DISSERTATION

DEVELOPING A STRATEGY FOR IDENTIFYING GENETICALLY IMPORTANT ANIMALS

Submitted by

Carrie S. Wilson

Department of Animal Sciences

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2023

Doctoral Committee:

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

Ronald Lewis
Esten Mason

ABSTRACT

DEVELOPING A STRATEGY FOR IDENTIFYING GENETICALLY IMPORTANT ANIMALS

Livestock researchers often need to sample animals within a breed to serve as a representative sample of the breed. Identifying the most relevant animals to include in research for genotyping, building a reference population, or inclusion in a gene bank is a complex issue. A suboptimal sampling strategy can lead to biased results, the need for additional sampling, and can be costly. When using public funds (e.g., federal grant or federal appropriations) or member fees (e.g., breed association funds), we have a responsibility to efficiently spend these investments in a wise manner, optimizing which animals are sampled before the research, genotyping, or gene banking begins.

The first objective was to develop a sampling strategy to maximize the genetic diversity captured for the sampled animals. Simulated data is ideal for this type of study as there is no limitation to the testing parameters. The primary benefit of simulation with this research was the opportunity to have known genotypes for every animal in the population. Since genotypes will almost never be available for the entire population in the real world, and identifying animals to genotype may in fact be the purpose of the sampling, pedigree-based sampling methods were chosen. Sampling methods tested included optimal contribution selection (OCS) and the genetic conservation index (GCI). The OCS selects parents based on constraining their co-ancestry rather than minimizing inbreeding. GCI seeks to maximize the number of founders in an animal's pedigree. The sampling strategy developed in Objective 1 was used to identify a subset of 100, 50, and 25 animals from each breed and the genetic diversity captured by each sampling method was assessed using both quantitative and molecular methods.

AlphaSimR was used to simulate the population for sampling. After an initial randomly mating founder population was developed, an additional 15 years of selection for phenotypic weaning weight

was simulated and resulted in a fully genotyped population with 13,662 animals per year. The simulation was designed to represent a sheep population. After the sampling strategies were applied to the simulated population, they were next applied to Suffolk sheep and Simmental beef populations for further assessment of their ability to capture genetic diversity. To assess population structure based on molecular data, the Suffolk and Simmental populations were limited to genotyped animals and their ancestors. The simulated population represented a large purebred population (n=204,930) with a moderate number of markers (n=53,901). The Suffolk population represented a small population (n=1,565) with many markers (n=606,006). Lastly, the Simmental population represented a large, admixed population (n=54,790) with a moderate number of markers (n=29,449).

For the second objective, the population structure of the full populations, comprised of genotyped animals, was assessed, and compared to the population structure of the animals from each sampling strategy. Each sampling strategy selected 100, 50, and 25 animals. The measure of success of capturing the genetic diversity of the population was a molecular-based measure defined by capturing the available alleles in the population. Other population structure measures included a comparison of a phenotypic trait, breeding values, inbreeding levels, heterozygosity, minor allele frequency (MAF) category classification, runs of homozygosity (ROH), N_e, and model-based population structure to visualize subpopulations.

While both sampling strategies were effective at capturing the available alleles in the population, OCS was more successful than GCI when comparing the same sample size. Success of capturing alleles decreased as sample size decreased from 100 to 50 to 25. Overall, OCS with a sample of 100 animals (OCS 100) was the most successful at capturing the available alleles in the population, capturing 96.5, 99.3, and 99.9 percent of the alleles for the simulated, Suffolk, and Simmental populations, respectively.

For a sampling strategy to be useful, it needs to be effective across a variety of species and breeds with a variety of breed histories and population sizes. The third objective was to compare the three populations evaluated in this research and compare the effectiveness of the sampling strategies across these populations. Population structure was compared for the three populations. Then, the effectiveness of OCS 100 was compared.

The three populations differed in population size and the amount of admixture present. The simulated population was characterized by a large number of low frequency alleles (n=5,339) that proved difficult to capture. The Suffolk population was small and consisted of 14 distinct subpopulations. The Simmental population had high levels of heterozygosity and less distinct subpopulation structure. Despite disparate populations, OCS 100 was the most robust across the three populations, consistently capturing the highest percentage of available alleles compared to the other sampling strategies.

In summary, OCS 100 was the most effective sampling strategy across three different populations. A low-cost pedigree-based sampling strategy can be used to capture the genetic diversity in a population. Researchers will need to weigh the risk of a greater loss of alleles when selecting a smaller population size. Risk could be further reduced by increasing the selected population size. Knowledge of the prevalence of low frequency alleles in the population and the value of capturing them should be considered.

ACKNOWLEDGEMENTS

This dissertation is dedicated to the sheep at the U.S. Sheep Experiment Station, who are a welcome reminder of why I began this journey in the first place. Dedicating my career to the genetic improvement of sheep is truly a dream come true and I am still in a state of awe that this is now my life. It has been a long journey to get to this point, which makes me all the more appreciative that I put myself out there to go back to graduate school and pursue my dream job. My family had no idea what I was getting into with this dream, but they have been supportive, nonetheless. My husband Mark, daughters Belle, Brynley, and Brooklyn, and dogs Hazel and Oliver have been my saving grace throughout the past 5 years of working on this degree. My girls probably don't remember a time when Mama wasn't going to school and working on her project. And my friend Beth, who I adore fiercely, and is the only other person in the world who can truly relate to all of this.

I would like to thank my committee for agreeing to take on a non-traditional Ph.D. student with a non-traditional project. I have known my advisors, Drs. Scott Speidel and Mark Enns, since I started my Master's degree more than 20 years ago. Dr. Esten Mason was so kind to join my committee and I value his expertise and insight. And finally, Dr. Ron Lewis. There is no possible way to thank you enough for joining my committee and everything you have helped with since that moment. I am aware that you are single-handedly responsible for my position at the USSES and there are no words to explain how much that has changed my life. I can only try to repay you by benefiting the sheep industry any way that I can.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND OBJECTIVES

1.1 Introduction

The proportion of the world's livestock breeds classified as being at risk of extinction increased from 15 percent to 17 percent between 2005 and 2014, with a further 58 percent of breeds classified as being of unknown risk status (FAO, 2015). Unprecedented improvements in reproductive technologies and genetic selection over the past 60 years have allowed U.S. livestock producers to dramatically increase meat production while simultaneously decreasing livestock inventory. This efficiency comes at a cost; the potential benefit of genomic selection is a doubling of genetic gain (van der Werf et al., 2014) which is also expected to double the rate of loss of within-breed genetic variation (Kristensen et al., 2015). Since genetic variation is the avenue for future genetic change, there is an urgent need to capture genetic variation before it is lost from a population.

Initiated in 1999, the mission of the USDA-ARS-National Animal Germplasm Program (NAGP) is to provide a ready supply of preserved germplasm for all major livestock species and develop a more comprehensive understanding of genetic diversity within U.S. breeds and within the NAGP collection. Capturing the existing genetic diversity within each breed is crucial before it is lost altogether. Within-breed genetic variability accounts for 83 to 93 percent of the total variation (Plante et al., 2007; Nicoloso et al., 2015) which demonstrates how critical it is to properly sample within a breed when that sample is meant to represent the population. With this research, various sampling strategies will be assessed to identify the most genetically important animals for inclusion in the gene bank, for research purposes, or for building a reference population. The sum of the identified animals will need to capture both the range of genetic diversity in the breed and the genetic merit. Sampling strategies identified for further

evaluation include optimal contribution selection (OCS) and the Genetic Conservation Index (GCI). OCS controls the relationship of the parents while maximizing the mean breeding value. The attribute that drives OCS is that the emphasis is placed on the selection of parents based on co-ancestry rather than future inbreeding (Meuwissen, 2009). The GCI maximizes the effective number of founders in the selected population. The GCI is computed using the proportion of genes of each founder in the animal's pedigree. A higher GCI is associated with maintaining higher genetic diversity within the breed (Alderson, 1992).

1.2 Objectives

Objective 1 develops a sampling strategy to maximize the genetic diversity of the sampled animals. The sampling strategy will be developed with simulated data and validated with real pedigree and breeding value data.

Objective 2 assesses the population structure for a simulated breed, a sheep breed (Suffolk), and a beef breed (Simmental). Capturing the genetic diversity of a breed in as few animals as possible is efficient and cost effective. Therefore, the sampling strategy developed in Objective 1 will be used to identify a subset of 100, 50, and 25 animals from each breed. Then, the population structure for each subset will be compared to the breed using SNP data to determine how much of the allelic genetic diversity has been captured.

Objective 3 quantifies how well the sampling strategy works for each breed and species. The sampling strategy will be quantified to determine if it is robust for breeds and species that are expected to have substantially different breed histories and subsequent population structures.

Since the sampling strategy could be validated using any livestock species, including a sheep breed provides the added value of drawing attention to a historically underserved industry and will

provide valuable information about the population structure and genotype data back to the industry. Assessing the population structure of U.S. sheep breeds has not been done at this level using a large quantity of SNP data (Ovine HD BeadChip = 606,000 SNP). This strategy is applicable for identifying animals for inclusion in research, gene banking, developing a reference population for genotyping, and for identifying redundant animals for culling. This project will produce a sampling procedure, population structure analyses for two breeds across two species, and validate with genotypic data to serve the livestock industry. This research will help efficiently secure animal genetic resources for the future and provide the genetic diversity for the livestock industry to continue to make reproductive and genetic gains.

CHAPTER II

REVIEW OF LITERATURE

2.1 Need for sampling strategy

Obtaining the correct data set is the key to conducting a proper analysis; otherwise, the results may be misleading or incomplete (Ott and Longnecker, 2015). Livestock researchers often need to sample animals within a breed to serve as a representative sample of the breed. Identifying the most relevant reference animals to include in research for genotyping, culling, or inclusion in a gene bank is a complex issue. A suboptimal sampling strategy can lead to biased results, the need for additional sampling, and can be costly. When using public funds (e.g., federal grant or federal appropriations) or member fees (e.g., breed association funds), we have a responsibility to efficiently spend these investments in a wise manner, optimizing which animals are sampled before the research, genotyping, or gene banking begins. A definitive, logical process ensures these limited funds are spent in an efficient manner. Applications for this research include deciding which animals to sample for inclusion in research, gene banks, forming a reference population for genotyping, or which group of animals to keep and cull in the case of drought or other need to reduce herd or flock size.

Much of conservation prioritization has focused on which breeds to conserve (Weitzman, 1992; Thaon D'arnoldi et al., 1998; Eding and Laval, 1999; Ruane, 1999; Karimi et al., 2016). However, within-breed genetic variability accounts for more diversity than between breeds. Researchers have reported within-breed genetic variability to account for 83 to 93 percent of the total variation (Table 2.1). This demonstrates how critical it is to properly sample within a breed when that sample is meant to represent the entire population.

Table 2.1 Summary of within-breed genetic variation across studies

Species	Population	Analytical Tool	Within-Breed Genetic Variation (%)	Source
Equine	24 Canadian populations	38 microsatellites	86.7	(Prystupa et al., 2012)
Equine	12 German breeds	30 microsatellites	88.4	(Aberle et al., 2004)
Equine	6 breeds	50 microsatellites	90.0	(Glowatzki-Mullis et al., 2006)
Equine	21 breeds	12 microsatellites	83.0	(Plante et al., 2007)
Caprine	14 Italian breeds	50K SNP chip	92.5	(Nicoloso et al., 2015)
Ovine	22 breeds	11 microsatellites	87.2	(Paiva et al., 2011)
Ovine	29 European breeds	23 microsatellites	87.0	(Handley et al., 2007)
Ovine	5 U.S. breeds	50K SNP chip	88.8	(Zhang et al., 2013)

Maximizing the genetic variation within a breed allows for adaptation to environmental change, whether that change is geographical, climate, management, or production environment. Part of the balance of maintaining genetic diversity while selecting for animals adapted to an environment is to minimize inbreeding, as inbreeding makes a population more homozygous, thereby reducing genetic variance within a population (Meuwissen, 2009). Selection for increased genetic merit also results in a loss of genetic diversity as favorable alleles are increased and unfavorable alleles are decreased or are eliminated from the population (Eynard et al., 2018b).

2.2 Genetic diversity

Genetic diversity exists as expressed (or adaptive) genetic diversity, which is at the coding regions of the genes, and neutral genetic diversity, which is at the non-coding regions (Eding and Laval, 1999; Windig and Engelsma, 2010). Genetic diversity is also measured in terms of allelic diversity, or allelic richness. The ability of a population to respond to selection or environmental pressure depends on having different alleles. Genetic diversity is maximized when all alleles at a locus have the same frequency, thereby minimizing the loss of alleles to genetic drift (Fernández et al., 2004). Because SNP

markers are biallelic, Fernández et al. (2016) recommended defining genomic allelic diversity based on haplotypes to maximize polymorphism. A common measure of genetic diversity is heterozygosity, which does not account for the number of alleles at a locus. The loss of a rare allele will generally not affect heterozygosity, but does reduce allelic richness (Falconer and Mackay, 1996; Greenbaum et al., 2014).

When determining the most genetically important animals to represent a breed, there is a conflict between maintaining the allelic frequencies that already exist in the breed or maximizing the allelic diversity, which will favor all alleles at the locus having the same frequency. The choice is between representing the existing population as it is or minimizing the chance of losing a rare allele (Fernández et al., 2004; Saura et al., 2008). Windig and Engelsma (2010) provide an example of selecting for genetic diversity resulting in an increase of the previously rare blond phenotype in a cattle population, which was considered a negative by the owners. For conservation programs, maintaining the allele frequencies as they exist in the population is preferred (Lacy, 2000).

2.3 Gene banking

With the world's population expected to increase to more than 9 billion people by 2050 and continued increases in per capita income, demand for livestock products is expected to dramatically increase (Thornton, 2010). By 2050, the global cattle population may increase from 1.5 billion to 2.6 billion, and the global goat and sheep population from 1.7 billion to 2.7 billion (Rosegrant et al., 2009). As mean climate change and increased climate variability are expected to occur, it is difficult to predict how a population will adapt to an altered environment. Adaptation is limited by the available genetic variation; hence ensuring genetic resources are available for the livestock industry to meet the world's food needs is of utmost importance (Thornton, 2010; Templeton, 2021).

As explained by Engelsma et al. (2011), since all animals within a breed cannot be included in a gene bank, a method to identify those animals that captures as much genetic diversity as possible to conserve is needed. An excellent description of gene banking sampling was provided by (Eynard et al., 2018b), “Sampling for gene bank collections should focus on collecting old individuals as representative as it can be of the former population, as well as individuals carrying unique diversity and to collect current individuals of potential interest for the future”. Gene banks are a major source of conservation of livestock genetic resources. The primary purpose of a gene bank is to decrease the risk for *in situ* live populations, which is why gene banks are logically a public entity because of the low present value associated with decreasing risk and capturing genetic diversity for the long-term (Oldenbroek, 1999; Weigel, 2001). In addition to the initial purpose of publicly owned gene banks, which is for reestablishment of breeds in case of a disease outbreak or other worst-case scenarios, the genetic resources can also be used for reintroduction of lost genetics, broadening the genetic base, research, and genetic studies. Additionally, in the U.S., live populations at land grant universities and other public entities have greatly been reduced, increasing the importance of having genetic resources in reserve for future research (Blackburn, 2018).

On the surface, there appears to be a conflict between conservation activities, which aim to maximize genetic diversity, and production agriculture, which supports intense selection intensity for traits of economic relevance, thus rewarding phenotypic uniformity; this reduction in variation inherently leads to reduced genetic diversity (Kristensen et al., 2015). As the use of reproductive technologies (AI, ET, sexed semen) increased starting in the 1960’s, much higher selection intensity for superior males and females has been practiced, leading to an increase in inbreeding and reduced effective population size. More recent advances in genomic selection in the past decade further increases selection intensity and decreases generation interval, further reducing effective population

size. While the potential benefit of genomic selection is a doubling of genetic gain (van der Werf et al., 2014), it may also double the rate of loss of within-breed genetic variation (Kristensen et al., 2015).

Despite these apparent conflicting interests between conservation activities of gene banks and production activities of the livestock industry, the opposite is true. Gene banks rely on contributions of genetics from the livestock industry to build a complete collection. If gene banks have sufficient genetic variation captured for each breed, industry will have cryopreserved genetics to draw from if additional genetic variation is needed in the future in the case of changing market demands, climate change, or deleterious mutations. The gene bank can then be replenished from the resulting offspring. This allows industry to maximize profit while also maintaining genetic variation. The livestock industry is primarily dominated by a few high-input/high-output breeds of high genetic merit. Maintaining the genetic diversity of these breeds is important to provide the opportunity to counter some of the negative effects in the future (e.g., selection for milk production in Holstein cattle has resulted in problems with fertility). Additionally, conservation of low-input/low-output breeds may benefit the high-input/high-output breeds as well (e.g., introgression of polled genes) (Windig and Engelsma, 2010).

2.4 Core collection and N_e

Blackburn (2009) described the animals selected for the Core Collection of the U.S. gene bank as taking into account genetic diversity, capturing rare alleles, and obtaining sufficient quantities for breed regeneration. For a gene bank, a Core Collection is the portion of the collection that is reserved for critical needs, including breed regeneration, reintroduction of genetic variation to overcome a bottleneck, or elimination of a deleterious mutation. Once the Core Collection has been established, it is meant to be dynamic and updated with current genetics to represent each breed. The Core Collection should contain 150% of the germplasm needed to reconstitute each breed (FAO, 2012).

If the Core Collection is needed for breed regeneration, the re-creation of a breed serves as the equivalent of a founder event, where a new population is established by only a few founders. General characteristics of a founder event include the loss of rare alleles, leading to decreased allelic diversity, reduced genetic variation, different allele frequencies than in the original population, and decreased heterozygosity (Halliburton and Halliburton, 2004; Campbell and Reece, 2008). Ideal sampling of animals for the Core Collection will minimize these negative attributes. Most populations behave as though they are smaller than their census size, a concept referred to as the effective population size (N_e). The effective population size is influenced by the breeding structure, sex ratio, mean and variance of number of offspring, variation in population size over time, and any other factor that causes the breed to deviate from the idealized population (Wright, 1931; Halliburton and Halliburton, 2004).

In livestock populations, Meuwissen and Woolliams (1994) are often cited as recommending a range of effective population size of 31 to 250. The United Nations Food and Agriculture Organization (FAO) set guidelines for breed conservation stating that achieving a minimum effective population size of 50 animals per generation should be the first priority for managers. This effective size of 50 will result in an inbreeding rate of one percent per generation (FAO, 1998). Given a consensus that 50-100 is a viable effective population size, Meuwissen (2009) recommends 100 to err on the side of caution. NAGP has set a minimum collection goal of 50 males per breed (Blackburn, 2018).

At present, gene banks throughout the world are primarily comprised of cryopreserved semen, as it is routinely collected and cryopreserved across most livestock species. Embryos pose a challenge because the mating decision has already been determined, reducing the flexibility of its future use. Additionally, embryos are more difficult and expensive to obtain when compared to semen (Blackburn, 2018). In the future, as cryopreservation of ooctyes becomes routine in the livestock industry, there will be a paradigm shift in how Core Collections for gene banks are constructed. Backcrossing over multiple generations will be replaced by reconstitution in a single generation. For now, emphasis in gene banks is

placed on selecting males for breed re-establishment via backcrossing. The Core Collection should include animals that maximize genetic variation; for semen only storage, this variation will be maximized by minimizing the average co-ancestry among males (Toro and Mäki-Tanila, 1999).

2.5 Sampling for research

It is rare for all animals of interest to be included in a research study; therefore, some form of sampling strategy must be implemented. For the U.S. Meat Animal Research Center 2000 Bull Project, the research included understanding breed composition and allele frequencies of 16 cattle breeds. Selection of bulls for inclusion in the research study was determined by each breed association and they were not provided any limitations on relatedness, offspring counts, or EPD accuracy (Kuehn et al., 2011). Because it is unlikely each breed association applied the same selection criteria, the results may be biased. Other animal genetics studies selected influential bulls from three dairy cattle breeds (Hozé et al., 2014), minimally related animals across farms from goat breeds (Nicoloso et al., 2015), randomly sampled animals within horse breeds (Prystupa et al., 2012), and based on their marginal contribution to the population for cattle breeds (Hozé et al., 2013).

2.6 Sampling for a reference population

Advances in the use of genomic data require a reference population that can be used for genotype imputation (Hozé et al., 2013; van der Werf et al., 2014; Neuditschko et al., 2017) and to maximize the accuracy for genomic selection (Hozé et al., 2013; Pszczola et al., 2014). Developing a reference population consists of determining how large it should be and which animals to include (van der Werf et al., 2014). Various strategies have been described for selecting animals for a reference

population. Neuditschko et al. (2017) criticized the standard method of selecting animals based on key ancestors through the use of pedigree or genomic relationships because of a disregard for population substructure and the most influential progeny; they instead proposed using the Eigenvalue Decomposition of a genomic relationship matrix to select key contributors. This strategy inherently requires genomic data to be available for individuals in order to be considered. For genomic selection, Pszczola et al. (2014) described the need to update the reference population over time to maintain high accuracy of genomic breeding values. Rather than randomly selecting animals for updating the reference population, they suggest that a reference population that is minimally related to each other yet maximally related to the target population will lead to the highest accuracy. For imputation purposes, Hozé (2013) concluded the accuracy of imputation was related to the size of the reference population and the relationship between the reference population and target populations. van der Werf et al. (2014) further suggested a balance between genetic merit and diversity, where progeny of young sires of high genetic merit are included given that they are lowly related to each other.

2.7 Sampling for culling

Just as genetically unique animals need to be identified for conservation, they also need to be identified in the case of population reduction, or culling. Genetically unique animals need to be identified for retention and other individuals can be selectively culled to reduce overall inbreeding and increase genetic diversity. Windig and Engelsma (2010) provided an example from the Netherlands where a semi-wild cattle population was maintained by selectively removing animals with redundant genetics (e.g., one of a pair of full sibs). The resulting inbreeding was less than was expected based on the effective population size. The use of genomics further enhances the ability to identify animals to

maintain a high level of genetic diversity, even with reduced population size (Windig and Engelsma, 2010).

2.8 Population genetics related to identifying genetically unique animals

Population structure is defined as the amount and distribution of genetic variation within and among both local populations and the entire species. The primary determinant of population structure is the balance between genetic drift and gene flow. Genetic variation provides the raw material that makes evolutionary change possible (Templeton, 2021).

Population genetics theory is driven by four factors that cause populations to change over time; those factors are random drift, selection, mutation, and migration (Eding and Laval, 1999). As described by Meuwissen (1997), selection based on performance traits will lead to increased inbreeding and increase the likelihood of deleterious genes drifting to high frequency for traits that are not under selection. In particular, reproductive traits suffer from inbreeding depression and many of the positive reproductive genes will be lost before the breeding objective is changed to correct the reduced reproductive performance. Ruane (1999) and Eding and Laval (1999) state that genetic drift plays a greater role in domesticated breeds than mutation since domestication was a relatively recent event and mutation takes many generations to have a measurable impact. Genetic variability is primarily explained by random drift (Eding and Laval, 1999). Halliburton (2004) describes the key results of genetic drift as 1) the direction is unpredictable; 2) the magnitude of change from generation to generation is larger with small populations; 3) the long-term effect is to reduce genetic variation; and 4) the populations will diverge from each other.

According to Eding and Laval (1999), capturing the genetic diversity within a breed should not focus on specific traits, but on overall diversity. This removes the emphasis from capturing every allele

per locus to capturing the diversity of genotypes instead. Their claim is that for polygenic traits, diversity is measured in terms of genotypes and not alleles. If subpopulations within a breed are isolated, e.g., geographical isolation without the use of A.I., even though selection in all populations may be in the same direction, random drift of neutral genes may still show divergence between the subpopulations. In this case, SNP are useful for measuring the genetic differentiation between subpopulations, more so than performance measures. Gene flow between subpopulations, i.e., migration, reduces the genetic differences between populations. Of the evolutionary forces in population genetics, drift and selection lead to a decrease in genetic variation while mutation and migration lead to an increase. Understanding the role of each of these forces in a population is important to assessing the genetic diversity of the population through its history, its state currently, and its likely future direction.

2.9 Features of sampled populations - real and simulated data

Genetic studies, and genomic research in particular, can be conducted using simulated or real data, or a combination of the two. Both approaches can be used for comparing analytical methods, comparing genomic breeding programs, evaluating genomic selection from the short- to long-term (Daetwyler et al., 2013), and modelling assumptions about population structures that impact genetic diversity (Hoggart et al., 2007). Real data is complex, while simulated data can be structured to allow for focused evaluation of a specific aspect of the population. Daetwyler et al. (2013) explain that exploring the source of variability, such as genetic drift, is better suited for simulated data than is real data. Additional limitations of real data are they provide a single, non-random sample of a population and a fixed sample size (Daetwyler et al., 2013).

Simulated data is a low-cost option to test a variety of hypotheses, including evaluating long-term impacts of a choice much more quickly than with real data (Daetwyler et al., 2013). Although

simulation is a “simplification of reality”, this can be beneficial by removing the noise of real data and highlighting important effects, particularly demonstrating these effects in graphical form to enhance understanding for livestock breeders (Leroy and Rognon, 2012; Windig and Oldenbroek, 2015). Windig and Oldenbroek (2015) discussed simulation uncovering unexpected results, specifically regarding ways to manage inbreeding within breeds. Windig and Doekes (2018) used simulation to investigate the effects of outcrossing to reduce inbreeding in a dog breed. They determined the impact of donor alleles being introduced to the recipient breed as a result of the outcross. These techniques can be utilized to identify outside alleles introduced into a breed as a result of a grading up program, as with Katahdin sheep in the U.S. Limitations of simulated data identified by researchers include the assumption that the population structure is consistent through time, that the population starts with unrelated, non-inbred animals, and that there is no selection against deleterious alleles. Other weaknesses identified include that a simulation may not consider the existence of subpopulations or lines, assumes an equal opportunity for mating between males and females, and all sires produce the same number of offspring (Windig et al., 2004; Windig and Oldenbroek, 2015).

Simulation can be applied to evaluate the effectiveness of conservation strategies. Saura et al. (2008) compared methods with the objective of maintaining the allele frequencies at each locus to match the original population for a conservation program. Simulated markers included both neutral loci and QTL. Various scenarios included minimizing either the molecular or the pedigree co-ancestry at various levels of linkage, varying selection intensity, and including all pedigree relationships or only those from the current population, assuming previous relationships were unknown. In this study, recommendations for *ex-situ* conservation were possible because simulation allowed many scenarios to be tested that would not be possible with real data.

