein calves using clinical scoring and lung ultrasound. J. Dairy Sci. 103:1632–1641.

Reverter, A., L. Porto-Neto, M. Fortes, R. McCulloch, R. Lyons, S. Moore, D. Nicol, J. Henshall, and S. Lehnert. 2016. Genomic analyses of tropical beef cattle fertility based on genotyping pools of Brahman cows with unknown pedigree. J. Anim. Sci. 94:4096–4108.

Rivas, M. A., M. Beaudoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, and N. Burtt. 2011. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat. Genet. 43:1066–1073.

Rolf, M. M., J. F. Taylor, R. D. Schnabel, S. D. McKay, M. C. McClure, S. L. Northcutt, M. S. Kerley, and R. L. Weaber. 2010. Impact of reduced marker set estimation of genomic relationship matrices on genomic selection for feed efficiency in Angus cattle. BMC Genet. 11:1–10.

Saleem, A., S. Saleem Bhat, F. A. Omonijo, N. A Ganai, E. M. Ibeagha-Awemu, and S. Mudasir Ahmad. 2024. Immunotherapy in mastitis: state of knowledge, research gaps and way forward. Vet. Q. 44:1–23. doi:10.1080/01652176.2024.2363626.

Sham, P., J. S. Bader, I. Craig, M. O'Donovan, and M. Owen. 2002. DNA pooling: a tool for large-scale association studies. Nat. Rev. Genet. 3:862–871.

Shaw, S. H., M. M. Carrasquillo, C. Kashuk, E. G. Puffenberger, and A. Chakravarti. 1998. Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes. Genome Res. 8:111–123.

Taylor, B. A., and S. J. Phillips. 1996. Detection of obesity QTLs on mouse chromosomes 1 and 7 by selective DNA pooling. Genomics. 34:389–398.

Timsit, E., N. Dendukuri, I. Schiller, and S. Buczinski. 2016. Diagnostic accuracy of clinical illness for bovine respiratory disease (BRD) diagnosis in beef cattle placed in feedlots: A systematic literature review and hierarchical Bayesian latent-class meta-analysis. Prev. Vet. Med. 135:67–73. doi:10.1016/j.prevetmed.2016.11.006.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

Vargas Jurado, N., L. A. Kuehn, J. W. Keele, and R. M. Lewis. 2021. Accuracy of GEBV of sires based on pooled allele frequency of their progeny. G3. 11:jkab231.