In order to thoroughly evaluate research strategies, researchers have incorporated both real and simulated data. Neuditschko et al. (2017) evaluated a method for identifying key contributors to a

population using a simulated dataset to represent a highly structured population and two real datasets to represent complex population structures. Sánchez-Molano et al. (2016) developed a procedure to identify ancestral haplotypes using simulated data and then applied it to real data from several cattle breeds. Other researchers used real pedigree data to develop a simulation for 'what if' mating strategies to see how the current pedigree compared to the alternative simulated strategies for controlling genetic disorders and managing genetic diversity (Leroy and Rognon, 2012). Windig et al. (2004) used existing data from sheep breeds to replicate the population structure and then simulated various approaches for selecting against scrapie sensitive alleles determining the impact on inbreeding levels and allele frequencies. Windig and Oldenbroek (2015) used existing pedigrees for Golden Retriever dogs to simulate a population and optimize breeding strategies to manage inbreeding levels.

Because of the complexity of simulating populations, animals, and their underlying genomes, there is no single model that can be applied for all research. Daetwyler et al. (2013) proposed ways to validate simulated data and standardize reporting of results so studies can be more readily compared. They recommend comparing population parameters for the simulated data, such as linkage disequilibrium and heterozygosity, to expectations reported in the literature. Additionally, the features of the genome and traits, the assumptions made, and the validation design should be described. To standardize reporting, population parameters, such as estimates of effective population size, genome length, number of phenotypes, family structure, and sample structure should be delineated.

2.10 Data simulation

AlphaSimR software was developed to allow running stochastic simulations of whole breeding programs over multiple generations to model long-term genetic gain. AlphaSimR uses both the coalescent and gene drop methods. The coalescent is used for backwards-in-time simulations, where

whole chromosome haplotypes are generated for founders that match a defined genetic model with linkage disequilibrium and allele frequencies determined by the user. The gene drop method is then used for forwards-in-time simulations to create new haplotypes from the original founders (Hickey and Gorjanc, 2012; Gaynor et al., 2021).

Founder haplotypes are simulated using the Markovian Coalescent Simulator (MaCS) (Chen et al., 2009), which simulates haplotypes from moving from one end of the sequence to the other and models dependencies between recombination events that are close together but handles recombination events that are farther apart as independent. New genotypes are created by simulating meiosis and genetic recombination using the gamma model, which accommodates crossover interference (McPeek and Speed, 1995). Traits are classified according to the biological effects being modeled and include additive, dominance, epistatic, and genotype-by-environment (GxE). Biological effects can be combined for a trait, e.g., additive+dominant+epistatic. Modeling of dominance allows for directional dominance as well as degree of dominance ranging from partial dominance to overdominance. Epistasis is modeled as additive-by-additive epistasis effects between discrete pairs of loci. GxE effects are modeled as an additive effect whose value is a function of a single environmental covariate (Gaynor et al., 2021).

Selection can be modeled using various criteria, including phenotypes, genetic values, breeding values, estimated breeding values, and can be on one trait or an index of multiple traits. Selection can occur within families, between families, and over the entire population. A wide range of mating schemes can also be applied (Gaynor et al., 2021).

2.11 Optimal Contribution Selection (OCS)

According to Doekes et al. (2018), OCS is the "gold standard" for maximizing genetic gain (mean EBV) while restricting mean relatedness. As described by Meuwissen (2009), OCS controls the relationship of the parents while maximizing the mean breeding value. Computationally, OCS:

maximizes: $G = c'EBV$

while restricting: $\bar{A}_p = c'Ac$

and restricting: $c'Q = 0.5$,

where G is the average genetic level of the parents weighted by their number of offspring; EBV is a vector of estimated breeding values of the parents; c is a vector of genetic contributions of the parents (fraction of offspring each parent obtains, scaled such that it sums to 0.5 for males and to 0.5 for females; c_i=0 implies that animal i is not selected); $\bar{A}_p$ is the desired value of the average genetic relationships of the parents; A=matrix of genetic relationships of the selection candidates; and Q is a design matrix indicating the gender of the candidates (first column is 1 for male, 0 for females; second column is 1 for female, 0 for males). The $c'Q = 0.5$ restriction mathematically ensures that the genetic contributions of the males (females) add to 0.5 (in real life this is ensured by nature). The contributions of the selection candidates c are optimized by an optimization algorithm, e.g., by the Lagrangian multiplier method (Meuwissen, 2009).

Meuwissen (1997) used simulation to compare OCS to Best Linear Unbiased Prediction (BLUP) selection and demonstrated that OCS resulted in higher genetic gain at every level of inbreeding than BLUP selection (range of 21 to 60% greater selection response). OCS is far superior to BLUP selections when low rates of inbreeding are required. The attribute that drives OCS is that the emphasis is placed on the selection of parents based on co-ancestry rather than future inbreeding (Meuwissen, 1997).

Proper selection of parents is more important than mating decisions (Meuwissen, 2009). Nonrandom mating will not prevent inbreeding in the long term because the relationships between selected parents will eventually be converted to inbreeding (Meuwissen, 1997). Avoiding the mating of close relatives delays the accumulation of inbreeding, yet the rate stays the same: if inbreeding is plotted over time, the curve is shifted downwards but the slope is unchanged (Meuwissen, 2009).

To demonstrate the importance of controlling the average relationship of the selected parents, Meuwissen (2009) provides an example where one male is unrelated to all females in a population, so he is chosen to mate to all females to produce non-inbred offspring. This strategy would be chosen if the objective was maximum avoidance of inbreeding in the current generation. However, the resulting offspring are all half sibs and the generations after that would have high levels of inbreeding. Using the OCS strategy, the male weighted by his number of offspring would be a heavy contributor to the average relationship of the parents, so selecting this single male would not be acceptable. In looking at the relationship of the selected parents in the current generation, the OCS strategy protects future generations from inbreeding.

In the short-term, excessive use of elite animals to rapidly improve performance will narrow the gene pool and reduce the available genetic diversity for future generations. Ultimately, this would limit the long-term genetic potential of the population. OCS helps to avoid this by finding a balance between short and long-term genetic gain (Gorjanc and Hickey, 2018).

2.12 Studies with OCS

Gourdine et al. (2012b) used a simulation study to test selection strategies incorporating selecting for a single EBV in a small pig population. They compared random selection, truncation selection for EBV, and OCS. While genetic gain was slightly higher for truncation selection than OCS,

inbreeding rate was much higher. The researchers concluded OCS could be used to improve genetic merit in a small population while also maintaining an acceptable level of inbreeding. Other researchers used simulation to compare OCS strategies to those currently used fish breeding programs; they concluded OCS should be implemented (Skaarud et al., 2011). Windig and Oldenbroek (2015) used simulation of a Dutch dog breed and found OCS to be the most efficient breeding strategy to manage inbreeding for this population. However, they acknowledged that OCS is difficult to implement in practice when many individual breeders are involved. Avendaño et al. (2003) applied OCS to real livestock breeds in the UK. For both the beef and sheep breed, OCS resulted in greater genetic gain at a defined inbreeding rate than truncation BLUP selection. Like concerns expressed by Windig and Oldenbroek (2015), they recognized the additional genetic gain available through the use of selection tools, but realization of the gain will require coordination among breeders. Howard et al. (2014) applied OCS with a commercial pig population and concluded OCS improved genetic gain as predicted and agreed with the theory and simulation results presented by others.

Selection based on QTL information causes diversity to vary over the genome. Simulations with OCS for a population selected using QTL information showed that the inbreeding rate in the region near the QTL was higher than the pedigree based inbreeding rate. The researchers suggested that selection based on overall average relatedness could result in a loss of diversity in and around QTLs under selection (Windig and Engelsma, 2010).

2.13 Genetic conservation index (GCI)

According to Alderson (1992), capturing the genetic diversity of a population is best accomplished by retaining all the alleles in the original founder population. Moreover, this is achieved by an animal receiving equal contributions from all the founder ancestors. Computationally, the GCI for an

animal is the effective number of founders in its pedigree from $1/\Sigma P_i^2$, where P_i is the proportion of genes of founder animal *i* in the pedigree (Alderson, 1992). Animals can be ranked based on GCI, with a higher GCI value representing a higher level of genetic diversity. The maximum GCI value is equal to the number of founder ancestors in the breed. Because GCI is based on the founder ancestors for each breed, the value cannot be compared across breeds. Because GCI is computed from pedigree data, incomplete to inaccurate pedigree recording will influence the results.

2.14 Studies with GCI

Alderson used GCI to make selection and mating decisions for the rare White Park cattle breed. Mating decisions are made by computing the GCI for the next generation based on the selected parents (Alderson, 1992). For the Pantaneiro Horse in Brazil, GCI has increased over time with the female trend lagging behind males until 2005 (McManus et al., 2013). In the Italian Maremmano horse, GCI trend also increased over time; however, GCI has decreased among males in recent years while continuing to increase among elite females (Giontella et al., 2019).

Increased emphasis on genetic conservation among breeders and research institutions of the Morada Nova sheep breed in Brazil has led to an increase in GCI over time as well as an increase in the range of GCI values. However, incomplete pedigree data for this breed may influence the results (McManus et al., 2019). Boer goats in Brazil had an increasing GCI over time with a large range of GCI values among animals (Menezes et al., 2015). GCI was low for the Spanish Murciano-Granadina goat, suggesting the founder population is not being well represented in the current population. The authors found few influential animals representing the genetic diversity of the population and suggested this could cause a loss of genetic diversity over time if changes in sire use policy are not implemented (Oliveira et al., 2016). The GCI for this breed may be complicated by the merging of the herdbooks of the

Granadina and the Murciana breeds in the 1970s and the ongoing controversy of whether they are a single breed or two separate breeds. This exemplifies the importance of pedigree records and breed history on computations such as GCI that are based on pedigree data alone (Martinez et al., 2010).

2.15 Pedigree- vs. marker-based diversity

Marker-based diversity and pedigree-based diversity are measured on different scales, where pedigree-based diversity is a measure of the probability of an allele being identical by descent (IBD) from the founder population, which is considered to be unrelated, and marker-based diversity, which does not assume a founder population and all markers that are identical by state (IBS) are assumed to be IBD. Marker-based diversity considers Mendelian sampling and allows for evaluation in more detail at specific genome regions, neither of which are possible with pedigree-based diversity. The reported correlation between pedigree and marker-based diversity ranged from 0.39 in humans to 0.92 in Iberian pigs; Mendelian sampling ensures the correlation will never reach 1. Specific genome regions where direct selection for QTL has led to high inbreeding in specific genome regions is of particular concern for loss of genetic diversity. If the pedigree information is missing or unreliable, marker-based diversity would be particularly useful in conserving genetic diversity (Engelsma et al., 2011).

2.16 The U.S. sheep industry

The U.S. lamb crop reached a record number of 32.6 million in 1941 and has been on the decline since that time, reaching a low of 3.2 million in 2020 (Figure 2.1) (NASS, 2021). The lambing rate was a low of 92% in 1978 and peaked at 113% in 2004-2005. The most recent reported lambing rate (2018) was 107% (Figure 2.2) (NASS, 2021). The top ten sheep producing states in 2021 are shown in Table

2.2(NASS, 2021). While most sheep are owned by operations with fewer than 100 ewes, most breeding ewes are part of large flocks (Figures 2.3 and 2.4, respectively). Public funding for sheep research declined by 30 percent between 2002 and 2014 (Miller et al., 2016). The U.S. sheep industry continues to be productive, contributing $2.02 billion in sheep-related products that generated a total impact of $5.8 billion to the economy in 2016 (Stepanek Shiflett, 2017). The top ten breeds for number of registrations in 2020 are shown in Table 3 (Banner Sheep Magazine, 2021).

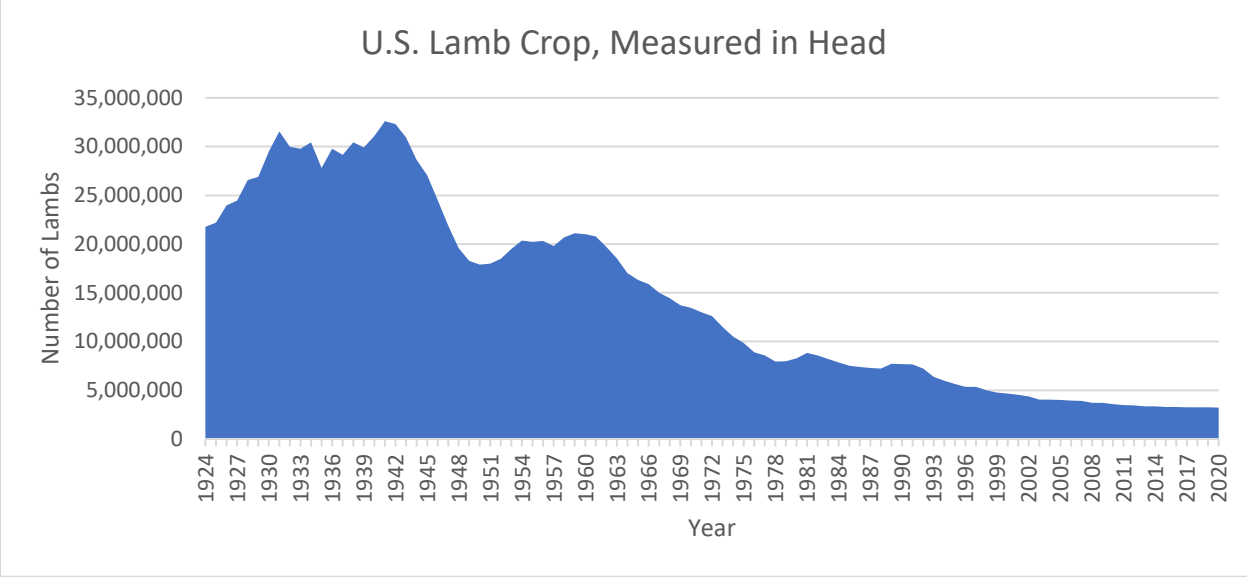

Figure 2.1 U.S. lamb crop by year (Data from NASS, accessed 12/1/2021)

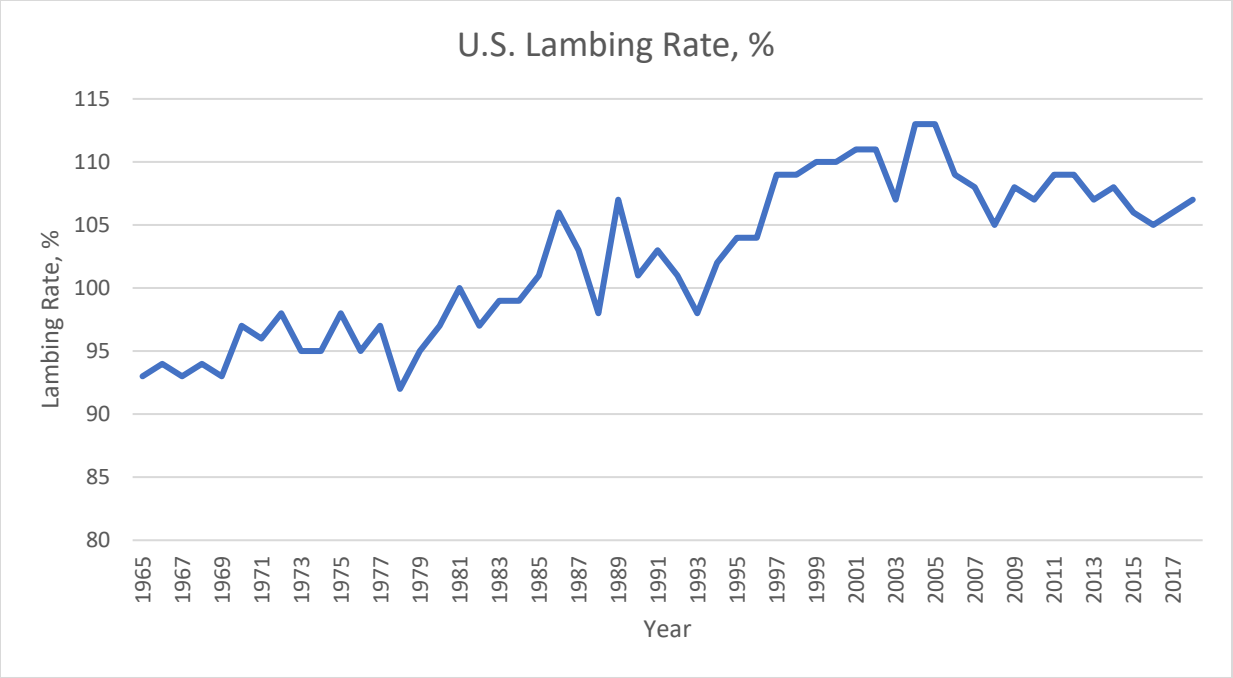

Figure 2.2 U.S. lambing rate by year (Data from NASS, accessed 12/1/2021)

Table 2.2 Top ten states for sheep inventory including lambs, 2021 (Data from NASS, accessed 12/1/2021)

State	**Sheep Inventory including Lambs**
Texas	730,000
California	555,000
Colorado	445,000
Wyoming	340,000
Utah	285,000
South Dakota	245,000
Idaho	230,000
Montana	200,000
Iowa	160,000
Oregon	155,000

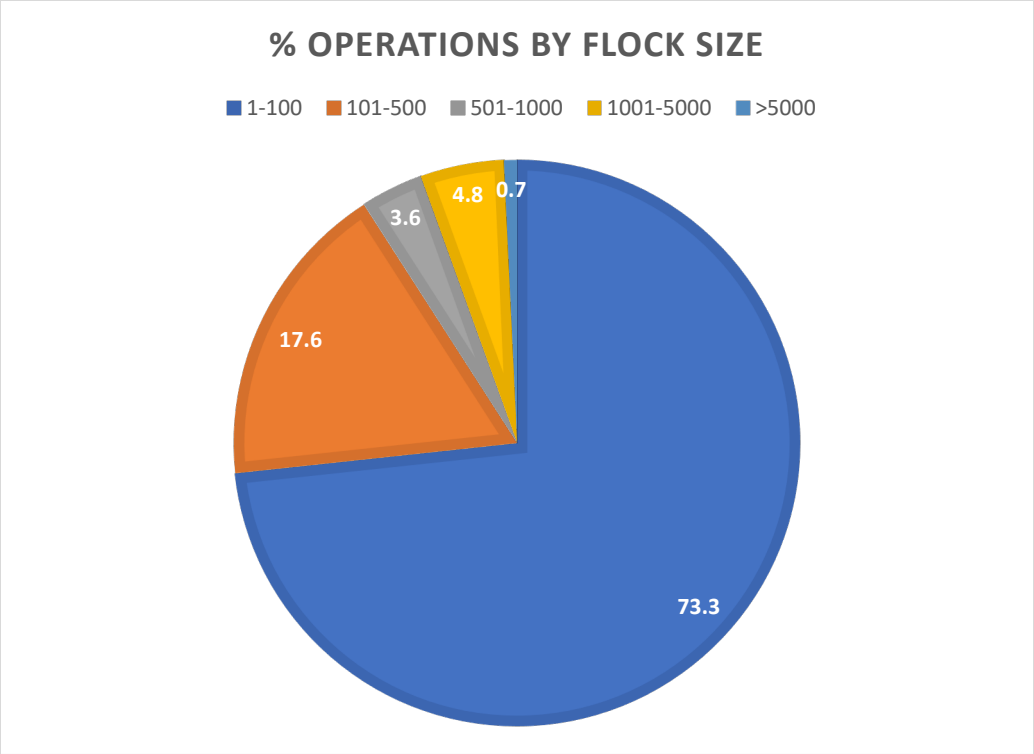

Figure 2.3 Percent sheep operations by flock size (adapted from Miller et al., 2016)

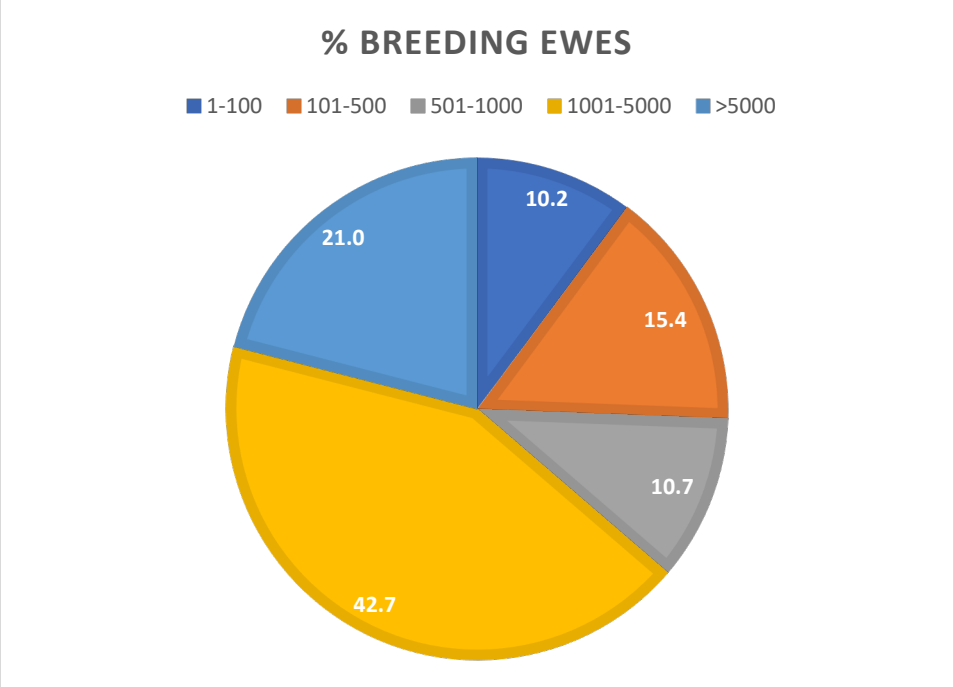

Figure 2.4 Percent breeding ewes by flock size (adapted from Miller et al., 2016)

Table 2.3 U.S. sheep breeds with highest number of registrations in 2020

Breed	Number of Registrations
Dorper	10,581
Katahdin	10,186
Hampshire	6,001
Dorset	5,625
Southdown	5,020
Suffolk	4,561
Shropshire	1,823
Shetland	1,743
Rambouillet	1,628
Babydoll Southdown	1,569

2.17 Suffolk sheep

The Suffolk breed was developed in England by crossing Southdown rams with Norfolk Horned ewes. They were imported into the U.S. in 1888. Rams weigh 115 to 160 kg and ewes weigh 80 to 115 kg. Suffolk are the primary terminal sire breed for the U.S. sheep industry (USSA, 2021). In a study comparing terminal sire breeds, Suffolk-sired lambs were heavier at each weigh date, grew faster, and reached body weight end points with fewer days on feed than other breeds (Kirschten et al., 2013). Suffolk are a medium wool breed with a fiber diameter of 25.5 to 33 microns and a staple length of 5 to 6.75 cm. An average ewe fleece weight is 2.25 to 3.6 kg. To meet the breed standard, Suffolk should have a distinctly black head, ears, and legs with a white fleece (USSA, 2021). In 2020, Suffolk were the 6th most registered sheep breed in the U.S. with 4,561 registrations (Banner Sheep Magazine, 2021).

Blackburn et al. (2011) used an FAO/ISAG panel of 31 microsatellites to assess the genetic diversity among and within 28 U.S. sheep breeds. The researchers found U.S. sheep breeds to have large amounts of observed heterozygosity and breeds were clustered more by production type than by geographical origin. Suffolk averaged 6.07 alleles per locus, which was higher than the average of 5.86 across all breeds. Inbreeding (F_{IS}), H_O, and H_E were reported as 0.139, 0.578, and 0.655, respectively. Of

the 28 breeds, only 8 had a higher F_{IS} than Suffolk. For H_O and H_E, Suffolk was intermediate. The authors concluded the U.S. has a wide range of genetic diversity in its sheep breeds based on evidence provided by number of alleles, heterozygosity, and genetic distance.

Zhang et al. (2013) used the Illumina OvineSNP50 BeadChip to compare within and between genetic diversity for five U.S. sheep breeds. The genetic differentiation using Wright's F_{ST} between Suffolk and Rambouillet was 0.16. A neighbor-joining tree also showed Suffolk distinctly separated from Rambouillet and Rambouillet-related breeds. They reported Suffolk to have the lowest gene diversity, heterozygosity, and polymorphism information content when compared to the other breeds.

The National Sheep Improvement Program (NSIP) computes EBV for the following traits for Suffolk: Birth Weight, Weaning Weight, Maternal Weaning Weight, Post Weaning Weight, Post Weaning Fat Depth, Post Weaning Eye Muscle Depth, Number of Lambs Born, Number of Lambs Weaned, Carcass Plus Index, and SRC$ Index. The Carcass Plus Index places emphasis on increasing post weaning weight and post weaning eye muscle depth while decreasing post weaning fat depth. The SRC$ Index is the Self Replacing Carcass Index and is designed for terminal breeds that keep replacements. Emphasis is placed on birth weight, weaning weight, maternal weaning weight, post weaning weight, post weaning fat, post weaning eye muscle depth, and number of lambs weaned (NSIP, 2021). Since 2004, Suffolk breeders have increased post weaning weight by 2 kg and 120 day weight by 3 kg (Wyoming Livestock Roundup, 2021).

2.18 The U.S. beef industry

Beef production in the U.S. has increased over time while beef cattle inventory has decreased due to increased reproductive efficiency and genetic selection (Figure 2.5) (FAS, 2021). The top ten beef producing states in 2021 are shown in Table 2.4 (NASS, 2021).

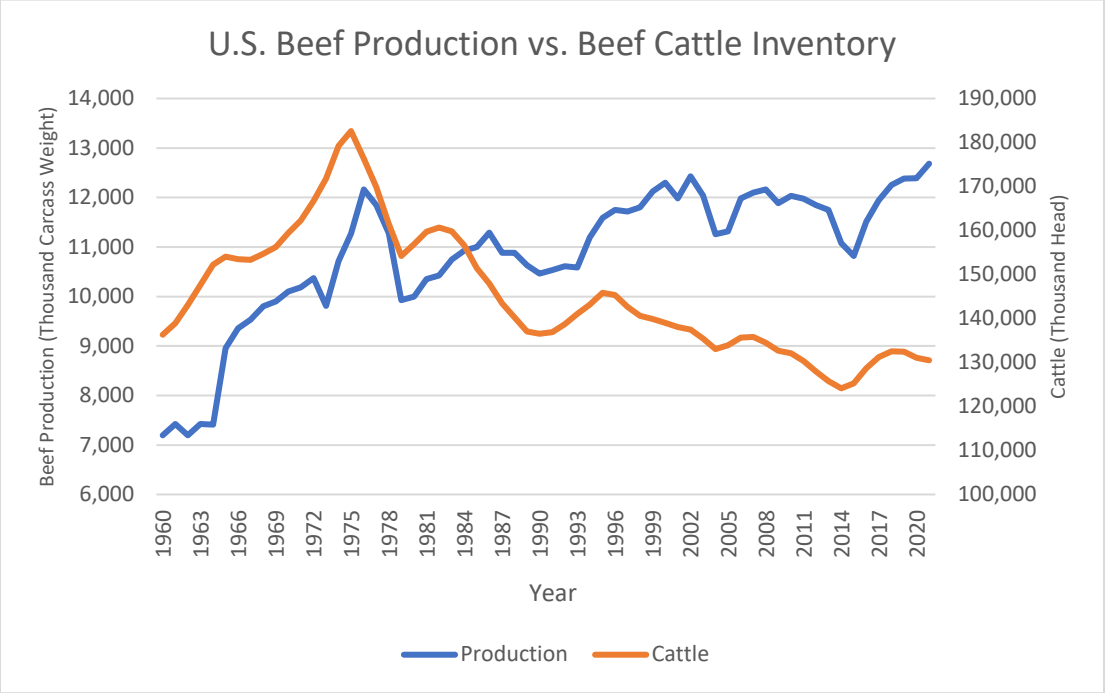

Figure 2.5 U.S. beef production and beef cattle inventory by year (Data from FAS, accessed 12/15/2021)

Table 2.4 Top ten states for cattle inventory including calves (beef and dairy), 2021 (Data from NASS, accessed 12/14/2021)

State	Cattle Inventory including Calves
Texas	13,100,000
Nebraska	6,850,000
Kansas	6,500,000
Oklahoma	5,300,000
California	5,150,000
Missouri	4,300,000
South Dakota	4,000,000
Iowa	3,650,000
Wisconsin	3,450,000
Colorado	2,650,000

2.19 Simmental cattle

Simmental cattle originated in Switzerland and are now distributed worldwide. There are 40 to 60 million Simmental throughout the world, second only to Brahman in total numbers. They were thought to have been imported into the U.S. in the late 1800's but did not have a lasting presence until the importation of the bull Parisien from France into Canada in 1967. Semen was imported into the U.S. in 1967 and the American Simmental Association (ASA) was started in 1968. The breed was originally multi-purpose, and was used for meat, milk, and draft. Currently, Simmental are dual-purpose throughout the world with the exception of the U.S. where emphasis is exclusively on beef production (ASA Beef Briefs, 2021). Registrations peaked at 89,730 in 1982, dropped to 43,054 in 1998, and increased to 75,122 in 2020. There are 129,859 animals with genomic information. Simmental rank second for U.S. beef semen sales and third in registrations after Angus and Hereford (ASA Annual Report, 2020; ASA, 2021).

In a pedigree analysis, Whitacre and Spangler (2012) determined Simmental has split into nucleus and multiplier levels, with few breeders driving genetic change for the breed. The top states for Simmental seedstock production are Montana, South Dakota, Texas, Kansas, North Dakota, and Nebraska. The authors suggested further improvements in generation interval would be valuable for improving the rate of genetic change.

The American Simmental Association participates in a multi-breed genetic evaluation for single-step genomic enhanced EPDs performed by International Genetic Solutions (IGS). EPDs are computed weekly. Traits evaluated include Calving Ease, Birth Weight, Weaning Weight, Yearling Weight, Average Daily Gain, Maternal Calving Ease, Maternal Milk, Maternal Weaning Weight, Stayability, Docility, Carcass Weight, Yield Grade, Marbling, Backfat, Ribeye Area, Shear Force, All-Purpose Index (Dollars per

cow exposed under an all-purpose-sire scenario), and Terminal Index (Dollars per cow exposed under a terminal-sire scenario) (ASA, 2021).

2.20 SNP chips

Given a sufficiently dense SNP chip, genetic diversity can be examined in greater detail than with pedigree data. With SNP data, a chromosomal region that may be differentiated can be compared between animals or breeds in a way that would not be possible with pedigree data alone. Specific markers, such as deleterious variants or signals of selection, can also be identified. SNP data provides the opportunity to identify genetically unique animals at the genomic level. High density SNP data provides a powerful tool to better understand breeds and estimate the relationship between individuals and breeds (Engelsma et al., 2012; Eynard et al., 2018a). The sheep and cattle genomes are estimated to be 2.86 Gb and 3 Gb in size, respectively (NIH, 2004; Zhang et al., 2013).

The Illumina BovineHD BeadChip consists of 777,962 evenly distributed SNP with a median < 3 kb gap spacing. The chip was developed using 20 Bos taurus, 3 Bos indicus, and 4 Bos taurus x Bos indicus breeds. Ten Simmental were included in the development of the SNP chip and had 624,820 polymorphic loci and a mean minor allele frequency (MAF) of 0.22 (Illumina, 2015). Of the Bos taurus breeds, only 2 breeds had more polymorphic loci. For Simmental, the linkage disequilibrium (LD) at 70 kb was 0.209; in comparison, LD at 70 kb for the breeds studied ranged from 0.16 to 0.26. The Illumina Bovine SNP50 includes 54,001 SNP and was developed using 576 animals from 21 cattle breeds and 6 outgroup species. Three Simmental were included in the development of the chip (Matukumalli et al., 2009). According to Wiggans et al. (2016), there are at least 18 different bovine chips that have been submitted for use in national genetic evaluation for dairy cattle.

The Illumina Infinium OvineHD BeadChip was developed by the International Sheep Genomics Consortium and includes 606,006 SNP with an average spacing of 5 kb. The chip was developed using 75 animals from 41 breeds and wild sheep and includes 30,000 putative functional variants (Kijas et al., 2014). Suffolk was not among the breeds included in the development of the chip, leading to the need to address potential ascertainment bias (Albrechtsen et al., 2010). One approach to removing ascertainment bias is to prune SNP in high LD (Kijas et al., 2012; Edea et al., 2017). The OvineHD chip has been used in genetic diversity studies (Edea et al., 2017), genomic selection (Brito et al., 2017), genome wide association studies (Kijas et al., 2016; Dolebo et al., 2019), and identification of QTL (Posbergh et al., 2019).

2.21 Reference population size

A reference population is generally discussed in terms of developing genomic enhanced breeding values for use in selection. Genomic selection will be feasible for sheep breeds only once sufficiently accurate genomic prediction is possible. This increase in accuracy can be achieved through a combination of increasing the reference population size, increasing marker density or using sequence data instead of SNP, improving statistical procedures, and incorporating phenotypic data. Most livestock breeds have a small N_e (< 200), which increases LD, thereby facilitating genomic selection (Goddard, 2012). Because sheep have high genetic diversity within and between breeds, a larger number of animals are needed for the reference population than with dairy cattle, for example (van der Werf et al., 2014). To evaluate the feasibility of using genomic selection in French dairy sheep, reference populations for each breed ranged from 281 to 2,887. A reference population of 2,000 animals was required for optimizing genetic improvement (Larroque et al., 2014). In French Simmental, (Hozé et al.,

2014) improved accuracy of selection by 0.06 to 0.08 over pedigree-based genetic evaluation using only 181 animals in the training population.

When considering a reference population to represent a breed for genetic diversity studies, there are no clear guidelines. David et al. (2018) analyzed 61 Morada Nova sheep to evaluate genetic diversity of the breed. Al Mamun et al. (2015) had a range of 231 to 265 animals for three breeds and two crossbred populations in Australia when assessing genetic diversity. To assess the population structure of 18 Welsh sheep breeds, a range of 6 to 24 animals were genotyped with the Illumina OvineSNP50 array (Beynon et al., 2015). For genetic diversity studies using microsatellites, Blackburn et al. (2011) included a range of 7 to 46 for each of 28 U.S. sheep breeds. Zhang et al. (2013) evaluated 16 to 22 animals per breed for five U.S. sheep breeds when assessing genetic diversity and differentiation. Kijas et al. (2012) evaluated 74 sheep breeds throughout the world, ranging from 3 to 120 animals per breed. To assess the origins of Caribbean hair sheep, Spangler et al. (2017) evaluated 10 sheep per breed for 29 breeds. Grasso et al. (2014) evaluated 10 to 110 animals for three sheep breeds in Uruguay. In goats, Brito et al. (2017) had a range of 48 to 403 for nine goat breeds to evaluate genetic diversity and detect signatures of selection. When evaluating 14 Italian goat breeds, Nicoloso et al. (2015) included a range of 15 to 32 goats per breed. Visser et al. (2016) determined genetic diversity and population structure of Angora goats from three countries, using 26 to 48 animals per population. In beef and dairy cattle, the genetic structure of 19 breeds was determined with a range of 12 to 53 animals per breed (Gibbs et al., 2009). Among 24 equine and pony populations in Canada, animals sampled ranged from 11 to 60 per breed (Prystupa et al., 2012). When compared to the number of animals required for genomic selection, genetic diversity studies have relied on far fewer animals. As an industry, a set of guidelines based on experimental design (minimum number of breeders, minimum number of animals) needs to be established for use in population genetics studies.

2.22 Measures of within breed genetic diversity/population structure

Expected heterozygosity (H_E) is the most used parameter for measuring genetic diversity within a population. Also called gene diversity, H_E is the expected proportion of heterozygotes if the population were in Hardy-Weinberg equilibrium (Fernández et al., 2004). The actual number of heterozygous animals, H_O, is related to inbreeding (Engelsma et al., 2012). A summary of H_E and H_O across cattle and sheep breeds is summarized in Table 2.5.

Table 2.5 Expected (H_E) and observed (H_O) heterozygosity across studies

Species	**Breed**	**H_E**	**H_O**	**SNP Chip**	**Source**
Cattle	Angus		0.27	BovineHD	(Porto-Neto et al., 2014)
Cattle	Hereford		0.31	BovineHD	(Porto-Neto et al., 2014)
Cattle	Limousin		0.30	BovineHD	(Porto-Neto et al., 2014)
Cattle	Shorthorn		0.25	BovineHD	(Porto-Neto et al., 2014)
Cattle	Simmental		0.338	Bovine SNP50	(Curik et al., 2010)
Cattle	Angus	0.387	0.385	BovineHD	(Kelleher et al., 2017)
Cattle	Charolais	0.381	0.381	BovineHD	(Kelleher et al., 2017)
Cattle	Hereford	0.392	0.388	BovineHD	(Kelleher et al., 2017)
Cattle	Limousin	0.380	0.379	BovineHD	(Kelleher et al., 2017)
Cattle	Simmental	0.386	0.384	BovineHD	(Kelleher et al., 2017)
Sheep	Rambouillet	0.36		Ovine SNP50	(Kijas et al., 2012)
Sheep	Suffolk (Australian)	0.37		Ovine SNP50	(Kijas et al., 2012)
Sheep	Suffolk (Irish)	0.33		Ovine SNP50	(Kijas et al., 2012)
Sheep	Suffolk	0.34		OvineHD	(Kijas et al., 2014)
Sheep	Border Leicester	0.32		OvineHD	(Kijas et al., 2014)
Sheep	Poll Dorset	0.33		OvineHD	(Kijas et al., 2014)
Sheep	Border Leicester	0.30	0.30	Ovine SNP50	(Al-Mamun et al., 2015)
Sheep	Poll Dorset	0.34	0.34	Ovine SNP50	(Al-Mamun et al., 2015)
Sheep	Merino	0.38	0.38	Ovine SNP50	(Al-Mamun et al., 2015)
Sheep	Corriedale	0.355	0.355	Ovine SNP50	(Grasso et al., 2014)
Sheep	Creole	0.258	0.285	Ovine SNP50	(Grasso et al., 2014)
Sheep	Merino	0.362	0.377	Ovine SNP50	(Grasso et al., 2014)
Sheep	Suffolk		0.33	Applied Biosystems Axiom Ovine (50K)	(Davenport et al., 2020)

The level of genetic diversity present within populations can be measured by the number of polymorphic loci and their allele frequencies distributions (Brito et al., 2015). Grasso et al. (2014) defined MAF categories as fixed (MAF = 0), rare (MAF < 0.01), and highly polymorphic (MAF 0.3 – 0.5). The MAF categories for three breeds are shown in Table 2.6. These results suggest either the Creole breed has less genetic diversity then the other two breeds or could be indicative of ascertainment bias of the Ovine SNP50 BeadChip.

Table 2.6 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for Corriedale, Creole, and Merino sheep using the Ovine SNP50 BeadChip (Grasso et al., 2014)

Breed	**0**	**< 0.01**	**0.3 – 0.5**
Corriedale	1.8	4.4	50.1
Creole	26.9	27.4	36.0
Merino	2.9	3.4	50.9

Runs of homozygosity (ROH) are contiguous lengths of homozygous genotypes that are due to parents transmitting identical haplotypes to their offspring. Long ROH represent recent inbreeding, while shorter ROH are attributed to historical events, such as breed founder effects. A typical approach to measuring ROH using SNP is a sliding window of 50 SNP with a minimum length of 500 kb, a minimum of 25 SNP, allowing for 2 missing SNP, and allowing a maximum of 1 possible heterozygous SNP. ROH can be reported as the mean sum of ROH per animal and summarized by ROH length categories. Angus and Hereford had a high number of short (< 5 Mb) ROH while Holstein had a high number of long (> 20 Mb) ROH (Purfield et al., 2012). In sheep, Border Leicester, Poll Dorset, and Merino had 12,561, 9,875, and 2,008 total ROH, respectively. Three Poll Dorset animals had a total ROH comprising almost 20% of the genome (Al-Mamun et al., 2015). Across the genome, there can be nonuniform patterns of ROH, referred to as hotspots and coldspots. Hotspots have frequent ROH and indicate a reduced level of diversity; coldspots have infrequent ROH. Possible reasons for these differences across the genome

include stochasticity in recombination events, demographic processes, and positive selection (Pemberton et al., 2012).

Linkage disequilibrium (LD) is the nonrandom association of alleles at two or more loci (Conner and Hartl, 2004), measured as the squared correlation (r^2) between alleles (Hill and Robertson, 1968). Linkage disequilibrium ranges from 0 (no LD) to 1 (complete LD) between two markers. Linkage disequilibrium is strongly influenced by population history, starting with the bottleneck of domestication and continuing through evolutionary forces including genetic drift, migration, mutation, and selection (Brito et al., 2015). Linkage disequilibrium is typically reported as the r^2 decay over distance or as the r^2 at a particular distance apart (e.g., 10 kb) (Al-Mamun et al., 2015). An r^2 of > 0.2 is typically considered sufficient for use in genomic selection. Populations with a low LD will require a higher density SNP chip for genomic selection than those with a high LD (Brito et al., 2015). In beef cattle, LD (r^2) at 10 kb was 0.46, 0.49, and 0.25 for Angus, Hereford, and Brahman, respectively. At 70kb, LD (r^2) had declined to 0.20 for Angus, 0.23 for Hereford, and 0.13 for Brahman. The authors concluded the higher LD for Bos taurus breeds was due to a smaller ancestral population and a stronger bottleneck during breed formation than Bos indicus breeds (Porto-Neto et al., 2014). For sheep breeds, LD (r^2) at 10 kb was 0.34, 0.27, and 0.33 for Border Leicester, Merino, and Poll Dorset, respectively (Al-Mamun et al., 2015). Zhang et al. (2013) reported the highest to lowest LD for Suffolk, Columbia, Rambouillet, Targhee, and Polypay. For sheep breeds, Kijas et al. (2012) explained the low LD typically found for sheep breeds as caused by a broad sampling of wild ancestors, less severe bottlenecks during breed formation, and low levels of selection intensity relative to other livestock species.

Reported N_e for cattle and sheep breeds are presented in Table 2.7. Of the three breeds reported by Al-Mamun et al. (2015), the authors stated smaller N_e for Border Leicester and Poll Dorset relative to Merino was associated with bottlenecks during breed formation.

Table 2.7 Cattle and sheep breed effective population size (N_e) across studies

Species	Breed	N_e	Source
Cattle	Angus	136	(Gibbs et al., 2009)
Cattle	Charolais	110	(Gibbs et al., 2009)
Cattle	Hereford	97	(Gibbs et al., 2009)
Cattle	Limousin	174	(Gibbs et al., 2009)
Cattle	Red Angus	85	(Gibbs et al., 2009)
Cattle	Charolais*	198 – 958	(Leroy et al., 2013)
Cattle	Holstein*	49 – 93	(Leroy et al., 2013)
Cattle	Limousin*	168 – 740	(Leroy et al., 2013)
Cattle	Salers*	51 – 323	(Leroy et al., 2013)
Cattle	Simmental*	110 – 169	(Leroy et al., 2013)
Sheep	Border Leicester	140	(Al-Mamun et al., 2015)
Sheep	Merino	152	(Al-Mamun et al., 2015)
Sheep	Poll Dorset	348	(Al-Mamun et al., 2015)
Sheep	Dorset*	21 – 68	(Leroy et al., 2013)
Sheep	Finn*	35 – 96	(Leroy et al., 2013)
Sheep	Hampshire*	68 – 145	(Leroy et al., 2013)
Sheep	Southdown*	42 – 109	(Leroy et al., 2013)
Sheep	Suffolk*	69 – 310	(Leroy et al., 2013)

*Range of N_e from multiple estimation methods

2.23 Validation across species and breeds

A sampling strategy designed to apply to all livestock species and breeds must consider the different breed histories and population structures for each population. Bos taurus breeds have been identified as having lower genetic diversity than Bos indicus breeds due to a smaller ancestral population and larger bottlenecks at breed formation (Gibbs et al., 2009). Sheep breeds are associated with higher genetic diversity than cattle breeds due to a large ancestral population and a lack of selection intensity (Kijas et al., 2012).

When assessing methods to compute N_e, Leroy et al. (2013) found a significant interaction between computation method and species. The authors recommended attention to species and specific population structure when considering computation method. For example, when inbreeding is used as

an indicator of genetic diversity, it often failed to consider the substructure within a population, leading to biased conclusions. Since artificial insemination (A.I.) is a common practice in cattle, but not in sheep, unbalanced progeny sizes can also influence the preferred computation method for N_e. An additional factor to consider is the extent of pedigree knowledge or errors, which was also identified by Engelsma et al. (2011).

For selection scenarios using markers with strong LD, the strategy was effective by "extending effects of the conservation criteria to the whole genome." Without strong LD, marker-based selection scenarios were inferior to pedigree-based ones. The extent of LD among breeds may produce different results (Saura et al., 2008).

To compare the effectiveness of selection strategies, Engelsma et al. (2011) compared the results for kinship, MAF, and percentage fixed alleles. These measures were computed for the entire genome and for each chromosome. The measures were also compared based on sampling different numbers of animals. The same measures to compare the effectiveness of a sampling strategy within a breed can be used to evaluate the effectiveness of a sampling strategy across species and breeds, including percentage of alleles captured, MAF, H_O, H_E, ROH, LD, and N_e.

Studies have assessed genetic diversity measures between sheep and cattle, including population structure and genetic merit (Avendaño et al., 2003) and effective population size (Leroy et al., 2013). Genetic diversity has been evaluated between sheep breeds in many studies (Handley et al., 2007; Grasso et al., 2014; Al-Mamun et al., 2015; Edea et al., 2017; Ahbara et al., 2018). Comparing cattle breeds for genetic diversity (Gibbs et al., 2009; Makina et al., 2015), population structure (Kelleher et al., 2017; Stronen et al., 2019), LD and effective number of founders (Hozé et al., 2013), and limiting migrant contributions (Wellmann et al., 2012) have all contributed to a greater understanding of how to evaluate differences between breeds.

2.24 Conclusion

Many sampling strategies have been developed and used for a wide variety of purposes across livestock species and breeds. Further, sampling strategies have been used to assess the capture of genetic diversity. Tools that use OCS have been developed and used to improve genetic merit in a small population (Gourdine et al., 2012b), in breeding programs (Skaarud et al., 2011), and to manage inbreeding (Windig and Oldenbroek, 2015). Computation of GCI has been incorporated into the ENDOG software package (Gutiérrez and Goyache, 2005) and has been used to evaluate the presence of founder alleles in the present population for cattle (Alderson, 1992), horses (McManus et al., 2013; Giontella et al., 2019), sheep (McManus et al., 2019), and goats (Menezes et al., 2015; Oliveira et al., 2016).

Although these sampling strategies have been used in a variety of ways, they have not been directly evaluated for the ability to capture all the genetic diversity of a population. The first objective of this research is to develop a sampling strategy using OCS and GCI using pedigree and breeding value data from a simulated population and validated with real data. The real data will include only animals with available pedigree and SNP data. The second objective uses the strategy developed in the first objective to identify a subset of 25, 50, and 100 animals from each breed. The population structure for each subset will be compared to the larger population using both quantitative and molecular measures to determine if the genetic diversity has been captured. In the third objective, the sampling strategies will be compared for efficacy across species and breeds.

CHAPTER III

DEVELOP A SAMPLING STRATEGY TO MAXIMIZE THE GENETIC DIVERSITY OF SAMPLED ANIMALS

3.1 Introduction

Research inherently requires a sampling of animals to be included in studies, whether it be within a herd or flock, or across a region, area, breed, or species. A formal sampling strategy to ensure the genetic diversity of the sampled animals is maximized will provide assurance that the sampled animals are representative of the entire population. Sampling strategies can be pedigree- or genomic-based, although it is rare for every candidate to have available genomic data. In fact, the sampling procedure may be used to identify animals to be genotyped. Readily available pedigree data for most U.S. breeds makes sampling strategies using pedigree-based methods a cost-effective option.

Data simulation provides a low-cost opportunity to test sampling strategies using data where every factor is known, including parentage, measured traits, and, most importantly, the underlying genomic variants. A variety of populations can be simulated and tested, and the long-term impacts can be assessed more easily than with a real population. Once sampling strategies have been evaluated using a simulated population, the lessons learned can then be applied to real data (Leroy and Rognon, 2012; Daetwyler et al., 2013; Windig and Oldenbroek, 2015).

Two sampling strategies designed to maintain genetic diversity in a population include optimal contribution selection (OCS) and the genetic conservation index (GCI). Traditionally, OCS has been used to restrict mean relatedness in the population while maximizing genetic gain. Studies have shown the superior combination of minimizing inbreeding while maximizing genetic gain for OCS when compared to other methods, such as truncation selection (Avendaño et al., 2003; Gourdine et al., 2012a; Howard et al., 2014). With OCS, emphasis is placed on selection of the parents based on constraining the extent

of co-ancestry. Attention to the current co-ancestry of parents is expected to minimize inbreeding in future generations (Meuwissen, 1997; Meuwissen, 2009). In this research, OCS was applied to manage co-ancestry without regard to the impact on genetic gain. The approach of GCI is to maintain the founder alleles in the population. If the founder alleles are maintained, the genetic diversity of the initial population will be maintained. The greater the number of founders represented in an animal's pedigree, the higher the GCI index. The highest possible GCI index would be achieved by an animal having an equal contribution from all founder ancestors (Alderson, 1992). GCI has been successfully applied to select bulls in White Park cattle (Alderson, 1992) and to evaluate changes in founder representation over time (McManus et al., 2013; Menezes et al., 2015; Giontella et al., 2019).

The objectives of this study were to 1) simulate a population to use OCS and GCI as sampling strategies to capture the genetic diversity of the population; and 2) to use both the GCI and OCS sampling strategies investigated to capture the genetic diversity of two real populations, a sheep breed (Suffolk) and a beef breed (Simmental).

3.2 Materials and Methods

Data simulation. Simulation of a sheep population was performed using AlphaSimR. Founder haplotypes were created by defining the length of each of the 26 autosomal chromosomes (Howe et al., 2020). The chromosome lengths ranged from 42,034,648 bp (chromosome 24) to 275,406,953 bp (chromosome 1). Number of markers and quantitative trait loci (QTL) were defined based on chromosome length and as described by Vargas Jurado et al. (2021) for a total of 53,901 markers and 2,449 QTL. Markers per chromosome ranged from 925 to 6,059 and QTL ranged from 42 to 275. A founder population of 1,000 animals was generated using Markovian Coalescent Simulation (MaCS) with an effective population size (N_e) of 250 (Chen et al., 2009). Weaning weight was simulated as a

phenotypic trait with a mean of 29 kg and a heritability of 0.15, which is the heritability used by the National Sheep Improvement Program in the sheep genetic evaluations. AlphaSimR first samples QTL effects from an initial standard normal distribution and then scales the values to the mean and variance defined by the user in the simulation (Gaynor et al., 2021).

The founder population was then increased in size over 30 years using random mating until reaching a population size of 13,880 lambs per year. Litter size was set to two with sex of lamb assigned randomly. Ewe lambs were randomly selected to replace 20 percent of the ewe flock each year plus an additional 10 percent were kept to increase the population size. There was one ram per 22 ewes, with a replacement rate of 60 percent. By year 30, there were 6,940 ewes and 315 rams in the flock.

The final data set was generated using phenotypic selection for weaning weight, which is representative of selection in the sheep industry for terminal sire breeds. Phenotypic selection was performed by sorting the phenotypic values from highest to lowest for the available pool of animals and selecting the required number of animals for each year based on the highest phenotypic values (Gaynor et al., 2021). Phenotypic selection for weaning weight was performed for 15 additional years after the founder population was established. Population size was fixed at 13,662 lambs per year with a replacement rate of 20 and 60 percent for ewes and rams, respectively, each year. Weaning weight estimated breeding values (EBV) and phenotypic weaning weight were reported for each animal. The average allele substitution effects were used to calculate the breeding values for each animal by summing the breeding values at each locus. The true genotypic value was calculated by summing its coded genetic value for its genotype across all quantitative trait nucleotide loci. A random residual deviate was added to each animal's true genetic value to determine the phenotypic value (Faux et al., 2016). Pedigree records were traced back until all ancestors were unknown, for a total of 211,951 records. There were 1,600 unique sires with a range of 68 to 1020 offspring and 25,819 unique dams with a range of 2 to 32 offspring. Inbreeding coefficients were calculated using the Animal Breeders

Toolkit (ABTK) (Golden et al., 1992) and the trend in inbreeding determined. The rate of inbreeding was computed as:

$$\Delta F_i = \frac{F_i - F_{i-1}}{1 - F_{i-1}},$$

where F_i and F_{i-1} were the average inbreeding at *i* and *i-1* years, respectively. The final dataset consisted of 204,930 animals, which included all lambs from 15 years of phenotypic selection.

Single nucleotide polymorphisms (SNP) for each animal were analyzed using PLINK (Purcell et al., 2007). The probability that an individual will be heterozygous at a given locus, expected (H_E) and observed heterozygosity (H_O), were computed for the entire population and by year. Molecular inbreeding, F_{IS}, measured as a heterozygote deficiency (or homozygote excess) across each sample (Wright, 1951), was computed as:

$$F_{IS} = (H_E - H_0)/H_E$$

The minor allele frequency (MAF) categories were determined as fixed (MAF = 0), rare (MAF < 0.01), moderate (MAF 0.01 - 0.3), and high (MAF 0.3 - 0.5) (Grasso et al., 2014; Wilson et al., 2022). Runs of homozygosity (ROH) were computed for year 15 animals using the detectRUNS package in R (Biscarini et al., 2019) with a sliding window of 50 SNP with a minimum length of 1,000 kb, a minimum of 30 SNP, allowing for 1 missing SNP, and allowing a maximum of 1 possible heterozygous SNP. The ROH class categories were determined as 1 - 5, 5 - 10, 10 - 20, 20 - 40, and > 40 Megabase pairs. Recent N_e was computed using the linkage disequilibrium method of Waples and Do (2008) as implemented in NeEstimator v2.1 (Do et al., 2014). The dataset was pruned to remove SNP with MAF < 0.01 and randomly reduce the number of markers by 75 percent using PLINK, leaving 10,202 markers, which were used to compute model-based population structure using ADMIXTURE (Alexander et al., 2009). The population substructure analysis used the genotype matrix to estimate the subpopulation proportions and the population allele frequencies to assign individuals to the subpopulations. The number of populations, K, can be determined using the lowest cross validation error compared to other K values

(Alexander et al., 2009). For the simulated population, the cross validation error continued to decline through K = 15. Because this was a within breed analysis, additional subpopulations were not computed. For each replicate of the co-ancestry coefficient matrix, Q, the pophelper package in R was used to align and merge the runs (Francis, 2017). The output was visualized using STRUCTURE PLOT (Ramasamy et al., 2014).

Optimal contribution selection. The optiSel package in R was used to select 100, 50, and 25 animals using OCS (Wellmann, 2019). While OCS can be applied to maximize genetic gain while minimizing kinship, the objective of this research was to maximize genetic diversity. Therefore, animals were selected to minimize the average kinship without regard to genetic gain. Computationally, OCS maximizes $G = c'EBV$, while restricting: $\bar{A}_p = c'Ac$, and restricting: $c'Q = 0.5$, where G was the average genetic level of the parents weighted by their number of offspring, EBV was a vector of estimated breeding values of the parents, c was a vector of genetic contributions of the parents (fraction of offspring each parent obtains, scaled such that it sums to 0.5 for males and to 0.5 for females; $c_i = 0$ implies that animal i was not selected), $\bar{A}_p$ was the desired value of the average genetic relationships of the parents where A was the numerator relationship matrix of the selection candidates, and Q was a design matrix indicating the gender of the candidates (first column is 1 for male, 0 for females; second column is 1 for female, 0 for males) (Meuwissen, 2009). OptiSel provides the option for five optimization solvers. The optimization solver selected was from the cccp package (Pfaff, 2014) for solving cone constrained convex programs and was recommended by the author for short run time and accuracy for minimizing kinships. Other solvers were recommended for different optimizations, for example, maximizing genetic gain at native alleles. The complete pedigree was included (n = 211,951), with all animals in year 15 (n = 13,662) being designated as candidates for selection for sampling.

Genetic conservation index. The ENDOG software package (Gutiérrez and Goyache, 2005) was used to compute the GCI for all animals in the pedigree (n = 211,951). The GCI for an animal was computed as $1/\Sigma P_i^2$, where P_i was the proportion of genes of founder animal *i* in the pedigree (Alderson, 1992). The highest 100 (GCI 100), 50 (GCI 50), and 25 (CGI 25) indexing animals from year 15 were selected (Gutiérrez and Goyache, 2005).

Suffolk sheep. Pedigree data were obtained from the National Sheep Improvement Program (NSIP). These data contained 244 Suffolk sheep genotyped with the Illumina OvineHD BeadChip, which includes 606,006 SNP markers (Illumina, 2015). Sheep for genotyping were selected based on pedigree relationships to maximize the genetic diversity available among the animals with an available DNA sample (Wilson et al., 2022). Pedigree records were traced back for the 244 sheep until all ancestors were unknown, resulting in a total of 1,565 animals in the full pedigree. There were 496 unique sires with a range from 1 to 20 offspring and 929 dams with a range from 1 to 8 offspring. To evaluate how well the subset of the Suffolk breed represented the entire Suffolk breed, pedigree-based measures of effective population size were compared for the subset and full breed for increase in inbreeding by maximum generations, complete generations, and equivalent generations (Gutiérrez and Goyache, 2005). Pedigree records for the full Suffolk breed included those recorded in NSIP, with birth years ranging from 1973 to 2019. Pedigree records were traced back until all ancestors were unknown, resulting in a total of 64,310 animals (Wilson et al., 2022).

Individual inbreeding coefficients were computed (Gutiérrez and Goyache, 2005) and inbreeding trend and rate of inbreeding were reported. Weaning weight EBV, Carcass Plus Index, and phenotypic weaning weight, obtained from NSIP, were plotted over time. The Carcass Plus Index includes post-weaning weight, fat depth, and eye muscle depth in a 60:20:20 ratio (Emenheiser and Notter, 2011).

Quality control measures performed for the molecular analyses were conducted using the PLINK software package (Purcell et al., 2007). Only autosomal chromosomes and markers mapped to the genome were retained for further analysis, leaving 577,401 markers. One animal was removed for having a call rate < 0.95. This data set was used to compute heterozygosity and molecular inbreeding (F_{IS}) measures with PLINK. Runs of homozygosity were computed using the detectRUNS package in R as described for the simulated population (Biscarini et al., 2019). Current N_e was estimated using the linkage disequilibrium method in NeEstimator (Waples and Do, 2008; Do et al., 2014). Minor allele frequencies were computed using PLINK with a dataset further reduced to remove markers with a call rate < 0.80, leaving 546,938 markers. The dataset was pruned in PLINK to randomly reduce the number of markers (n = 49,773) and was used to compute model-based population structure using ADMIXTURE (Purcell et al., 2007; Alexander et al., 2009).

Molecular analyses used to quantify the full population included measures of heterozygosity (H_E and H_O), Wright's inbreeding coefficient (F_{IS}), MAF categories, N_e, ROH by size class, ROH by chromosome, ROH by animal, and model-based population substructure using ADMIXTURE. For the substructure analysis, the number of populations, K, was determined using the lowest cross validation error compared to other K values (Alexander et al., 2009). For each replicate of the co-ancestry coefficient matrix, Q, produced by ADMIXTURE, the CLUMPP program using CLUMPAK software was used to permute the matrices to find a close match among iterated runs (Jakobsson and Rosenberg, 2007; Kopelman et al., 2015). The output from CLUMPAK was summarized using STRUCTURE PLOT, allowing visualization of the results in bar plots (Ramasamy et al., 2014).

Simmental cattle. Pedigree, performance, and genotypic (SNP) data were obtained from the American Simmental Association (ASA). Animals that were 84 percent and higher Simmental were included in the analyses; lower percentage Simmental were used in pedigree building and computation

of inbreeding coefficients. Genotypic data was obtained from a variety of SNP chips. All genotypic data files were merged and overlapping SNP across chips were identified by comparing the associated map files for each chip, with as many SNP and animals as possible retained. After merging, there were 5,613 Simmental genotyped with 29,449 SNP markers. Pedigree records were traced back for the 5,613 cattle until all ancestors were unknown, resulting in a total of 54,790 animals in the full pedigree. Pedigree animals less than 84 percent Simmental were excluded from further computations, e.g., mean weaning weight, but were included as pedigree animals. There were 3,645 breeder codes present in the full dataset. There were 9,118 males and 25,344 females included in the pedigree that were 84 percent Simmental and above. There were 8,512 unique sires with a range of 1 to 349 offspring and 23,906 unique dams with a range of 1 to 24 offspring. The unique sires and dams could be a lower percentage animal with a higher percentage mate.

Individual inbreeding coefficients were computed using the ENDOG software package (Gutiérrez and Goyache, 2005) and the inbreeding trend and rate of inbreeding was calculated. Weaning weight Expected Progeny Difference (EPD), All-Purpose Index (API), and phenotypic weaning weight were plotted over time. The API is for use in herds where daughters are kept as replacements with the rest of the heifers and steers sold based on yield and grade and includes birth weight, calving ease, weaning weight, and yearling weight (Saad et al., 2020; ASA Beef Briefs, 2021).

Quality control measures performed for the molecular analyses were conducted using PLINK.The final sifted data set consisting of 5,613 individuals was used to compute heterozygosity and molecular inbreeding (F_{IS}) measures with PLINK. Runs of homozygosity measures were computed using the detectRUNS package in R (Biscarini et al., 2019). Current N_e was computed using the linkage disequilibrium method in NeEstimator v2.1 (Waples and Do, 2008; Do et al., 2014). The data set was also used to compute the model-based population structure using ADMIXTURE (Alexander et al., 2009).

The population structure of the full Simmental population was assessed by measuring the levels of heterozygosity, Wright's inbreeding coefficient, MAF categories, N_e, and ROH by size class, chromosome, and animal. Population substructure was determined using ADMIXTURE. The number of populations, K, can be determined using the lowest cross validation error compared to other K values (Alexander et al., 2009). For the Simmental population, the cross validation error continued to decline through K = 15; because this was a within breed analysis, additional subpopulations were not computed. For each replicate of the co-ancestry coefficient matrix, Q, produced by ADMIXTURE, the pophelper package in R was used to align and merge the runs (Francis, 2017). The CLUMPAK package, which was used for the Suffolk ADMIXTURE analysis, is limited to 5,000 animals. The output from pophelper was summarized using STRUCTURE PLOT, allowing visualization of the results in bar plots (Ramasamy et al., 2014).

3.3 Results

Data simulation. Chromosome length (bp), number of markers, and number of QTL per chromosome are shown in Table 3.1. Inbreeding was low, with a mean of 0.002 and a range of 0 to 0.31. The rate of inbreeding was 0.0002 per year. Inbreeding trend by year is shown in Figure 3.1. Weaning weight EBV and phenotypic weaning weight trend by year are shown in Figures 3.2 and 3.3, respectively. A linear regression of weaning weight EBV on year was calculated. The estimated slope corresponding to weaning weight EBV was 0.123 kg (SE = 0.003, $P < 0.001$). Similarly, a linear regression of phenotypic weaning weight on year was calculated. The estimated slope for weaning weight was 0.124 kg (SE = 0.003, $P < 0.001$).

Table 3.1 Chromosome length (bp), number of markers, and number of QTL per chromosome for the simulated population

Chromosome	Chromosome length	Number of markers	Number of QTL
1	275406953	6059	275
2	248966461	5477	249
3	223996068	4928	224
4	119216639	2623	119
5	107836144	2372	108
6	116888256	2572	117
7	100009711	2200	100
8	90615088	1994	91
9	94583238	2081	95
10	86377204	1900	86
11	62170480	1368	62
12	79028859	1739	79
13	83079144	1828	83
14	62568341	1377	63
15	80783214	1777	81
16	71693149	1577	72
17	72251135	1590	72
18	68494538	1507	68
19	60445663	1330	60
20	51176841	1126	51
21	49987992	1100	50
22	50780147	1117	51
23	62282865	1370	62
24	42034648	925	42
25	45223504	995	45
26	44047080	969	44

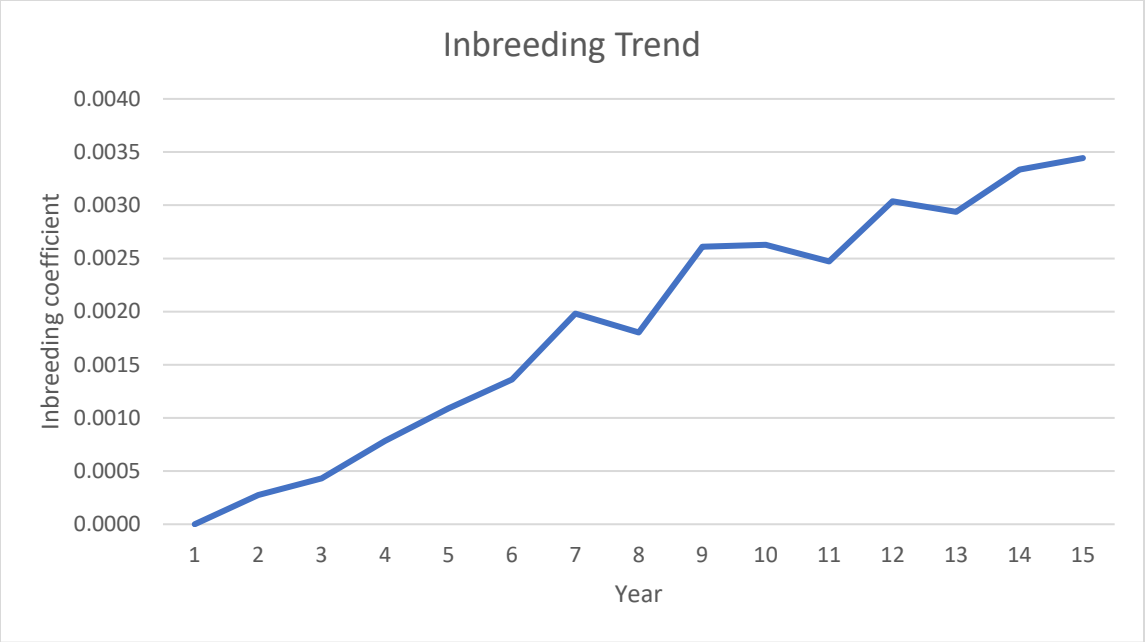

Figure 3.1 Inbreeding trend by year for the simulated population

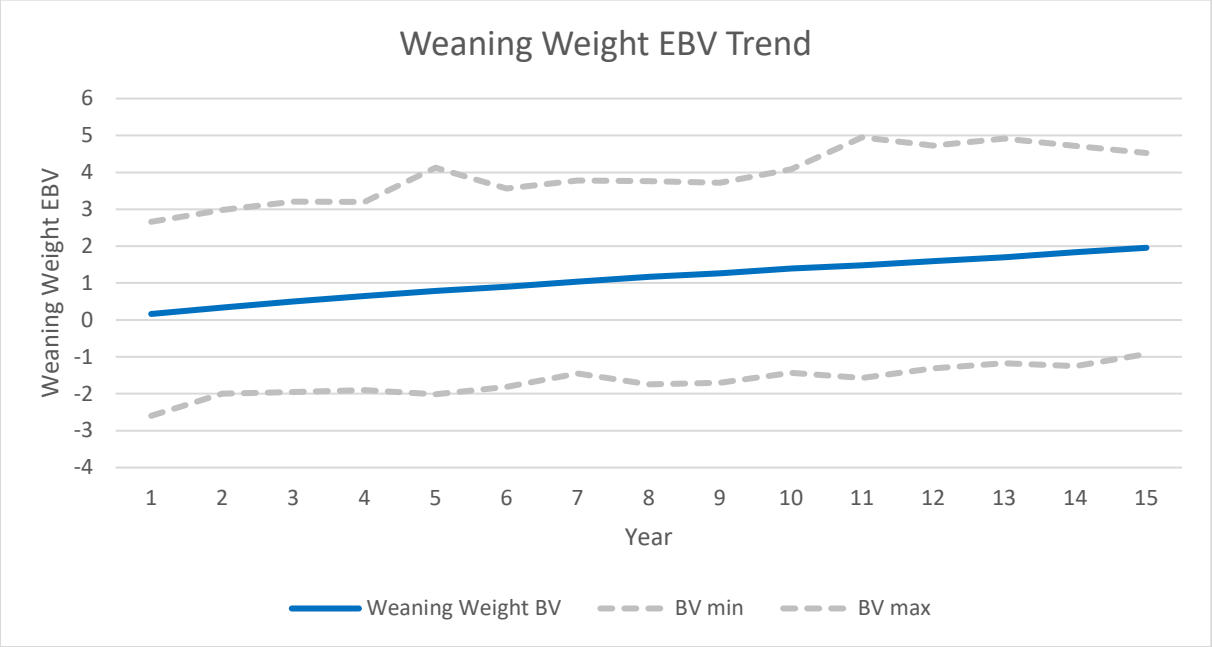

Figure 3.2 Weaning weight estimated breeding value (EBV) trend by year for the simulated population

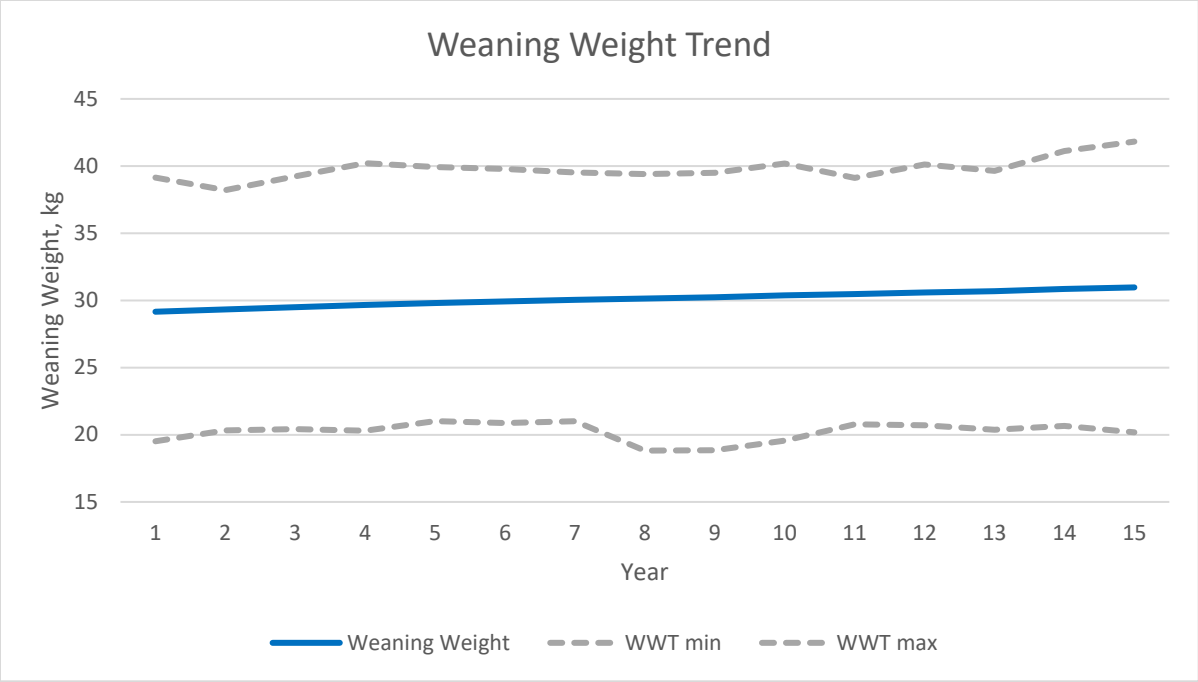

Figure 3.3 Phenotypic weaning weight trend by year for the simulated population

Overall, H_E and H_O were both 0.257; the values were expected to be the same for a simulated population. Over time, both H_E and H_O increased from 0.255 in year 1 to 0.259 in year 15. The number of fixed alleles increased each year, as expected in a population undergoing selection (Table 3.2) (Eynard et al., 2018b). Molecular inbreeding (F_{IS}) for year 15 ranged from -0.10 to 0.31 with a mean of 0.00, where negative values indicate a heterozygote excess and positive values indicate a heterozygote deficiency. The SNP in MAF categories for year 15 are presented in Table 3.3. There were 14.2 percent of SNP in the fixed category (MAF = 0) and 9.9 percent in the rare category (< 0.01). The majority of SNP were in the moderate and high categories. The ROH were summarized by class size (Mbps) and percentage (Table 3.4) and by chromosome for year 15, where chromosome 1 contained the most total runs (13.4%) and chromosome 24 had the fewest (1.0%) (Table 3.5). The majority of ROH were short runs of less than 5 Mbps with less than 1 percent of the runs being 20 Mbps or greater. The ROH per animal ranged from 75 to 140 with a mean of 107.6. The current N_e was computed as 289.8 for the population. Fifteen

subpopulations were visualized for year 15 animals where the proportional assignment of each animal was represented as a column (Figure 3.4).

Table 3.2 Number of fixed alleles by year for the simulated population

Generation	No. fixed alleles
1	6,647
2	6,704
3	6,838
4	6,965
5	7,065
6	7,123
7	7,205
8	7,306
9	7,337
10	7,433
11	7,471
12	7,522
13	7,534
14	7,607
15	7,644

Table 3.3 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for the simulated population for year 15

MAF Category	% of SNP
Fixed (0)	14.2
Rare (< 0.01)	9.9
Moderate (0.01 – 0.3)	52.1
High (0.3 - 0.5)	23.8

Table 3.4 Runs of homozygosity (ROH) by size class and total ROH percentage for the simulated population for year 15

ROH Class Category (Mbps)	No. ROH	Percent Total ROH
1 - 5	1,161,438	79.0
5 - 10	241,548	16.4
10 - 20	56,807	3.9
20 - 40	9,644	0.7
> 40	1,221	0.1

Table 3.5 Runs of homozygosity (ROH) count and percentage by chromosome for year 15

Chromosome	ROH Count	ROH Percentage
1	197,233	13.4
2	190,456	13.0
3	187,092	12.7
4	79,550	5.4
5	62,340	4.2
6	71,146	4.8
7	63,814	4.3
8	44,376	3.0
9	51,722	3.5
10	52,542	3.6
11	28,509	1.9
12	39,895	2.7
13	56,154	3.8
14	30,208	2.1
15	40,477	2.8
16	39,758	2.7
17	30,295	2.1
18	32,241	2.2
19	25,733	1.7
20	23,344	1.6
21	18,449	1.3
22	21,479	1.5
23	32,553	2.2
24	14,982	1.0
25	20,148	1.4
26	16,162	1.1

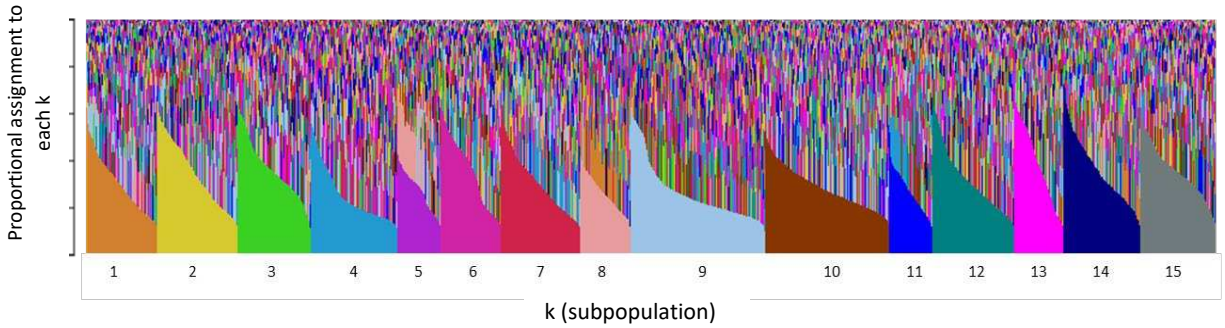

Figure 3.4 Model-based population structure of the simulated population (n = 13,662), displaying the proportional assignment of each animal as a column and sorted by highest proportional assignment to a subpopulation

Optimal contribution selection. The OCS optimization converged after 28 iterations. From the 13,662 selection candidates, the animals with the highest 25, 50, and 100 optimal contributions were identified. The optimal contributions ranged from 0.003 to 0.008 for OCS 100, 0.004 to 0.008 for OCS 50, and 0.005 to 0.008 for OCS 25.

Genetic conservation index. The GCI for the entire population was computed and animals were selected from year 15. The top GCI value was 93.1 for all GCI samples and the low was 69.1, 73.8, and 78.1 for GCI 100, GCI 50, and GCI 25, respectively. In comparison, the lowest value for year 15 was 2.0.

Suffolk sheep. The full pedigree represents 36 flocks. Genotyped animals included 115 ewes and 129 rams. The mean pedigree-based inbreeding coefficient was 0.011 with a range from 0 to 0.257. The inbreeding trend by birth year was shown in Figure 3.5. Inbreeding was low, with a slight increase over time. Unknown parentage in early years contributed to low and fluctuating inbreeding levels in the first few years. The rate of inbreeding was 0.0003 per year. Weaning weight EBV, Carcass Plus Index, and phenotypic weaning weight trend by birth year were shown in Figures 3.6, 3.7, and 3.8, respectively. Of the 1,565 animals in the pedigree, 894 contributed phenotypic data to the EBV and Index calculations. A

linear regression of each trait on year was calculated. The estimated slope for weaning weight EBV was 0.146 (SE = 3.030, $P < 0.001$). The estimated slope corresponding to Carcass Plus Index was 0.428 (SE = 0.193; $P < 0.001$). The estimated slope for phenotypic weaning weight was 0.008 (SE = 0.781; $P = 0.82$). The average Carcass Plus Index trended upward with a maximum value of 162.46 in the most recent year. Phenotypic weaning weight trend remained relatively consistent.

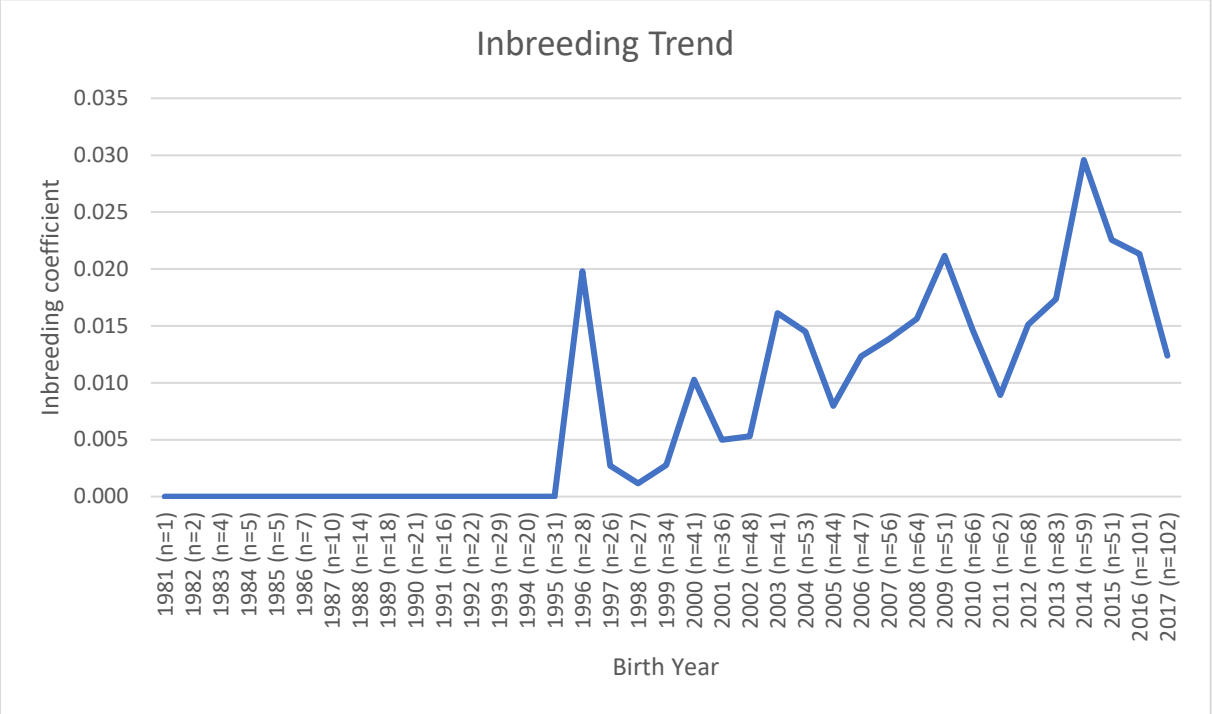

Figure 3.5 Suffolk inbreeding trend by birth year

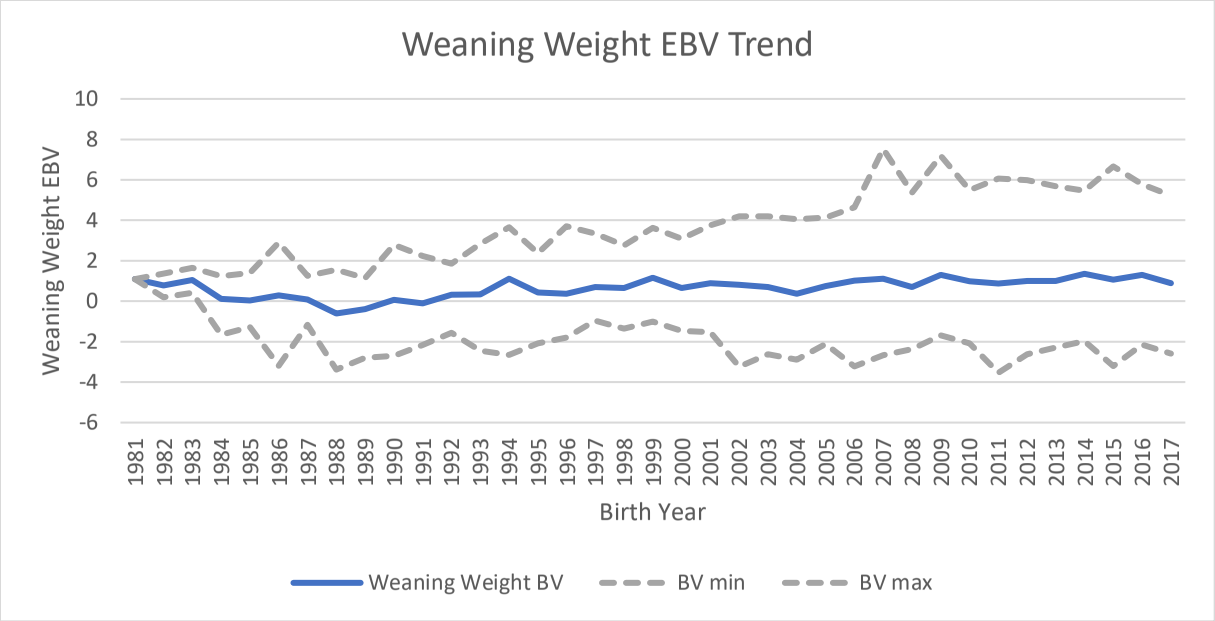

Figure 3.6 Suffolk weaning weight estimated breeding value (EBV) trend by birth year

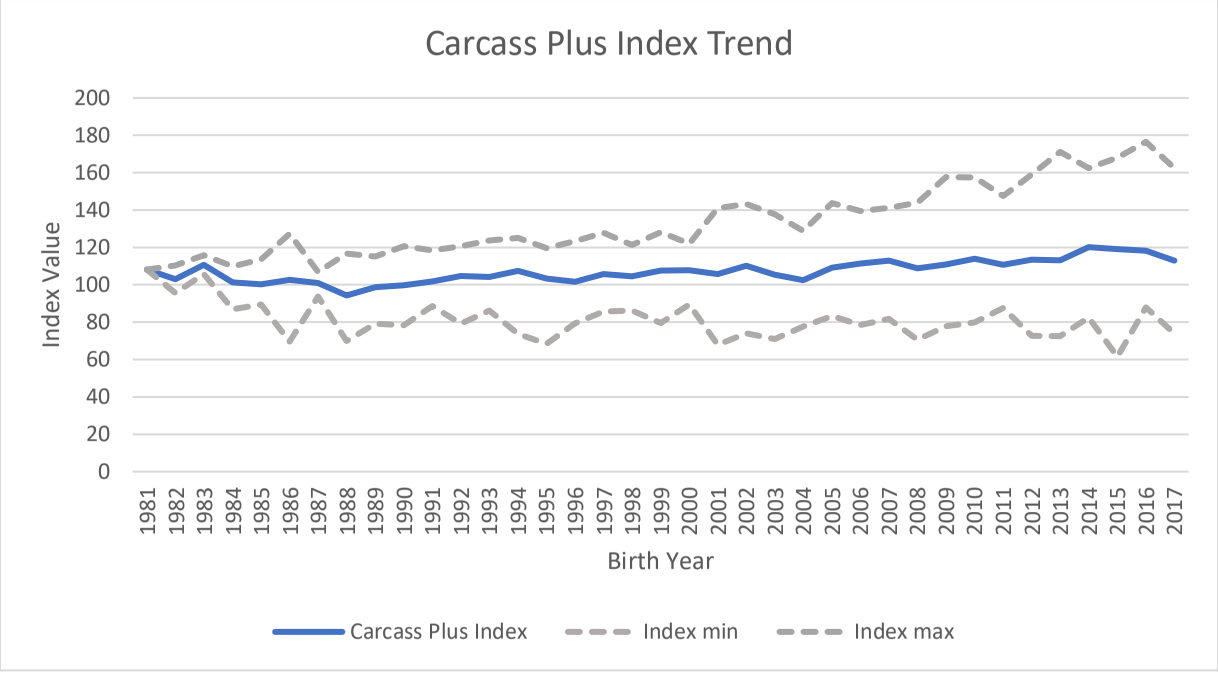

Figure 3.7 Suffolk Carcass Plus Index trend by birth year

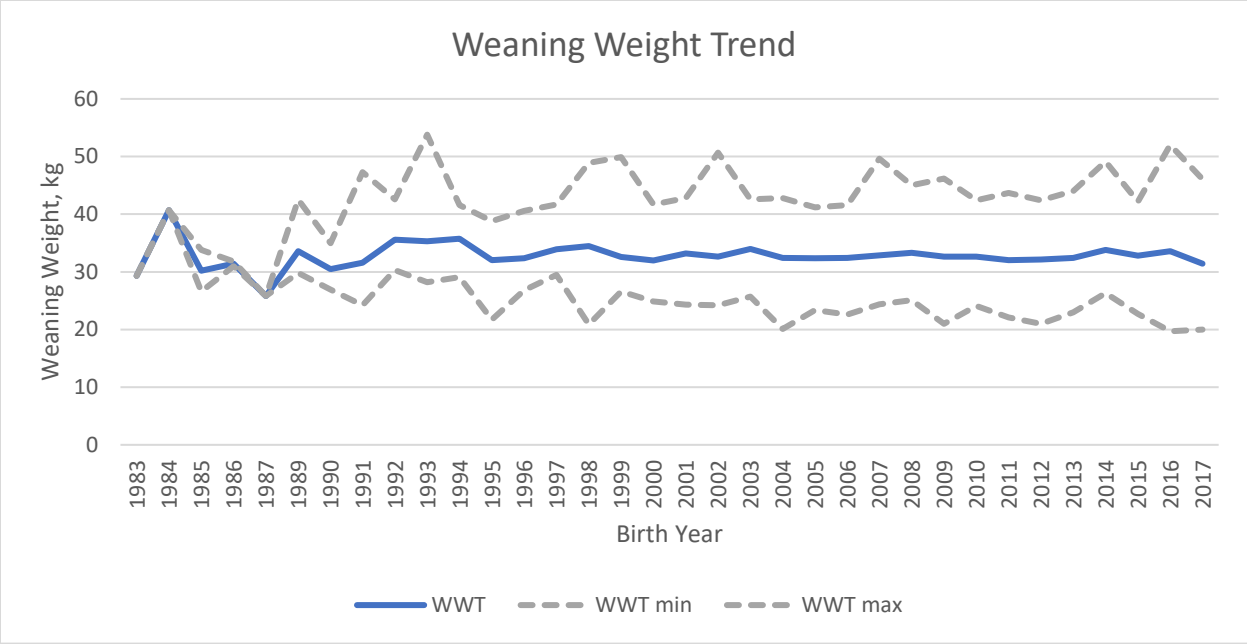

Figure 3.8 Suffolk phenotypic weaning weight trend by birth year

For the 244 genotyped sheep, H_E was 0.318 and H_O was 0.308. Overall, F_{IS} was 0.03 with a range of -0.16 to 0.23. MAF categories were summarized in Table 3.6. Of the MAF categories, 7.2 percent were fixed (MAF = 0) and 3.4 percent were present in the rare category (< 0.01). Most of the SNP were in the moderate and high categories. The ROH were summarized by class size (Mbps) and percentage (Table 3.7), and by chromosome (Table 3.8). Chromosome 1 contained the most total runs (12.2%) and chromosome 20 had the fewest (1.7%). The majority of ROH were short runs of less than 5 Mbps (95.9%) with only 0.3 percent of the runs in the 10 to 20 Mbps category. High recombination rates have been reported in sheep relative to other livestock species, making long runs of ROH less likely to persist in the population (Fröhlich et al., 2015). ROH per animal ranged from 38 to 252 with an average of 114.9. The current N_e was computed as 58.3. Model-based population structure determined there were 14 subpopulations within the genotyped Suffolk population. A visual representation of the proportional assignment of each animal (column) is represented in Figure 3.9.

Table 3.6 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for the Suffolk population

MAF Category	% of SNP
Fixed (0)	7.2
Rare (< 0.01)	3.4
Moderate (0.01 – 0.3)	55.2
High (0.3 - 0.5)	34.3

Table 3.7 Runs of homozygosity (ROH) by size class and total ROH percentage for the Suffolk population

ROH Class Category (Mbps)	No. ROH	Percent Total ROH
1 - 5	26,770	95.9
5 - 10	1,072	3.8
10 - 20	71	0.3

Table 3.8 Runs of homozygosity (ROH) count and percentage by chromosome for the Suffolk population

Chromosome	ROH Count	ROH Percentage
1	3,402	12.2
2	2,906	10.4
3	2,705	9.7
4	1,462	5.2
5	991	3.6
6	1,383	5.0
7	973	3.5
8	943	3.4
9	1,034	3.7
10	1,105	4.0
11	735	2.6
12	807	2.9
13	972	3.5
14	753	2.7
15	857	3.1
16	960	3.4
17	651	2.3
18	811	2.9
19	658	2.4
20	462	1.7
21	532	1.9
22	585	2.1
23	577	2.1
24	492	1.8
25	595	2.1
26	562	2.0

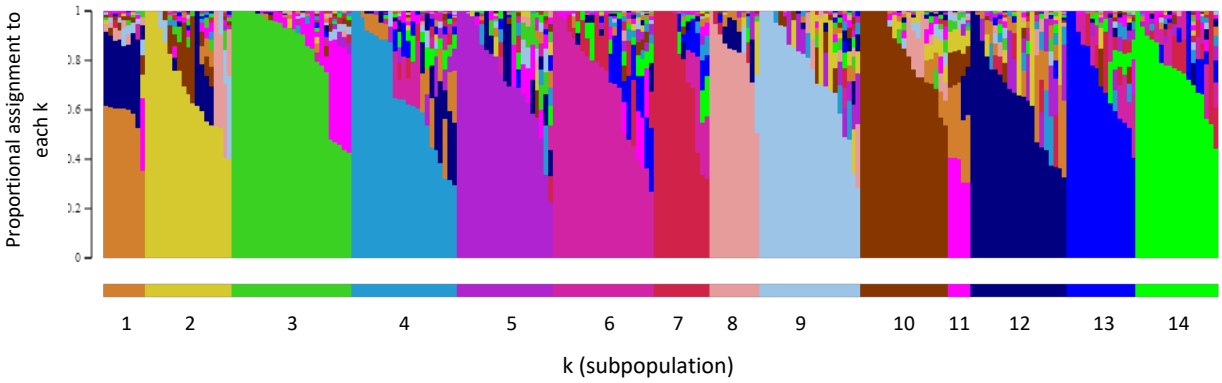

Figure 3.9 Model-based population structure of Suffolk (n = 244), where the proportional assignment of each animal was represented as a column and the animals are sorted by highest proportional assignment to a subpopulation

Optimal contribution selection. The OCS optimization converged after 17 iterations. Animals with the highest optimal contribution values were selected for the highest 100 (OCS 100), 50 (OCS 50), and 25 (OCS 25) animals. From the 244 selection candidates, the OCS 100 values ranged from 0.0007 to 0.0362, OCS 50 ranged from 0.0074 to 0.0362, and OCS 25 ranged from 0.0128 to 0.0362.

Genetic conservation index. Animals and their corresponding GCI index were extracted from the ENDOG results. Animals were ranked by index and subset into the highest 100 (GCI 100), 50 (GCI 50), and 25 (GCI 25). Because this sampling procedure is a ranking by index, animals present in GCI 50 were also present in GCI 100; similarly, animals present in GCI 25 were also present in GCI 100 and GCI 50. The GCI 100 ranged from 4.1 to 16.2, GCI 50 ranged from 6.2 to 16.2, and GCI 25 values ranged from7.9 to 16.2; in comparison, the lowest GCI value in the population was 1.0.

Simmental cattle. Genotyped animals in the final dataset included 949 males and 4,664 females. The mean pedigree-based inbreeding coefficient was 0.024 with a range from 0 to 0.375. The rate of inbreeding was computed as 0.0007 per year, or 0.0032 per generation, which is less than the critical

value of 1 percent per generation as defined by the FAO (1998). Inbreeding trend by birth year is shown in Figure 3.10. Inbreeding remained low but steadily increased after 1972. Prior to that time, unknown parentage information is reflected in very low levels of inbreeding. Few weaning weights were reported before 1974 with a total of 17,757 weaning weights available for the 84 percent and greater Simmental population (n = 34,462). Weaning weight EPD, API, and phenotypic weaning weight trend by birth year are shown in Figures 3.11, 3.12, and 3.13, respectively. Weaning weight EPD and API increased over time while phenotypic weaning weight remained flat. A linear regression of each trait on birth year was calculated. The estimated slope for weaning weight EPD was 0.629 (SE = 0.018, $P < 0.001$). The estimated slope corresponding to API was 0.698 (SE = 0.029, $P < 0.001$). The estimated slope for phenotypic weaning weight was 0.553 (SE = 0.124, $P < 0.001$).

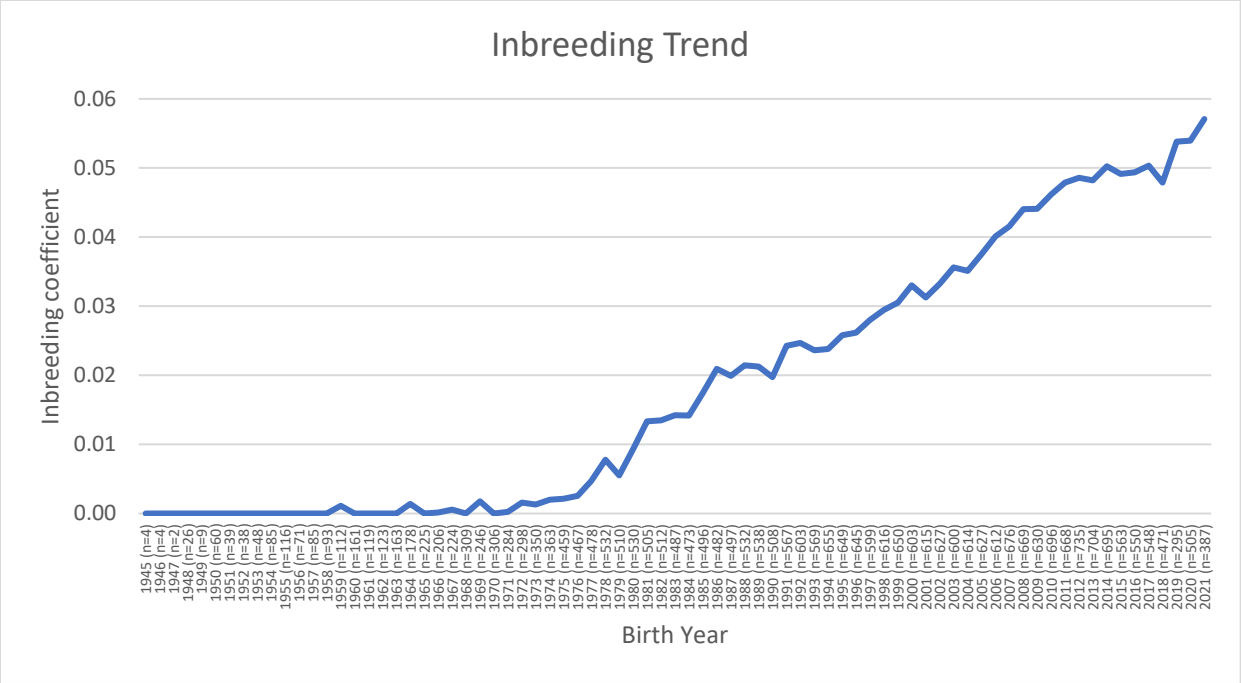

Figure 3.10 Simmental inbreeding trend by birth year

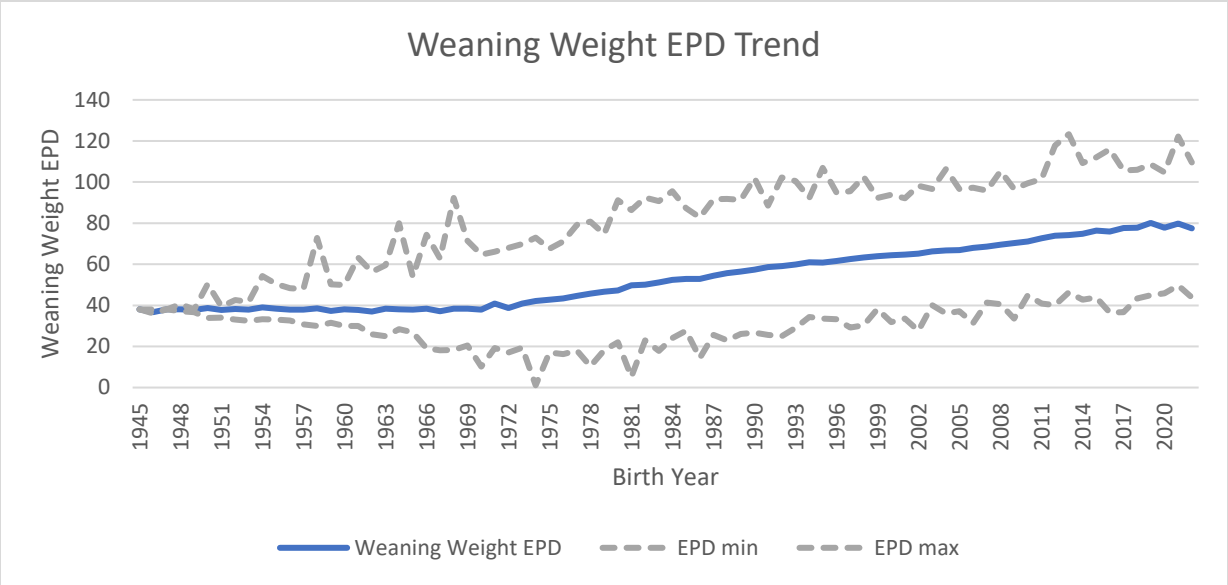

Figure 3.11 Simmental weaning weight expected progeny differences (EPD) trend

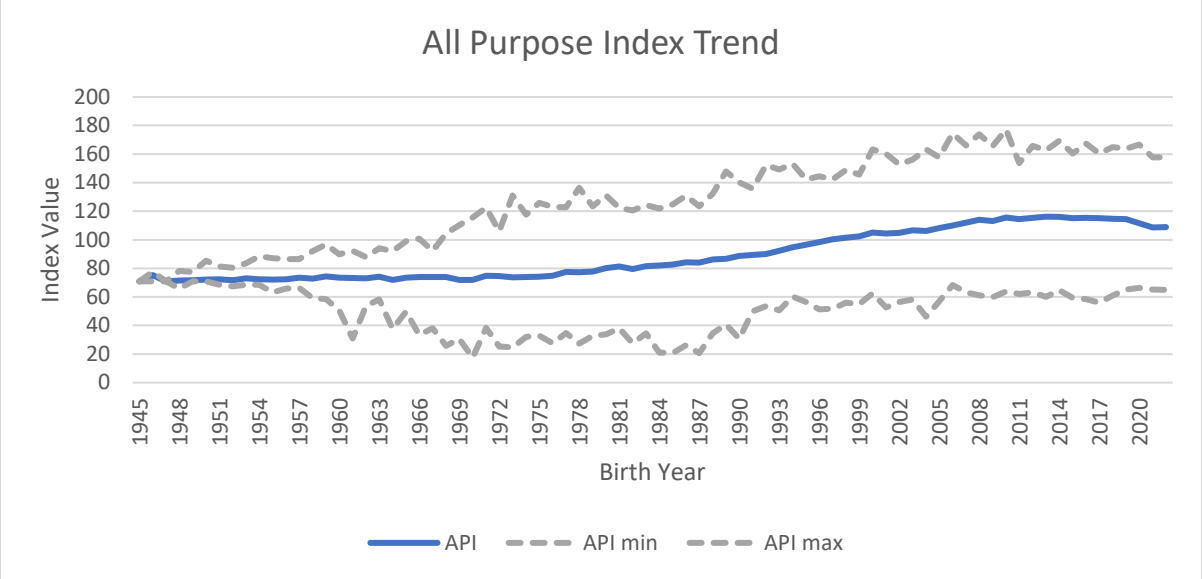

Figure 3.12 Simmental All Purpose Index (API) trend

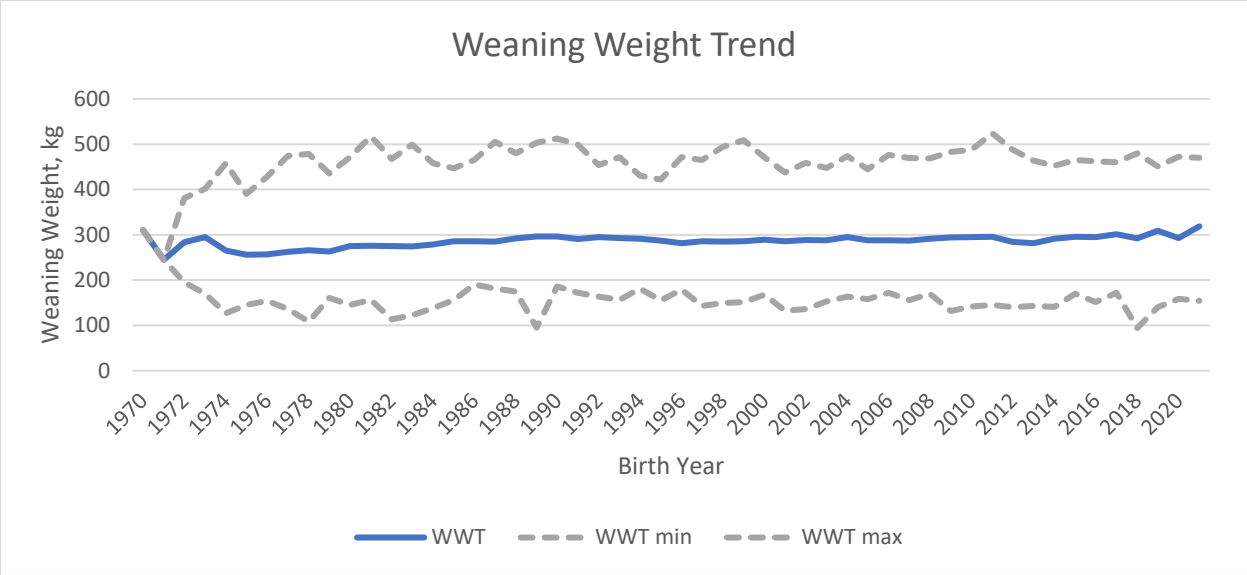

Figure 3.13 Simmental weaning weight (kg) trend

For the 5,613 genotyped Simmental, the overall H_E was 0.416 and H_O was 0.408. Overall, F_{IS} was 0.02 and ranged from -0.72 to 0.41. Table 3.9 summarizes the MAF categories. Few SNP were in the fixed (0.05 percent) and rare (0.56 percent) categories with the vast majority in the high category. The ROH were summarized by class size (Mbps) and percentage (Table 3.10) and by chromosome (Table 3.11). Chromosome 5 had the most total runs (8.4%) and chromosome 27 had the fewest (0.8%). Only 1.1 percent of the ROH were in the 10 to 20 Mbps category with 30.5 percent in the 5 to 10 Mbps category. The majority of the ROH were short runs in the 1 to 5 category (68.4%). The ROH per animal averaged 5.9 with a range from 0 to 44. Current N_e was computed as 153.8. Fifteen subpopulations were determined from the model-based population structure within the genotyped Simmental population. A visual representation of the proportional assignment of each animal (column) is represented in Figure 3.14.

Table 3.9 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for the Simmental population

MAF Category	% of SNP
Fixed (0)	0.05
Rare (< 0.01)	0.56
Moderate (0.01 – 0.3)	34.06
High (0.3 - 0.5)	65.33

Table 3.10 Runs of homozygosity (ROH) by size class and total ROH percentage for the Simmental population

ROH Class Category (Mbps)	No. ROH	Percent Total ROH
1 - 5	21,253	68.4
5 - 10	9,475	30.5
10 - 20	353	1.1

Table 3.11 Runs of homozygosity (ROH) count and percentage by chromosome for the Simmental population

Chromosome	ROH Count	ROH Percentage
1	1,951	6.3
2	2,239	7.2
3	1,863	6.0
4	746	2.4
5	2,603	8.4
6	1,486	4.8
7	1,698	5.5
8	1,197	3.9
9	1,006	3.2
10	2,083	6.7
11	576	1.9
12	816	2.6
13	1,207	3.9
14	1,315	4.2
15	538	1.7
16	1,162	3.7
17	682	2.2
18	525	1.7
19	1,338	4.3
20	1,007	3.2
21	400	1.3
22	388	1.2
23	794	2.6
24	778	2.5
25	524	1.7
26	735	2.4
27	262	0.8
28	708	2.3
29	454	1.5

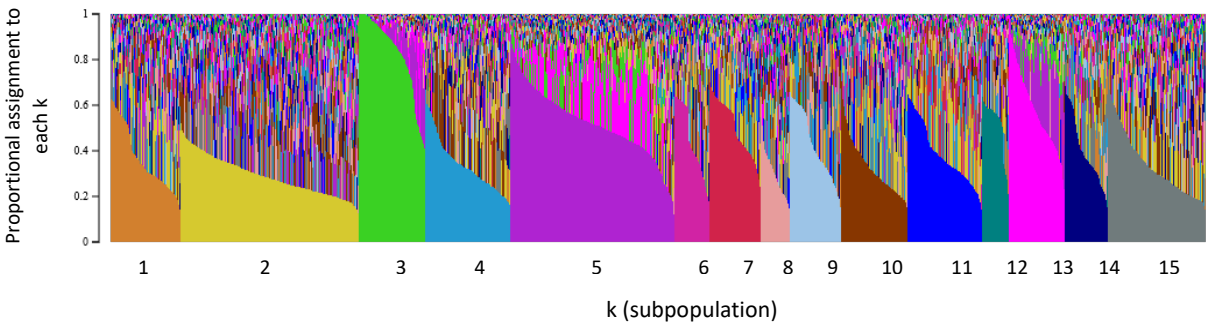

Figure 3.14 Model-based population structure of Simmental (n = 5,613), where the proportional assignment of each animal was represented as a column and the animals are sorted by highest proportional assignment to a subpopulation

Optimal contribution selection. The OCS optimization converged after 24 iterations. From the 5,613 selection candidates, the animals with the highest 100, 50, and 25 optimal contribution values were selected. The optimal contribution values ranged from 0.0000002 to 0.037 for OCS 100, 0.008 to 0.037 for OCS 50, and 0.015 to 0.037 for OCS 25. The low optimal contribution for most animals suggests there are few unique animals based on pedigree that are necessary to select.

Genetic conservation index. Selection candidates were sorted by GCI and the top 100, 50, and 25 animals were selected. The GCI 100 ranged from 182.6 to 323.0, GCI 50 ranged from 208.6 to 323.4, and GCI 25 values ranged from 232.8 to 323.4. The lowest GCI value among the selection candidates was 1.1.

3.4 Discussion

Data simulation. Simulation of the sheep population created the underlying genome for the entire population, which is the primary advantage of simulated over real populations. Recommended reporting for simulated populations included the size of the genome, number of markers, number of

QTL, distribution of QTL effects, simulation of genetic values, chosen heritability, heterozygosity, and number of generations of mating (Daetwyler et al., 2013). In addition to reporting these values, the N_e, MAF categories, ROH, inbreeding levels, relationships among animals, and model-based population structure were evaluated.

The N_e was simulated to be 250. Post hoc evaluation of the final year of selection estimated molecular-based N_e as 289.8. Phenotypic weaning weight was initially set at 29 kg for the randomly mating population. After 30 years of random mating, the mean weaning weight was 29.2 kg. Phenotypic selection for 15 years for weaning weight resulted in a positive trend for both weaning weight and weaning weight EBV. The number of fixed alleles increased over time, as expected in a population undergoing selection. MAF categories show how the allele frequencies of the population were distributed, and the simulated population had 14.2 percent classified as fixed and 9.9 percent of alleles classified as rare. Other sheep breeds were reported to have fixed alleles ranging from 3.9 percent for Merinos to 23.3 percent for Border Leicester (Kijas et al., 2014). The fixed and rare allele percentages for the simulated population were higher than Corriedale and Merino but lower than Creole sheep reported by Grasso et al. (2014). For a simulated dataset using similar parameters to this simulation, Vargas Jurado et al. (2021), removed 26 percent of the markers because they had a MAF < 0.01. A high number of fixed and rare alleles are typical in simulated datasets and are typically managed by removing these markers as a quality control step. The low frequency alleles were retained here to not to be more representative of a real breed, but to increase the challenge of capturing the alleles in a population. For the ADMIXTURE analysis, there were high levels of admixture of subpopulations for most animals; the highest proportional assignment of an animal to any subpopulation was 0.89 (Figure 3.3). While the goal was not to model a specific sheep breed, it was to generate a population structure that could be defined and to evaluate the ability of sampling procedures to capture that population structure.

There were no animals in common for OCS and GCI sampling strategies. While the methods for selecting animals differ between the two strategies, at least some genetically important animals in common would be expected to be identified by both methods.

Suffolk sheep. The subset of the Suffolk population analyzed in this study was characterized by low levels of inbreeding and high levels of polymorphic alleles. Weaning weight EBV and phenotypic weaning weight show minimal improvement over time for this breed while the Carcass Plus Index has seen more improvement. Since a high level of selection leads to decreased genetic variation, the lack of intense selection for the measured traits may have contributed to the high level of genetic diversity present in the population.

Heterozygosity measures, H_E and H_O, were 0.318 and 0.308, respectively. The percentage of fixed and rare alleles were 7.2 and 3.4, respectively. High heterozygosity and low fixed alleles suggest the population has plenty of diversity. It also makes sampling the population to retain all alleles less challenging. ROH were dominated by short ROH, which are attributed to ancient inbreeding or founder breed effects and influenced by high recombination rates observed in sheep (Fröhlich et al., 2015). Since they have been maintained in the population throughout time, it is important to make sure the sampled populations reflect a similar pattern of ROH. Additionally, a lack of long ROH suggest no major recent inbreeding events have occurred (Purfield et al., 2012).

N_e was computed as 58.3. Given a range of N_e of 50 to 100 for a population to be considered stable, the Suffolk population has sufficient genetic diversity (FAO, 1998; Meuwissen, 2009). Wilson et al. (2022) reported an N_e of 79.5 using a larger Suffolk population from which this data is a subset; this also demonstrates the importance of selecting animals for research purposes to minimize biased results. Caution should be exercised when using relatively few genotyped animals as representative of an entire breed. Fourteen subpopulations or ancestral populations were identified in the model-based population

structure analysis; Wilson et al. (2022) identified 7 subpopulations. The presence of distinct subpopulations within the breed maintains different allelic combinations across the breed. Effective sampling of the breed will need to capture each of these subpopulations. There were 36 animals in common for OCS 100 and GCI 100 sampling strategies.

***Simmental cattle*.** The Simmental population included in this analysis had a low overall inbreeding level, with an average inbreeding coefficient of 0.05 in the most recent generation. Inbreeding began to increase in the 1970s when more parents were known and steadily increased after that point. The sires of this population had a range from 1 to 349 offspring and the dams had a range from 1 to 24. The upper end of these numbers indicates the use of artificial reproductive techniques, such as artificial insemination and embryo transfer. Extensive use of such techniques can lead to fewer sires and dams, a narrowing of the genetic base, and decreased genetic diversity; this is not yet indicated in this population (Funk, 2006; Melka et al., 2013). Weaning weight EPD has increased steadily over time, doubling during the time period evaluated. Weaning weight EPD in early years had a narrow range of values and increased over time. The API trend has increased over time with a wider range of values in recent years. Phenotypic weaning weight has remained stable over time even as weaning weight EPD has increased.

Heterozygosity in this population was high for both H_E and H_O, with values of 0.416 and 0.408, respectively. High levels of heterozygosity were observed in other Simmental populations (Curik et al., 2010; Kelleher et al., 2017). F_{IS} was 0.02, which closely matches the pedigree inbreeding value of 0.024. For the MAF categories, 0.05 and 0.56 percent of the alleles were fixed and rare, respectively. Because the SNP dataset included markers from multiple SNP chips, the overlapping markers across chips may be those that are highly polymorphic and, therefore, chosen with the potential to be more informative than less polymorphic SNP. The moderate category included 34.1 percent of the SNP while the high category

included 65.3 percent. These high frequency alleles in the population make it more likely to capture the genetic diversity in the population when compared to those with many low frequency alleles. The majority of ROH were in the 1 - 5 category (68.4%) with an additional 30.5 percent in the 5 - 10 category. Since shorter runs are generally associated with ancient inbreeding, this population has a mix of ancient and more recent inbreeding. This increase in recent inbreeding indicated by ROH aligns with the inbreeding trend that is linearly increasing in recent generations.

N_e was reported as 153.8. This value is well above the minimum of 50 recommended by the FAO (1998) and 100 recommended by Meuwissen (2009). Based on the results from this research, the more conservative estimate of 100 is recommended moving forward. Including Simmental above 84 percent rather than strictly limiting to fullbloods likely contributes to added genetic diversity of the population and prevents a greater loss of N_e. The inclusion of other breeds in the Simmental registry may also explain why there are many subpopulations within the breed. For practical purposes, the ADMIXTURE analysis was limited to 15 subpopulations, which all have much admixture within the subpopulations rather than being assigned to a single distinct population (Figure 3.14). Less than 1 percent of the animals (n = 55) had a proportional assignment above 0.99 to any subpopulation. High levels of heterozygosity coupled with admixture in the Simmental population are expected to be favorable for capturing the alleles present in the population. There were no animals in common for OCS and GCI sampling strategies.

3.5 Conclusion

A simulated population was developed to allow for a comparison of sampling strategies for identifying genetically important animals to capture the genetic diversity of the population. The two sampling strategies, OCS and GCI, use pedigree-based selection strategies and the assessment of the

success of capturing the genetic diversity available in the population was measured using both molecular and quantitative methods. The primary advantage of the simulated population was the ability to know the underlying genotypes for every animal in the population. A thorough assessment of the entire population was performed by evaluating a phenotypic trait, breeding values, inbreeding, and molecular measures of heterozygosity, minor allele frequency categories, effective population size, and population substructure. Then, the sampling strategies were applied, with 100, 50, and 25 animals sampled from the respective strategies. The simulated population represented a large purebred sheep population with a moderate number of markers.

Once the simulated population was sampled, the real populations were evaluated and included a sheep population and a beef population. The Suffolk sheep population was small but included a large number of markers to capture the allelic diversity from. The Simmental cattle population was of moderate size with an admixed population and a moderate number of markers to capture.

The simulated population had low levels of inbreeding and a low rate of inbreeding. This was similar for the Suffolk and Simmental populations. The simulated population had the lowest levels of heterozygosity of the three populations, making capturing the genetic diversity in the simulated population the most challenging. Similarly, the simulated population had the highest percentage of fixed and rare alleles of the three populations. Capturing rare alleles are the most challenging and comprised 9.9 percent of the simulated population and only 3.4 and 0.6 percent of the Suffolk and Simmental populations, respectively. The simulated population had the majority of ROH as short runs (< 5 megabase pairs) as was the case for the Suffolk and Simmental populations. In the model-based population structure analysis, the simulated population had many subpopulations. The Suffolk population had much more distinct subpopulations while the Simmental subpopulations were similar to the simulated population where the subpopulations were much less defined.

Sampling of the simulated population resulted in OCS samples with low levels of contributions with the highest contribution of 0.008. The highest contributions were 0.036 and 0.037 for Suffolk and Simmental, respectively. Since GCI is based on the number of founders, direct comparisons across populations cannot be made. Only the Suffolk population had selected animals in common between OCS 100 and GCI 100. This may have been due to the small population size since no animals were in common for the larger populations. The OCS sampling strategy places emphasis on minimizing the kinship of selected animals while the GCI sampling strategy maximizes the representation of founder animals in the selected animals. Because these selection strategies differ in their purpose, it perhaps should not be surprising that there are not more overlapping selected animals between the two strategies. However, intuitively, the most genetically important animals selected to represent a population would be expected to be the same. The simulated population provided a structure for testing the two sampling strategies, which were then applied to the two real populations.

CHAPTER IV

ASSESS THE POPULATION STRUCTURE FOR A SIMULATED BREED, A SHEEP BREED (SUFFOLK), AND A BEEF BREED (SIMMENTAL)

4.1 Introduction

Population structure is defined by the amount and distribution of genetic variation in a population. Changes to population structure occur through random genetic drift, selection, mutation, and migration, all of which drive the amount of genetic variation within the population (Eding and Laval, 1999). Various molecular tools and measures that help to define population structure include observed (H_O) and expected heterozygosity (H_E), minor allele frequencies (MAF), runs of homozygosity (ROH), and parametric subpopulation structure. For quantitative measures, the variation of measured traits and trends over time can be examined.

The effective population size, N_e, is defined as the number of individuals that would give rise to the same rate of inbreeding if they bred in the manner of the idealized population. Another definition is the size of an "ideal" population of animals that would have the same decrease in genetic diversity due to genetic drift as the real population of interest (Wright, 1931; Halliburton and Halliburton, 2004). The United Nations Food and Agriculture Organization (FAO) recommends breed conservation efforts maintain a minimum N_e of 50, which is expected to result in an inbreeding rate of one percent per generation while Meuwissen (2009) recommends a more conservative N_e of 100.

Genetic diversity can be measured as pedigree-based or marker-based. Pedigree-based diversity is a measure of the probability of an allele being identical by descent (IBD) from the founder population, which is assumed unrelated. Marker-based diversity does not assume a founder population and considers all markers that are identical by state (IBS) to be IBD. This research uses a pedigree-based

sampling strategy to identify animals and then uses marker-based population assessments to evaluate sampling success.

This research has the objective of capturing all allelic genetic diversity in a population; however, there is not strict agreement among conservationists about what capturing the genetic diversity of a population means. Eding and Laval (1999) argue that overall diversity is more important than capturing specific alleles and an emphasis on diversity of genotypes was more important for polygenic traits than the alleles that make up those traits. The objective of this research is to capture the genetic diversity of the population with as few animals as possible and determine success by assessing the population structure of the selected animals compared to the full population.

4.2 Materials and Methods

Optimal contribution selection (OCS) and Genetic Conservation Index (GCI). For the simulated population followed by the real populations of Suffolk sheep and Simmental cattle, sampling strategies included optimal contribution selection (OCS) (Meuwissen, 2009) and the Genetic Conservation Index (GCI) (Alderson, 1992). In the respective sampling strategies, 100, 50, and 25 animals were selected. The six samples are reported as OCS 100, OCS 50, OCS 25, GCI 100, GCI 50, and GCI 25, referring to the sampling strategy and the number of animals sampled.

The optiSel package in R was used to select 100, 50, and 25 animals using OCS (Wellmann, 2019). OCS typically is used to select animals to minimize the average kinship for selected animals while maximizing genetic gain. Computationally, OCS maximizes $G = c'EBV$, while restricting: $\bar{A}_p = c'Ac$, and restricting: $c'Q = 0.5$, where G was the average genetic level of the parents weighted by their number of offspring, EBV was a vector of estimated breeding values of the parents, c was a vector of genetic contributions of the parents (fraction of offspring each parent obtains, scaled such that it sums to 0.5 for

males and to 0.5 for females; c_i=0 implies that animal i was not selected), $\bar{A}_p$ was the desired value of the average genetic relationships between the parents where A is the numerator relationship matrix of the selection candidates, and Q was a design matrix indicating the gender of the candidates (first column is 1 for male, 0 for females; second column is 1 for female, 0 for males) (Meuwissen, 2009). Since maximizing genetic gain was not the goal of the sampling strategy, only minimizing kinship was considered here.

The ENDOG package was used to compute GCI (Gutiérrez and Goyache, 2005), which seeks to maintain equal contributions from all the founder ancestors in each animal. Computationally, the GCI for an animal is the effective number of founders in its pedigree from $1/\Sigma P_i^2$, where P_i is the proportion of genes of founder animal *i* in the pedigree (Alderson, 1992). Animals can be ranked based on GCI, with a higher GCI value representing a higher level of genetic diversity.

Simulated population. Weaning weight means, ranges, standard deviations, and coefficients of variation from the OCS and GCI samples were compared to the year 15 population (n = 13,662). Weaning weight EBV statistics were also compared. Differences between the means of weaning weight for year 15 and the OCS and GCI samples were analyzed using a one-way ANOVA test. Pairwise comparisons of the means were performed using the Tukey-Kramer procedure. Similar analyses were conducted for weaning weight EBV.

The SNP for each animal were analyzed using PLINK (Purcell et al., 2007). Expected (H_E) and observed heterozygosity (H_O) were computed as well as molecular inbreeding, F_{IS}. Pairwise comparisons of the means were performed using the Tukey-Kramer test for H_E, H_O, and F_{IS}. The minor allele frequency (MAF) categories were determined as fixed (MAF = 0), rare (MAF < 0.01), moderate (MAF 0.01 - 0.3), and high (MAF 0.3 - 0.5) (Grasso et al., 2014; Wilson et al., 2022). A Chi-square Goodness of Fit test was performed to determine whether the proportions in the MAF categories were equal between year 15

and each OCS and GCI sampled population. Runs of homozygosity (ROH) were computed using the detectRUNS package in R (Biscarini et al., 2019) with a sliding window of 50 SNP with a minimum length of 1,000 kb, a minimum of 30 SNP, allowing for 1 missing SNP, and allowing a maximum of 1 possible heterozygous SNP. The ROH class categories were determined as 1 - 5, 5 - 10, 10 - 20, 20 - 40, and > 40 Megabase pairs. A Chi-square Goodness of Fit test was performed to determine whether the proportions in the ROH categories were equal between year 15 and each OCS and GCI sampled population. Recent N_e was computed for each sampled population using the linkage disequilibrium method of Waples and Do (2008) as implemented in NeEstimator v2.1 (Do et al., 2014) for comparison with the full population. The alleles not captured from each sampling strategy were computed from the results of this analysis based on the entire year 15 population and when only alleles with a MAF > 0.01 were considered. The allelic frequencies of the missing alleles were computed. Model-based population structure was compared for the full population and the sampled animals.

Suffolk sheep. Phenotypic weaning weight means, ranges, standard deviations, and coefficients of variation from the OCS and GCI samples were compared to the entire genotyped Suffolk population (n = 244). Weaning weight EBV and Carcass Plus Index statistics were also compared. A one-way ANOVA was used to compare differences between the means for the full population and each OCS or GCI population, followed pairwise comparisons of the means using the Tukey-Kramer procedure for weaning weight, weaning weight EBV, and Carcass Index Plus.

The H_E, H_O, F_{IS}, and MAF for the sampled populations were computed using PLINK (Purcell et al., 2007). Pairwise comparisons of the means for H_E, H_O, and F_{IS} were performed using the Tukey-Kramer test. The MAF for each marker were compared to the full population for each sample method to compute the missing alleles. The allele frequencies of the missing alleles were computed. An additional count of missing alleles was performed, considering only alleles with a MAF > 0.01. The MAF categories

were compared for the full population and each sampling strategy using A Chi-square Goodness of Fit test. The ROH classes were compared for the full population and each sampling strategy using A Chi-square Goodness of Fit test. Recent N_e was compared for the full population and each sampled population. Model-based population structure was compared for the full population and the sampled animals.

Simmental cattle. Means, ranges, standard deviations, and coefficients of variation from the OCS and GCI samples were compared to the entire genotyped Simmental population (n = 5,613) for phenotypic weaning weight, weaning weight EPD and the All-Purpose Index. For each trait, a one-way ANOVA was used to compare the means for the full population and each OCS or GCI population. Then, the Tukey-Kramer test was used to perform pairwise comparisons of the means for the full population and each OCS or GCI population.

The H_E, H_O, F_{IS}, and MAF for the sampled populations were computed using PLINK (Purcell et al., 2007). Pairwise comparisons of the means for H_E, H_O, and F_{IS} were performed using the Tukey-Kramer test. Missing alleles from the sampled populations were computed using MAF for each marker when compared to the full population. The allele frequencies of the missing alleles were computed for all markers and when including only alleles with a MAF > 0.01. Assignment to MAF categories was compared for the full population and each sampling strategy. A Chi-square Goodness of Fit test was performed to determine whether the proportions in the MAF categories were equal between the full population and each OCS and GCI sampled population. A Chi-square Goodness of Fit test was performed to determine whether the proportions in the ROH categories were equal between the full population and each OCS and GCI sampled population. Recent N_e was computed for each sampled population and compared to the full population. Model-based population structure was compared for the full population and the sampled animals.

4.3 Results

Simulated population. A comparison of the minimum, mean, maximum, standard deviation, and coefficient of variation for phenotypic weaning weight and weaning weight EBV for the full year 15 population and OCS 100, OCS 50, and OCS 25 sampled populations was presented in Table 4.1. Based on both ANOVA and the Tukey-Kramer tests, mean weaning weight EBV in the year 15 population differed from that in each OCS population ($P < 0.001$). Phenotypic weaning weight in year 15 differed from that in OCS 100 ($P < 0.001$), but not OCS 50 ($P = 0.09$), and OCS 25 ($P = 0.07$). The full year 15 population consisted of 6,817 males and 6,845 females. OCS 100 selected 59 rams and 41 ewes, OCS 50 selected 26 rams and 24 ewes, and OCS 25 selected 14 rams and 11 ewes. The H_E was 0.259 for generation 15 and for all three sampled populations. The H_O was 0.259 for generation 15 and was 0.260, 0.259, and 0.261 for OCS 100, OCS 50, and OCS 25, respectively. Molecular inbreeding, measured as F_{IS}, averaged 0 for all sampled and the year 15 population. Based on the ANOVA and Tukey-Kramer tests, the means for H_E, H_O, and F_{IS} did not differ between populations, indicating the sampled populations represented the full population for these measures (Table 4.2).

Table 4.1 Minimum (Min), mean, maximum (Max), standard deviation (St Dev), and coefficient of variation (CV %) for weaning weight (WWT) estimated breeding values (EBV) and phenotypic weaning weight (WWT) for the year 15 population and the three optimal contribution selection (OCS) sampled populations

	Year 15	OCS 100	OCS 50	OCS 25
Min WWT EBV	28.1	28.4	28.4	28.4
Mean WWT EBV	31.0[a]	30.1[b]	30.0[b]	30.0[b]
Max WWT EBV	33.5	32.0	32.0	32.0
Std Dev WWT EBV	0.8	0.7	0.7	0.8
CV % WWT EBV	2.5	2.2	2.3	2.6
Min WWT	20.2	21.5	25.5	25.5
Mean WWT	31.0[a]	30.0[b]	30.1[ab]	30.0[ab]
Max WWT	41.8	34.6	33.8	33.7
Std Dev WWT	2.5	2.4	2.2	2.2
CV % WWT	8.1	8.0	7.4	7.4

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

Table 4.2 Average expected heterozygosity (H_E), observed heterozygosity (H_O), and molecular inbreeding (F_{IS}) for the year 15 population and the three optimal contribution selection (OCS) sampled populations

	Year 15	OCS 100	OCS 50	OCS 25
H_E	0.259[a]	0.259[a]	0.259[a]	0.259[a]
H_O	0.259[a]	0.260[a]	0.259[a]	0.261[a]
F_{IS}	-0.001[a]	-0.006[a]	-0.002[a]	-0.006[a]

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

The aim of the sampling is to capture the full range of alleles in the population. Table 4.3 shows the percentage of year 15 alleles that were captured by each sampling method and the frequency of the alleles that were not captured. When alleles with a MAF < 0.01 were excluded, 99.6, 98.5, and 96.7 percent of alleles were captured by OCS 100, OCS 50, and OCS 25, respectively. Table 4.4 shows the MAF categories for year 15 and for the OCS samples. All sampling methods had a higher percentage of fixed alleles than the full population, indicating a loss of variation. The percentage of fixed alleles increased as the number of animals sampled decreased. The percentage of alleles in the moderate and high

categories is similar across sampling methods, with the rare category being the most impacted. Based on Chi-square analyses, the proportions differed by MAF category when the full population was compared to each OCS sampled population ($P < 0.001$).

Table 4.3 Percentage of year 15 alleles captured and frequency of missing alleles using optimal contribution selection (OCS) sampled populations

		Frequency of Missing Alleles		
Selection Category	% Alleles Captured	Minimum	Average	Maximum
OCS 100	96.4	0.00004	0.0035	0.0497
OCS 50	94.7	0.00004	0.0073	0.1302
OCS 25	92.5	0.00004	0.0135	0.1945

Table 4.4 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for year 15 and the optimal contribution selection (OCS) sampled populations

MAF Category: % of SNP	Year 15	OCS 100	OCS 50	OCS 25
Fixed (0)	14.2	20.8	24.0	28.2
Rare (< 0.01)	9.9	2.7	0.0	0.0
Moderate (0.01 – 0.3)	52.1	52.5	52.0	47.3
High (0.3 - 0.5)	23.8	24.0	24.0	24.5

For year 15 and the sampled populations, the shortest ROH class has the highest percentage of runs (Table 4.5). There are few runs in the longest ROH classes. Short runs are generally attributed to ancient inbreeding while long runs are due to more recent inbreeding. Since the ancient inbreeding has persisted over time, it is considered less of a concern than more recent inbreeding. Based on Chi-square analyses, the proportions of total ROH in each class size did not differ for year 15 and each sampled population ($P > 0.05$). The average number of ROH per animal was also similar between year 15 and the sampled populations (Table 4.6). Recent N_e was computed as 289.8 for the full population and the OCS sampled populations were 88.5, 52.2, and 23.6 for OCS 100, OCS 50, and OCS 25, respectively.

Table 4.5 Runs of homozygosity (ROH) by size class and total ROH percentage for year 15 and the optimal contribution selection (OCS) sampled populations

ROH Class Category (Mbps)	Percent Total ROH in each Class Size			
	Year 15	OCS 100	OCS 50	OCS 25
1 – 5	79.0	79.6	79.5	79.3
5 – 10	16.4	15.7	15.8	16.2
10 – 20	3.9	3.9	3.9	3.8
20 – 40	0.7	0.8	0.8	0.7
> 40	0.1	0.1	0.0	0.0

Table 4.6 Average number of runs of homozygosity (ROH) per animal for year 15 and the optimal contribution selection (OCS) sampled populations

	Average ROH
Year 15	107.6
OCS 100	106.5
OCS 50	107.7
OCS 25	107.0

Model-based population structure was compared for the full genotyped population and each OCS sampled population. All 15 subpopulations identified in the model-based population structure analysis were represented in each of the OCS sampled populations, but each subpopulation becomes less distinct as the sampling progresses from the full population to OCS 100, OCS 50, and OCS 25. Figure 4.1 shows the proportional assignment of each animal for the full population (Figure 4.1(a)) and each OCS sampled population (Figure 4.1(b-d)).

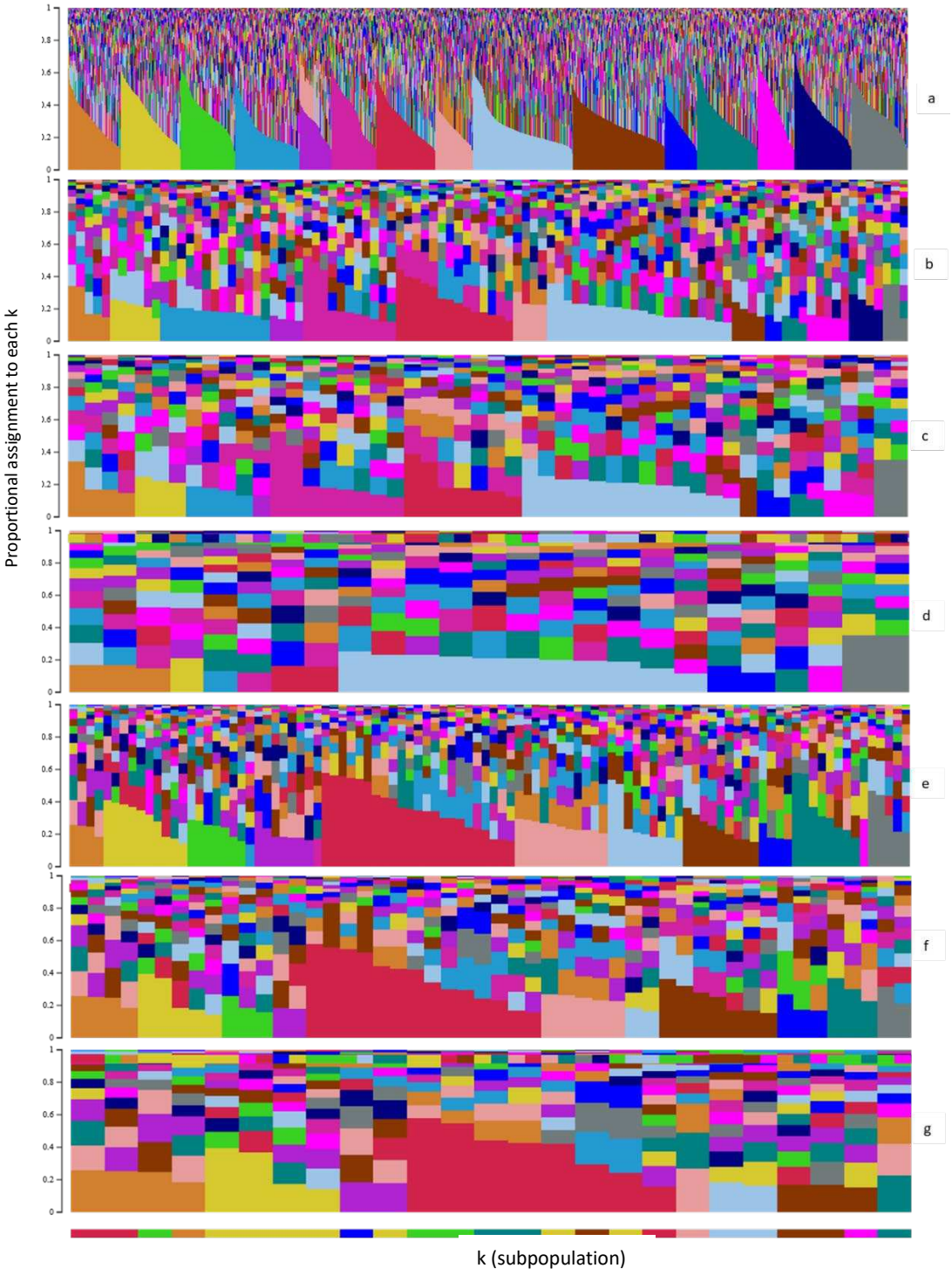

Figure 4.1 Model-based population structure for the simulated population for optimal contribution (OCS) and Genetic Conservation Index (GCI) strategies for the full population (a), OCS 100 (b), OCS 50 (c), OCS 25 (d), GCI 100 (e), GCI 50 (f), and GCI 25 (g) displaying the proportional assignment of each animal as a column and sorted by highest proportional assignment to a subpopulation

Phenotypic weaning weight and weaning weight EBV were compared for the full year 15 population and the GCI sampled populations in Table 4.7. Means differed between the year 15 and each GCI population for weaning weight EBV for both the one-way ANOVA and the Tukey-Kramer test ($P < 0.001$). For phenotypic weaning weight, the Tukey-Kramer test indicated different means for year 15 when compared to GCI 100 ($P = 0.01$), but not GCI 50 ($P = 0.39$) or GCI 25 ($P = 0.52$). The GCI 100 included 53 females and 47 males, GCI 50 included 28 females and 22 males, and GCI 25 included 15 females and 10 males. The H_E was 0.259 for year 15 and for all three GCI sampled populations. The H_O was 0.259 for year 15 and 0.259, 0.258, and 0.258 for GCI 100, GCI 50, and GCI 25, respectively. Molecular inbreeding (F_{IS}) averaged 0 for all sampled populations. Based on the ANOVA and Tukey-Kramer tests, the means for H_E, H_O, and F_{IS} did not differ between populations (Table 4.8).

Table 4.7 Minimum (Min), mean, maximum (Max), standard deviation (St Dev), and coefficient of variation (CV %) for weaning weight (WWT) estimated breeding values (EBV) and phenotypic weaning weight (WWT) for the year 15 population and the three Genetic Conservation Index (GCI) sampled populations

	Year 15	GCI 100	GCI 50	GCI 25
Min WWT EBV	28.1	30.1	30.2	30.5
Mean WWT EBV	31.0[a]	31.7[b]	31.8[b]	31.9[b]
Max WWT EBV	33.5	33.5	33.3	33.3
Std Dev WWT EBV	0.8	0.7	0.7	0.7
CV % WWT EBV	2.5	2.3	2.1	2.3
Min WWT	20.2	23.9	26.4	28.4
Mean WWT	31.0[a]	31.7[b]	31.5[ab]	31.7[ab]
Max WWT	41.8	38.2	34.4	34.4
Std Dev WWT	2.5	2.4	1.7	1.7
CV % WWT	8.1	7.7	5.5	5.4

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

Table 4.8 Average expected heterozygosity (H_E), observed heterozygosity (H_O), and molecular inbreeding (F_{IS}) for the year 15 population and the three Genetic Conservation Index (GCI) sampled populations

	Year 15	GCI 100	GCI 50	GCI 25
H_E	0.259[a]	0.259[a]	0.259[a]	0.259[a]
H_O	0.259[a]	0.259[a]	0.258[a]	0.258[a]
F_{IS}	-0.001[a]	0.001[a]	0.004[a]	0.003[a]

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

Table 4.9 shows the percentage of year 15 alleles that were captured by each GCI sampling method and the frequency of the alleles that were not captured. When alleles with a MAF < 0.01 were excluded, 99.3, 98.1, and 95.9 percent of the alleles were captured by GCI 100, GCI 50, and GCI 25, respectively. The MAF categories for year 15 and for the GCI samples are shown in Table 4.10. The percentage of fixed alleles increased as the number of sampled animals decreased. This indicates a loss of alleles in the sampled populations. The increase in fixed alleles corresponded to a decrease in the rare and moderate alleles categories. Based on Chi-square analyses, the proportions differed by MAF category when the full population was compared to each GCI sampled population ($P < 0.001$).

Table 4.9 Percentage of year 15 alleles captured and frequency of missing alleles from the Genetic Conservation Index (GCI) sampling strategies

		Frequency of Missing Alleles		
Selection Category	% Alleles Captured	Minimum	Average	Maximum
GCI 100	95.2	0.00004	0.0043	0.0665
GCI 50	93.6	0.00004	0.0085	0.0914
GCI 25	91.4	0.00004	0.0157	0.2145

Table 4.10 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for year 15 and the Genetic Conservation Index (GCI) sampled populations

MAF Category: % of SNP	Year 15	GCI 100	GCI 50	GCI 25
Fixed (0)	14.2	23.1	26.2	30.2
Rare (< 0.01)	9.9	2.0	0.0	0.0
Moderate (0.01 – 0.3)	52.1	51.1	49.8	45.3
High (0.3 - 0.5)	23.8	23.7	24.0	24.4

The percentage of ROH in each class size were comparable among the full population and each sampled population, as confirmed by Chi-square analyses ($P > 0.05$) (Table 4.11). The majority of ROH were in the 1 - 5 Mbps category with less than 1 percent in the 20 Mbps and higher classes. The number of ROH per animal were also comparable among the full population and the sampled populations (Table 4.12). Recent N_e for the GCI 100, GCI 50, and GCI 25 was 50.3, 31.3, and 17.4, respectively, while the full population had an N_e of 289.8.

Table 4.11 Runs of homozygosity (ROH) by size class and total ROH percentage for year 15 and the Genetic Conservation Index (GCI) samples

ROH Class Category (Mbps)	Percent Total ROH in each Class Size			
	Year 15	GCI 100	GCI 50	GCI 25
1 - 5	79.0	79.3	79.4	80.0
5 - 10	16.4	16.1	15.6	15.2
10 - 20	3.9	3.9	4.1	4.2
20 - 40	0.7	0.7	0.8	0.5
> 40	0.1	0.0	0.1	0.0

Table 4.12 Average number of runs of homozygosity (ROH) per animal for year 15 and the Genetic Conservation Index (GCI) samples

	Average ROH
Year 15	107.6
GCI 100	107.2
GCI 50	107.7
GCI 25	109.5

Model-based population structure was compared for the full genotyped population and each GCI sampled population. All 15 subpopulations identified by the model-based population structure analysis were represented in each of the GCI sampled populations. The proportional assignment of each animal for the full population and each GCI sampled population is represented in Figure 4.1. When

compared to the full population (Figure 4.1(a)), some of the GCI sampled populations had minimal representation (4, 6, and 13) while subpopulation 7 was well represented (Figure 4.1(e-g)).

Suffolk sheep. For the weaning weight metrics (EBV and phenotype) and for the Carcass Plus Index, the full genotyped population was compared to the animals selected in OCS 100, OCS 50, and OCS 25 to determine if the full range of variation was captured (Table 4.13). Based on both the ANOVA and the pairwise comparison of means using the Tukey-Kramer test, means did not differ for the full population and the OCS sampled populations ($P > 0.05$) for weaning weight EBV, Carcass Plus Index, or phenotypic weaning weight. The average coefficient of relationship among the genotyped population was 0.02. In comparison, the OCS 100, OCS 50, and OCS 25 animals had an average relationship amongst themselves of 0.01, 0.01, and 0.00, respectively. The OCS 100, OCS 50, and OCS 25 selected 59, 29, and 13 males and 41, 21, and 12 females, respectively.

Table 4.13 Minimum (Min), mean, maximum (Max), standard deviation (St Dev), and coefficient of variation (CV %) for weaning weight (WWT) estimated breeding values (EBV), phenotypic weaning weight (WWT), and Carcass Plus Index (CPI) for the genotyped Suffolk population and the three optimal contribution selection (OCS) sampled populations

	Suffolk population	OCS 100	OCS 50	OCS 25
Min WWT EBV	-3.7	-2.6	-1.8	-1.3
Mean WWT EBV	0.8[a]	0.8[a]	0.6[a]	0.7[a]
Max WWT EBV	7.5	4.8	3.3	2.3
Std Dev WWT EBV	1.6	1.3	1.0	0.9
CV % WWT EBV	200.1	162.0	152.5	129.3
Min WWT	19.7	19.7	19.7	30.3
Mean WWT	32.7[a]	32.7[a]	32.4[a]	33.5[a]
Max WWT	53.8	52.0	46.1	37.3
Std Dev WWT	5.0	5.3	5.2	2.1
CV % WWT	15.1	16.2	16.1	6.2
Min CPI	59.4	83.5	85.4	89.2
Mean CPI	110.0[a]	113.0[a]	111.7[a]	110.5[a]
Max CPI	176.5	162.5	162.5	132.2
Std Dev CPI	17.8	15.2	15.5	10.1
CV % CPI	16.1	13.5	13.8	9.2

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

Average H_E and H_O were both higher for the sampled populations than for the genotyped Suffolk population, which was confirmed by ANOVA and Tukey-Kramer tests. Similarly, average F_{IS} was lower for the sampled populations (Table 4.14). The success of the sampling strategy is measured by capturing the alleles available in the full population. The OCS 100, OCS 50, and OCS 25 captured 99.3, 98.7, and 97.4 percent of the alleles in the population (Table 4.15). If MAF < 0.01 were excluded, OCS 100, OCS 50, and OCS 25 captured 99.9, 99.6, and 98.6 percent of the alleles, respectively.

Table 4.14 Average expected heterozygosity (H_E), observed heterozygosity (H_O), and molecular inbreeding (F_{IS}) for the genotyped Suffolk population and the three optimal contribution selection (OCS) sampled populations

	Suffolk population	OCS 100	OCS 50	OCS 25
H_E	0.318[a]	0.321[b]	0.325[c]	0.330[d]
H_O	0.308[a]	0.315[b]	0.321[bc]	0.328[c]
F_{IS}	0.030[a]	0.019[ab]	0.012[b]	0.006[b]

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

Table 4.15 Percentage of Suffolk population alleles captured and frequency of missing alleles from the optimal contribution selection (OCS) sampling strategies

		Frequency of Missing Alleles		
Selection Category	% Alleles Captured	Minimum	Average	Maximum
OCS 100	99.33	0.0020	0.0045	0.0512
OCS 50	98.72	0.0020	0.0089	0.1414
OCS 25	97.35	0.0020	0.0170	0.1926

The MAF categories show the percentage of fixed, rare, moderate, and high frequency SNP for the full and OCS sampled populations (Table 4.16). The sampled populations had more fixed alleles than the full population with a higher percentage of fixed alleles when fewer animals were sampled. Across the full and sampled populations, the moderate and high categories contained the highest percentage of alleles. Based on Chi-square analyses, the proportions differed by MAF category when the full population was compared to each OCS sampled population ($P < 0.001$).

Table 4.16 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for the genotyped Suffolk population and the optimal contribution (OCS) sampled populations

MAF Category: % of SNP	Suffolk population	OCS 100	OCS 50	OCS 25
Fixed (0)	7.2	8.5	9.6	12.3
Rare (< 0.01)	3.4	1.8	0	0
Moderate (0.01 – 0.3)	55.2	55.3	55.5	52.4
High (0.3 - 0.5)	34.3	34.5	34.9	35.3

The ROH were compared by size class for the full population and the OCS sampled populations (Table 4.17). The majority of ROH were in the smallest class size across all populations. The largest class size was 10 to 20 Mbps and included 0.3 percent of the full genotyped population and 0.1 to 0.2 percent of the OCS sampled populations. Chi-square analyses indicated the proportions for each ROH class differed for OCS 50 when compared to the full population ($P = 0.03$), but not for OCS 100 or OCS 25 ($P > 0.05$). The average number of ROH per animal was 114.9 for the full genotyped population and ranged from 104.6 to 109.1 for the OCS sampled populations (Table 4.18). The range of ROH per animal was variable, with a range of 38 to 252 ROH for the full population. Both OCS 100 and OCS 50 had a range of 38 to 201 per animal, while OCS 25 had a range of 70 to 201 runs per animal. For the full population, recent N_e was computed as 58.3 and the OCS sampled populations were 52.1, 52.9, and 51.6 for OCS 100, OCS 50, and OCS 25, respectively.

Table 4.17 Runs of homozygosity (ROH) by size class and total ROH percentage for the genotyped Suffolk population and the optimal contribution selection (OCS) sampled populations

ROH Class Category (Mbps)	Percent Total ROH in each Class Size			
	Suffolk population	OCS 100	OCS 50	OCS 25
1 - 5	95.9	96.2	96.6	96.8
5 - 10	3.8	3.6	3.2	3.1
10 - 20	0.3	0.2	0.2	0.1

Table 4.18 Average number of runs of homozygosity (ROH) per animal for the genotyped Suffolk population and the optimal contribution selection (OCS) sampled populations

	Average ROH
Suffolk population	114.9
OCS 100	109.1
OCS 50	104.6
OCS 25	108.7

Model-based population structure was compared for the full genotyped population and each OCS sampled population. All 14 subpopulations were represented in each of the OCS sampled populations. For the full genotyped population, there were 30 animals assigned to only one subpopulation. For OCS 100, OCS 50, and OCS 25, there were 8, 4, and 2 animals assigned to only one subpopulation, respectively. Figure 4.2 shows the proportional assignment of each animal for the full population (Figure 4.2(a)) and each OCS sampled population (Figure 4.2(b-d)). Although all 14 subpopulations are represented in each OCS sampled population, the graph shows how the proportions have changed with sampling when compared to the full population, with minimal representation of some populations in the OCS 25 sampling.

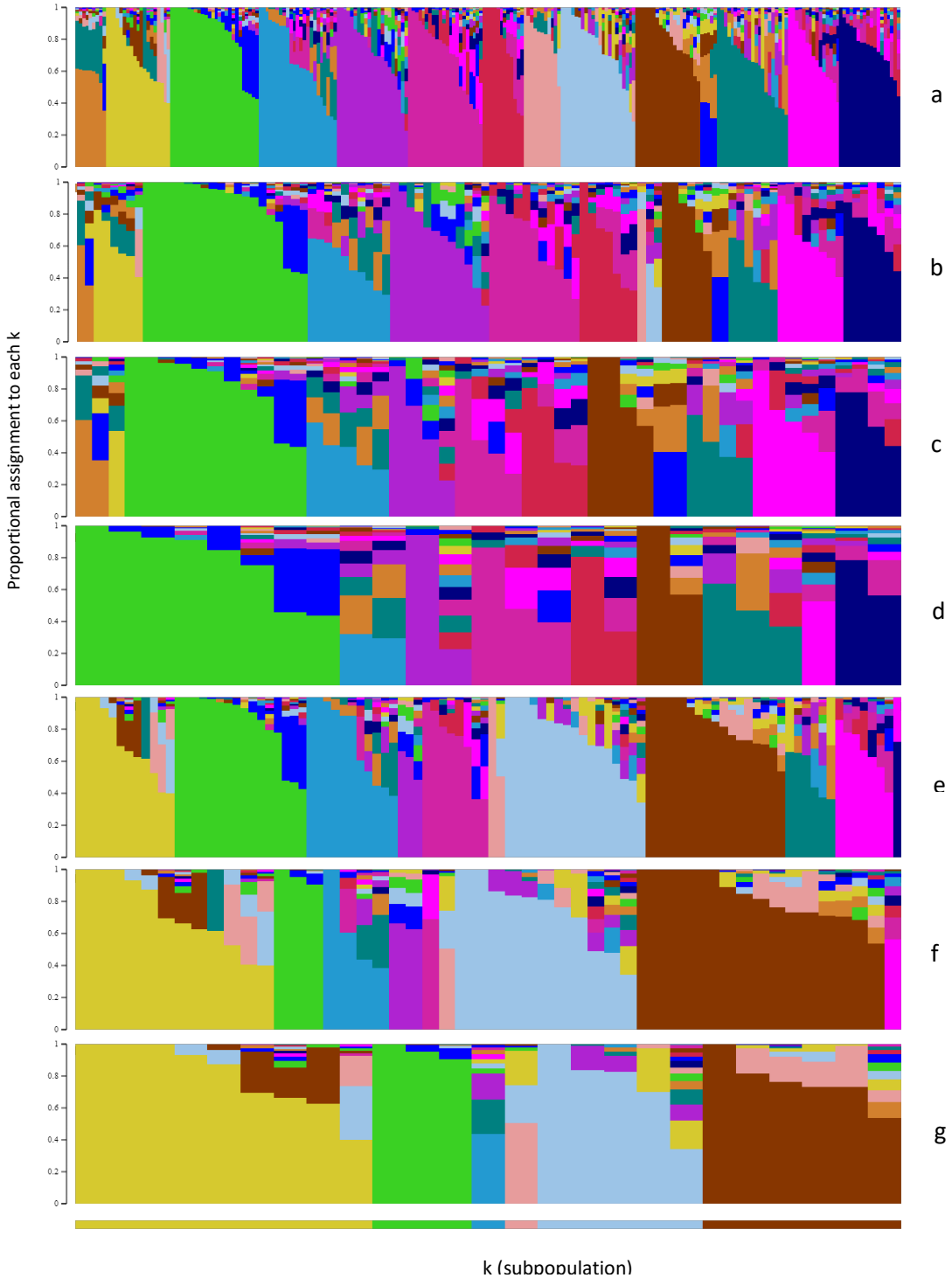

Figure 4.2 Model-based population structure for the genotyped Suffolk population for the optimal contribution (OCS) and Genetic Conservation Index (GCI) strategies for the full population (a), OCS 100 (b), OCS 50 (c), OCS 25 (d), GCI 100 (e), GCI 50 (f), and GCI 25 (g), displaying the proportional assignment of each animal as a column and sorted by highest proportional assignment to a subpopulation

Comparisons of the full genotyped Suffolk population and each GCI sampled population are shown in Table 4.19. Based on both ANOVA and Tukey-Kramer tests, mean weaning weight EBV in the full population differed from that in each GCI population ($P < 0.001$). Similarly, the Carcass Plus Index in the full population differed from that in each GCI population ($P < 0.001$). In contrast, the means between phenotypic weaning weight from the full population did not differ from each GCI population ($P > 0.05$). The average coefficient of relationship between the genotyped population was 0.02; the GCI sampled populations have higher relationships ranging from 0.03 for GCI 100 to 0.09 for GCI 25. For the selected populations, GCI 100 had 59 males and 41 females, GCI 50 had 31 males and 19 females, and GCI 25 had 15 males and 10 females.

Table 4.19 Minimum (Min), mean, maximum (Max), standard deviation (St Dev), and coefficient of variation (CV %) for weaning weight (WWT) estimated breeding values (EBV), phenotypic weaning weight (WWT), and Carcass Plus Index (CPI) for the genotyped Suffolk population and the three Genetic Conservation Index (GCI) sampled populations

	Suffolk population	GCI 100	GCI 50	GCI 25
Min WWT EBV	-3.7	-3.2	-3.2	-0.5
Mean WWT EBV	0.8[a]	1.6[b]	2.3[bc]	2.8[c]
Max WWT EBV	7.5	6.7	6.7	5.9
Std Dev WWT EBV	1.6	2.0	2.1	2.0
CV % WWT EBV	200.1	121.2	92.5	69.4
Min WWT	19.7	19.7	19.7	27.5
Mean WWT	32.7[a]	32.8[a]	33.4[a]	34.4[a]
Max WWT	53.8	51.3	51.3	51.3
Std Dev WWT	5.0	5.3	5.4	5.4
CV % WWT	15.1	16.2	16.1	15.8
Min CPI	59.4	86.6	86.6	86.6
Mean CPI	110.0[a]	121.4[b]	121.8[b]	127.1[b]
Max CPI	176.5	167.3	162.2	155.8
Std Dev CPI	17.8	21.4	23.4	20.7
CV % CPI	16.1	17.6	19.2	16.3

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

Heterozygosity of the GCI sampled populations was higher than the full population for both H_E and H_O (Table 4.20). These differences were confirmed by ANOVA and Tukey-Kramer tests with the exception of H_O when comparing the full population and GCI 100 (P = 0.72). Only GCI 25 differed from the full population for molecular inbreeding (*P* = 0.05). The percentage of alleles captured by GCI 100, GCI 50, and GCI 25 was 99.0, 97.6, and 95.7, respectively (Table 4.21). If MAF < 0.01 were excluded, the percentage of alleles captured increased to 99.8, 98.8, and 97.1 for GCI 100, GCI 50, and GCI 25, respectively.

Table 4.20 Average expected heterozygosity (H_E), observed heterozygosity (H_O), and molecular inbreeding (F_{IS}) for the genotyped Suffolk population and the three Genetic Conservation Index (GCI) sampled populations

	Suffolk population	GCI 100	GCI 50	GCI 25
H_E	0.318^{a}	0.321^{b}	0.323^{c}	0.327^{d}
H_O	0.308^{a}	0.310^{ab}	0.316^{b}	0.327^{c}
F_{IS}	0.030^{a}	0.034^{a}	0.023^{ab}	0.002^{b}

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, *P* < 0.05)

Table 4.21 Percentage of Suffolk population alleles captured and frequency of missing alleles from the Genetic Conservation Index (GCI) sampling strategies

		Frequency of Missing Alleles		
Selection Category	% Alleles Captured	Minimum	Average	Maximum
GCI 100	99.02	0.0020	0.0060	0.0779
GCI 50	97.56	0.0020	0.0146	0.1578
GCI 25	95.67	0.0020	0.0310	0.4978

The MAF categories were compared for the full Suffolk population and the GCI sampled animals (Table 4.22). Based on Chi-square analyses, the proportions differed by MAF category when the full population was compared to each GCI sampled population (*P* < 0.001). The percentage of fixed alleles was higher for the sampled populations and increased as the number of sampled animals decreased.

Moderately and highly polymorphic SNP made up most of the Suffolk population and the GCI sampled populations.

Table 4.22 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for the genotyped Suffolk population and the Genetic Conservation Index (GCI) sampled populations

MAF Category: % of SNP	Suffolk population	GCI 100	GCI 50	GCI 25
Fixed (0)	7.2	9.1	11.9	15.5
Rare (< 0.01)	3.4	1.8	0.0	0.0
Moderate (0.01 – 0.3)	55.2	54.8	54.2	50.8
High (0.3 - 0.5)	34.3	34.3	33.9	33.7

The ROH by size class were comparable between the full population and each GCI sampled population (Table 4.23) and did not differ based on Chi-square analyses comparing the proportions within each ROH class ($P > 0.05$). However, the average number of ROH per animal were higher for the GCI sampled populations than the full Suffolk population (Table 4.24). Recent N_e was computed as 58.3 for the full population and the GCI sampled populations were 39.8, 29.5, and 21.6 for GCI 100, GCI 50, and GCI 25, respectively.

Table 4.23 Runs of homozygosity (ROH) by size class and total ROH percentage for the genotyped Suffolk population and the Genetic Conservation Index (GCI) sampled populations

ROH Class Category (Mbps)	Percent Total ROH in each Class Size			
	Suffolk population	GCI 100	GCI 50	GCI 25
1 – 5	95.9	95.6	95.5	95.2
5 – 10	3.8	4.1	4.1	4.5
10 – 20	0.3	0.3	0.3	0.3

Table 4.24 Average number of runs of homozygosity (ROH) per animal for the genotyped Suffolk population and the Genetic Conservation Index (GCI) sampled populations

	Average ROH
Suffolk population	114.9
GCI 100	123.7
GCI 50	132.5
GCI 25	136.6

Model-based population structure was compared for the full genotyped population and the GCI sampled populations. Each of the GCI sampled populations included all 14 subpopulations. For the full genotyped population, there were 30 animals assigned to only one subpopulation. For GCI 100, GCI 50, and GCI 25, there were 16, 12, and 6 animals assigned to only one subpopulation, respectively. The proportional assignment of each animal for the full population and each GCI sampled population is shown in Figure 4.2(a, e-g). From the graphs, it is clear the proportional assignment to each subpopulation changes as the number of sampled animals decreases with GCI 25 having only four predominant subpopulations.

Simmental cattle. For the weaning weight (EPD and phenotype) and the All-Purpose Index metrics, the full genotyped population was compared to the animals selected in OCS 100, OCS 50, and OCS 25 to see if the full range of variation was captured (Table 4.25). The mean of the Simmental population differed from each OCS sampled population based on both ANOVA and Tukey-Kramer tests ($P \leq 0.01$) for phenotypic weaning weight. For weaning weight EPD, the Simmental population mean differed only from the OCS 100 ($P = 0.03$), but not OCS 50 ($P = 0.96$) or OCS 25 ($P = 0.23$). For API, the mean of the Simmental population differed from each OCS sampled population based on both ANOVA and Tukey-Kramer tests ($P < 0.001$). The average coefficient of relationship among the genotyped population was 0.03. In comparison, the OCS 100, OCS 50, and OCS 25 animals had an average

relationship of 0.01, 0.01, and 0.00, respectively. The genotyped population included 949 males and 4,664 females. OCS 100, OCS 50, and OCS 25 selected 11, 6, and 4 males and 89, 44, and 21 females, respectively.

Table 4.25 Minimum (Min), mean, maximum (Max), standard deviation (St Dev), and coefficient of variation (CV %) for weaning weight (WWT) expected progeny differences (EPD), phenotypic weaning weight (WWT), and All-Purpose Index (API) for the genotyped Simmental population and the three optimal contribution selection (OCS) sampled populations

	Simmental population	OCS 100	OCS 50	OCS 25
Min WWT EPD	36.7	49.3	49.3	55.6
Mean WWT EPD	75.0[a]	71.8[b]	75.8[ab]	79.2[a]
Max WWT EPD	123.3	104.6	104.6	104.6
Std Dev WWT EPD	11.1	13.0	15.1	15.1
CV % WWT EPD	14.7	18.1	19.9	19.0
Min WWT	94.5	167.7	172.7	172.7
Mean WWT	292.0[a]	261.2[b]	253.9[b]	234.6[b]
Max WWT	473.2	363.6	322.7	272.7
Std Dev WWT	49.2	44.5	34.4	34.2
CV % WWT	16.9	17.0	13.5	14.6
Min API	46.0	56.0	66.2	73.5
Mean API	115.5[a]	90.5[b]	92.4[b]	94.0[b]
Max API	169.4	116.3	116.3	116.3
Std Dev API	18.2	13.1	12.4	13.1
CV % API	15.7	14.5	13.4	13.9

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

Both ANOVA and Tukey-Kramer tests confirmed differences in the means for the full population and the sampled populations for H_E and H_O ($P > 0.05$). Average H_E and H_O were both higher for the sampled populations than for the Simmental population. Average F_{IS} did not differ for the sampled and full genotyped populations ($P > 0.05$). (Table 4.26). The OCS 100, OCS 50, and OCS 25 captured 99.93, 99.78, and 99.71 percent of the alleles in the population (Table 4.27). If SNP with a MAF < 0.01 were excluded, OCS 100, OCS 50, and OCS 25 captured 99.99, 99.96, and 99.91 percent of the alleles, respectively.

Table 4.26 Average expected heterozygosity (H_E), observed heterozygosity (H_O), and molecular inbreeding (F_{IS}) for the genotyped Simmental population and the three optimal contribution selection (OCS) sampled populations

	Simmental population	OCS 100	OCS 50	OCS 25
H_E	0.416^a	0.422^b	0.429^c	0.431^d
H_O	0.408^a	0.413^b	0.419^{bc}	0.426^c
F_{IS}	0.019^a	0.021^a	0.024^a	0.012^a

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

Table 4.27 Percentage of Simmental population alleles captured and frequency of missing alleles from the optimal contribution selection (OCS)sampling strategies

		Frequency of Missing Alleles		
Selection Category	% Alleles Captured	Minimum	Average	Maximum
OCS 100	99.93	0.0128	0.0186	0.0244
OCS 50	99.78	0.0117	0.0243	0.0679
OCS 25	99.71	0.0100	0.0384	0.3587

The MAF categories show the percentage of fixed, rare, moderate, and high SNP for the full and OCS sampled populations (Table 4.28). The full genotyped population and the sampled populations had few fixed alleles. However, the smaller the sampled population, the higher the percentage of fixed alleles. Across the full and sampled populations, the moderate and high categories contained the highest percentage of alleles. Based on Chi-square analyses comparing the MAF categories of the full population to each sample population, the proportions differed ($P > 0.05$).

Table 4.28 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for the genotyped Simmental population and the optimal contribution selection (OCS) sampled populations

MAF Category: % of SNP	Simmental population	OCS 100	OCS 50	OCS 25
Fixed (0)	0.05	0.18	0.48	0.62
Rare (< 0.01)	0.56	0.12	0.00	0.00
Moderate (0.01 – 0.3)	34.06	32.30	28.54	26.69
High (0.3 - 0.5)	65.33	67.39	70.97	72.68

Table 4.29 shows the ROH by size class for the full population and the OCS sampled populations. The smallest class size included approximately 70 percent of the ROH and the intermediate class size included approximately 30 percent across all populations. The largest class size was 10 to 20 Mbps and included 1.1 percent of the full genotyped population and 0.5 to 1.1 percent of the OCS sampled populations. When comparing the ROH classes from the full population to each sampled population, the proportions did not differ ($P > 0.05$). The average number of ROH was 5.9 for the full genotyped population and ranged from 3.6 to 4.5 for the OCS sampled populations (Table 4.30). The range of ROH per animal was variable, with a range of 0 to 44 ROH for the full population. Both OCS 100 and OCS 50 had a range of 0 to 17 per animal while OCS 25 had a range of 0 to 11 runs per animal. Recent N_e was computed as 153.8 for the full population and the OCS sampled populations were 89.2, 61.9, and 58.9 for OCS 100, OCS 50, and OCS 25, respectively.

Table 4.29 Runs of homozygosity (ROH) by size class and total ROH percentage for the genotyped Simmental population and the optimal contribution selection (OCS) sampled populations

ROH Class Category (Mbps)	Percent Total ROH in each Class Size			
	Simmental population	OCS 100	OCS 50	OCS 25
1 – 5	68.4	69.2	71.0	69.6
5 – 10	30.5	29.7	28.5	29.5
10 – 20	1.1	1.1	0.5	0.9

Table 4.30 Average number of runs of homozygosity (ROH) per animal for the genotyped Simmental population and the optimal contribution selection (OCS) sampled populations

	Average ROH
Simmental population	5.9
OCS 100	3.6
OCS 50	4.4
OCS 25	4.5

Model-based population structure is shown in Figure 4.3 and compares the full genotyped population (Figure 4.3(a)) and each OCS sampled population (Figure 4.3(b-d)). All 15 subpopulations

were represented in each of the OCS sampled populations. For the full genotyped population, there were 55 animals assigned to only one subpopulation, but no animals from the sampled populations were assigned to only one subpopulation. Although all 15 subpopulations are represented in each OCS sampled population, the figure shows how the proportions have changed with sampling when compared to the full population. Subpopulations 5, 8, and 13 have a high level of representation in the sampled populations. Subpopulation 8 includes a high number of lower percentage (admixed) Simmental animals.

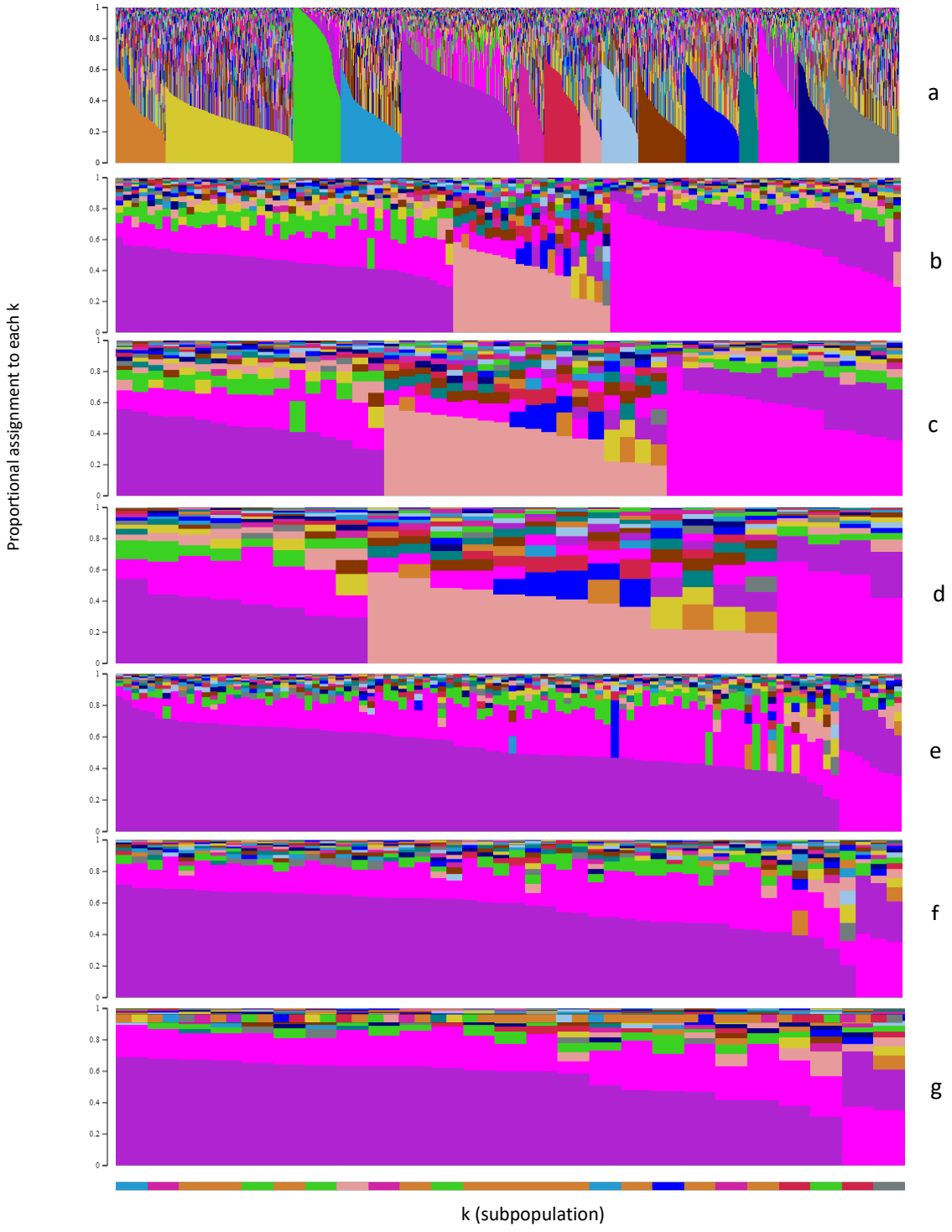

Figure 4.3 Model-based population structure for the genotyped Simmental population for the optimal contribution (OCS) and Genetic Conservation Index (GCI) strategies for the full population (a), OCS 100 (b), OCS 50 (c), OCS 25 (d), GCI 100 (e), GCI 50 (f), and GCI 25 (g), displaying the proportional assignment of each animal as a column and sorted by highest proportional assignment to a subpopulation

The full genotyped Simmental population and the GCI sampled populations were compared for the quantitative traits of weaning weight EPD, phenotypic weaning weight, and All-Purpose Index (Table 4.31). For both phenotypic weaning weight and weaning weight EPD, the mean of the Simmental population differed from the GCI 100 population but not GCI 50 or GCI 25. For the All Purpose Index, the mean of the Simmental population differed from each of the GCI populations ($P < 0.001$). The average coefficient of relationship between the genotyped population was 0.03 while the GCI sampled populations have higher relationships ranging from 0.06 for GCI 100 to 0.09 for GCI 25. For the selected populations, GCI 100 had 28 males and 72 females, GCI 50 had 15 males and 35 females, and GCI 25 had 9 males and 16 females.

Table 4.31 Minimum (Min), mean, maximum (Max), standard deviation (St Dev), and coefficient of variation (CV %) for weaning weight (WWT) expected progeny differences (EPD), phenotypic weaning weight (WWT), and All Purpose Index (API) for the genotyped Simmental population and the three Genetic Conservation Index (GCI) sampled populations

	Simmental population	GCI 100	GCI 50	GCI 25
Min WWT EPD	36.7	52.4	52.8	54.5
Mean WWT EPD	75.0[a]	70.3[b]	71.6[ab]	71.2[ab]
Max WWT EPD	123.3	95.2	95.2	95.2
Std Dev WWT EPD	11.1	9.4	10.4	10.5
CV % WWT EPD	14.7	13.3	14.5	14.8
Min WWT	94.5	184.5	184.5	186.4
Mean WWT	292.0[a]	264.6[b]	274.1[ab]	289.5[ab]
Max WWT	473.2	381.8	381.8	381.8
Std Dev WWT	49.2	43.8	48.9	67.1
CV % WWT	16.9	16.5	17.9	23.2
Min API	46.0	63.1	76.7	76.7
Mean API	115.5[a]	99.0[b]	99.2[b]	97.2[b]
Max API	169.4	153.2	123.0	121.2
Std Dev API	18.2	13.6	10.4	10.9
CV % API	15.7	13.8	10.5	11.2

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

The H_E for the GCI sampled populations was lower than the full genotyped population ($P < 0.05$) (Table 4.32). The H_O differed between the full population and the GCI 25 sampled population ($P = 0.008$) but not for GCI 100 ($P = 0.40$) or GCI 50 ($P = 0.24$). Molecular inbreeding was lower for the GCI sampled populations than the full genotyped population ($P < 0.05$). The percentage of alleles captured by GCI 100, GCI 50, and GCI 25 was 99.7, 99.5, and 99.2, respectively (Table 4.33). If MAF < 0.01 were excluded, the percentage of alleles captured increased to 99.9, 99.7, and 99.4 for GCI 100, GCI 50, and GCI 25, respectively.

Table 4.32 Average expected heterozygosity (H_E), observed heterozygosity (H_O), and molecular inbreeding (F_{IS}) for the genotyped Simmental population and the three Genetic Conservation Index (GCI) sampled populations

	Simmental population	GCI 100	GCI 50	GCI 25
H_E	0.416^{a}	0.403^{b}	0.399^{c}	0.396^{d}
H_O	0.408^{a}	0.411^{ab}	0.413^{ab}	0.421^{b}
F_{IS}	0.019^{a}	-0.019^{b}	-0.036^{bc}	-0.062^{c}

Means sharing the same superscript within a row are not significantly different from each other (Tukey-Kramer, $P < 0.05$)

Table 4.33 Percentage of Simmental population alleles captured and frequency of missing alleles from the Genetic Conservation Index (GCI)sampling strategies

		Frequency of Missing Alleles		
Selection Category	% Alleles Captured	Minimum	Average	Maximum
GCI 100	99.7	0.0100	0.0365	0.1162
GCI 50	99.5	0.0100	0.0517	0.2521
GCI 25	99.2	0.0100	0.0742	0.3261

Table 4.34 shows the MAF categories for the percentage of fixed, rare, moderate, and high SNP for the full and GCI sampled populations. The full genotyped population and the sampled populations had few fixed alleles, but the number of fixed alleles increased as the number of sampled animals decreased. All populations had most of their alleles in the moderate and high categories. Based on Chi-

square analyses comparing the MAF categories of the full population to each sample population, the proportions differed ($P > 0.05$).

Table 4.34 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for the genotyped Simmental population and the Genetic Conservation Index (GCI) sampled populations

MAF Category: % of SNP	Simmental population	GCI 100	GCI 50	GCI 25
Fixed (0)	0.05	0.65	1.08	1.61
Rare (< 0.01)	0.56	0.32	0.00	0.00
Moderate (0.01 – 0.3)	34.06	38.42	39.84	40.10
High (0.3 - 0.5)	65.33	60.61	59.08	58.29

When comparing the ROH classes from the full population to each sampled population, the proportions did not differ ($P > 0.05$) (Table 4.35). The average number of ROH was 5.9 for the full genotyped population and ranged from 0 to 44 (Table 4.36). The ROH per animal was similar for GCI 100 and GCI 50 with an average of 3.6 per animal with a range of 0 to 16. The GCI 25 had an average of 2.4 runs per animal with a range of 0 to 8. Recent N_e was 57.9, 30.6, and 31.5 for GCI 100, GCI 50, and GCI 25, respectively; in comparison, the full population had an N_e of 153.8.

Table 4.35 Runs of homozygosity (ROH) by size class and total ROH percentage for the genotyped Simmental population and the Genetic Conservation Index (GCI) sampled populations

ROH Class Category (Mbps)	Percent Total ROH in each Class Size			
	Simmental population	GCI 100	GCI 50	GCI 25
1 – 5	68.4	72.1	77.7	83.6
5 – 10	30.5	26.8	20.7	16.4
10 – 20	1.1	1.1	1.7	0.0

Table 4.36 Average number of runs of homozygosity (ROH) per animal for the genotyped Simmental population and the Genetic Conservation Index (GCI) sampled populations

	Average ROH
Simmental population	5.9
GCI 100	3.6
GCI 50	3.6
GCI 25	2.4

Model-based population structure is shown in Figure 4.3 and compares the full genotyped population (Figure 4.3(a)) and each GCI sampled population (Figure 4.3(e-g)). All 15 subpopulations were represented in each of the GCI sampled populations, but they are dominated by subpopulations 5 and 13 with minimal representation of the other subpopulations. In contrast to the OCS sampled populations, assignment to subpopulations was not influenced by percentage Simmental. This is consistent across GCI 100, GCI 50, and GCI 25 sampled populations. There were no animals from the sampled populations assigned to only one subpopulation.

4.4 Discussion

Simulated population. The success of the sampling methods depends upon capturing the available variation in the population, including the full range of phenotypic traits. In comparison to the GCI sampling, the OCS sampling performed better at capturing the lower end of the weaning weight range with OCS 100 performing the best. The GCI sampling performed better at capturing the upper end of the weaning weight range with GCI 100 performing the best. Similarly, the OCS sampling performed better at sampling the lower end of the weaning weight EBV range and the GCI sampling performed better at sampling the higher end. Significant differences between the means of the year 15 animals and the OCS sampled populations indicate the sampling procedures did not fully reflect the full population for weaning weight and weaning weight EBV.

Measures of gene diversity via H_E and H_O were higher for all sampled populations than the full populations. Higher heterozygosity would be expected from a successful genetic diversity sampling strategy. This suggests overall heterozygosity is being maintained in the sampled populations. ROH samples were similar when comparing the full population and all sampled populations. Subsampling of this population has maintained the ROH within the population. The MAF category proportions differed between the full population and each sampled population with the fixed alleles increasing as the sampled population size decreased, indicating a loss of alleles in the sampled populations. Recent Ne was lower for all sampled populations when compared to the full population and decreased as the number of sampled animals decreased. At each sampling size, OCS had a higher N_e than GCI.

The primary indicator of sampling success is capturing all available alleles in the population. In the simulated population, there was a high percentage of low frequency alleles. The OCS sampling did not capture all the available alleles, with OCS 100, OCS 50, and OCS 25 capturing 95.4, 93.4, and 90.6 percent of the alleles, respectively. If rare alleles were excluded (MAF < 0.01), the percentage of alleles captured increased to 99.6, 98.5, and 96.7 for OCS 100, OCS 50, and OCS 25, respectively. In a real population, a MAF < 0.01 would typically be filtered during the quality control process because it is difficult to determine if these are real rare alleles or genotyping errors. In this simulated population, these are true rare alleles because there are no genotyping errors. The GCI sampling also did not capture all available alleles, with 95.2, 93.6, and 91.4 percent of the alleles being captured by GCI 100, GCI 50, and GCI 25, respectively, With MAF < 0.01 excluded, 99.3, 98.1, and 95.9 percent of the alleles were captured by GCI 100, GCI 50, and GCI 25, respectively. Overall, OCS 100 captured the most alleles whether MAF < 0.01 was excluded or not. When comparing OCS and GCI for the same number of sampled animals (e.g., OCS 100 vs GCI 100), OCS captured more alleles than the GCI counterpart across all three sample sizes whether MAF < 0.01 was excluded or not.

For the model-based population structure (Figure 4.1), all sampling methods included animals from the 15 subpopulations. The OCS 100 (b) most closely resembles the subpopulation structure of the full population (a) followed by GCI 100 (e). When comparing the OCS plots (b – d) to the GCI plots (e – g), it is clear the subpopulations are represented in different proportions depending on the sampling method.

Suffolk sheep. A comparison of means for the full genotyped Suffolk population and the OCS sampled populations were not significantly different for weaning weight EBV, Carcass Plus Index, or phenotypic weaning weight. The GCI sampled populations differed from the full population for weaning weight EBV and Carcass Plus Index, but not phenotypic weaning weight. This suggests the OCS sampled populations better reflected the full population than the GCI sampled populations. The average coefficient of relationship among each of the OCS sampled animals was lower than the average coefficient of relationship of the full population while the GCI sampled animals were higher. Since the OCS sampling strategy places emphasis on minimizing kinship, a below average coefficient of relationship was expected. Alternatively, GCI focuses on maximizing founder alleles in each animal without regard to relationship.

Heterozygosity (H_E and H_O) were higher for the sampled populations than the full population with the exception of GCI 100 for H_O. This suggests overall heterozygosity is being maintained in the sampled populations. The MAF categories differed for each sampled population when compared to the full population. This was expected as the number of fixed alleles increased as alleles were lost from each successively smaller sampled population. The ROH class sizes were similar when comparing the full population and all sampled populations, except for OCS 50. The average number of ROH per animal for the GCI sampled populations were higher than the full population while the OCS sampled populations

were lower. For recent Ne, the OCS sampled populations were similar to the full population, and the GCI sampled populations were lower.

The primary measure of interest for sampling success is capturing all the alleles available in the full population. The OCS 100, OCS 50, and OCS 25 captured 99.3, 98.7, and 97.4 percent of the available alleles, respectively. The GCI 100, GCI 50, and GCI 25 captured 99.0, 97.6, and 95.7 percent of the available alleles, respectively. Of the missing alleles, the GCI sampled populations missed alleles with a higher average and maximum allele frequency than the OCS sampled populations. For example, GCI 25 did not capture an allele that was present at an allele frequency of 0.50. The highest missing allele frequency for OCS 25 was 0.19. If MAF < 0.01 was excluded, OCS 100, OCS 50, and OCS 25 captured 99.9, 99.6, and 98.6 percent of the available alleles, respectively while GCI 100, GCI 50, and GCI 25 captured 99.8, 98.8, and 97.1 percent of the available alleles, respectively. When comparing the sampling methods for the same number of selected animals, OCS outperformed GCI at each level. However, all sampling methods performed well at capturing the available alleles.

When comparing model-based population structure (Figure 4.2), all sampling methods included animals from the 14 subpopulations. Of the 30 animals from the full population that were assigned to a single subpopulation, OCS 100, OCS 50, OCS 25, GCI 100, GCI 50, and GCI 25 had 8, 4, 2, 16, 12, and 6 animals assigned to a single subpopulation, respectively. Because the sampling strategy defining GCI includes maximizing founders for each animal, more admixed animals from many subpopulations was anticipated. However, the opposite occurred, where OCS strategies sampled more broadly from the subpopulations. Figure 4.2 shows OCS 100 (b) most closely resembles the subpopulation structure of the full population (a).

Simmental cattle. For weaning weight EPD, phenotypic weaning weight, and API, the sampled populations did not capture the full range of variation available in the population. When means were

compared for the full population and each OCS or GCI sampled population, only OCS 100 and GCI 100 differed from the full population for weaning weight EPD. For phenotypic weaning weight, all OCS populations and GCI 100 differed from the full population. For API, all OCS and GCI populations differed from the full population. The population average was 0.03 for the pedigree-based coefficient of relationship. The three OCS sampled populations were lower than the full genotyped population, while the three GCI sampled populations were higher.

The H_E was higher for the OCS sampled populations and lower for the GCI sampled populations when compared to the full genotyped Simmental population. The H_O was higher for the OCS and GCI sampled populations than the full genotyped Simmental population. High levels of heterozygosity exist in the full and sampled populations. Pedigree-based inbreeding levels and molecular-based inbreeding (F_{IS}) were similar for both the full genotyped Simmental population and the OCS sampled populations. In contrast the pedigree-based GCI inbreeding levels ranged from 0.06 to 0.09 while the F_{IS} values for these populations were negative. This suggests the observed SNP for these animals showed more heterozygosity than expected based on their pedigrees. The percentage of ROH by class size were similar for the full genotyped Simmental population and the OCS sampled populations. The percentage of ROH for the intermediate class size was lower for the GCI sampled populations. Since shorter ROH tend to represent historical inbreeding and breed founder effects, it is important for the sampled populations to closely mirror the full population for the smaller class sizes. The full and sampled populations collectively had few ROH per animal. The N_e of the full population was higher than the sampled populations, but the OCS sampled populations were higher than the GCI sampled populations at each sampling size.

All the OCS and GCI sampled populations captured more than 99 percent of the available alleles in the population. This shows a population with a high level of heterozygosity and few fixed alleles can be captured in as few as 25 animals with strategic sampling. Of the missing alleles for the OCS samples,

the maximum MAF was 0.02, 0.07, and 0.36 for OCS 100, OCS 50, and OCS 25, respectively. Of the missing alleles for the GCI samples, the maximum MAF was 0.12, 0.25, and 0.33 for GCI 100, GCI 50, and GCI 25, respectively. OCS 25, GCI 50, and GCI 25 sampled populations all missed alleles that exist at a frequency of 0.25 and higher. The OCS 100 captured the highest percentage of available alleles and only missed alleles that existed at a MAF of 0.02 or less. While OCS 100 is the best sampling strategy for this population, any of the OCS and GCI sampling strategies would capture at least 99 percent of the available alleles in the population.

Model-based population structure (Figure 4.3) shows the proportional assignment of each animal for full genotyped population, the OCS sampled populations, and the GCI sampled populations. While all 15 subpopulations are represented in each of the sampled populations, the proportional assignments of the sampled populations do not match the full genotyped Simmental population. The three OCS sampled populations (b-d) are primarily represented by subpopulations 5, 8, and 13 while the three GCI sampled populations (e-g) are primarily represented by subpopulations 5 and 13. Although more than 99 percent of the available alleles were captured by all sampling strategies, the model-based population structure suggests the subpopulations were not proportionally represented.

4.5 Conclusion

While overall population structure was assessed for each selection strategy within each breed, the measure of success was capturing the available alleles in the population. For the simulated population, OCS 100 captured the most alleles followed by GCI 100. GCI 25 was the least successful and captured only 91.4 percent of the alleles. For the Suffolk population, OCS 100 captured the most alleles followed by GCI 100. GCI 25 captured the fewest alleles. For the Simmental population, all sampling strategies captured more than 99 percent of the available alleles. The OCS 100 captured the most alleles

followed by OCS 50. The GCI 25 was the least successful. Across recent N_e values, the OCS 100 samples were the highest when compared to all other sampled populations.

While no studies have compared these two selection strategies, Engelsma et al. (2011) used OCS to sample a range of 5 to 80 animals and compare the selected animals to the full population. The authors evaluated kinship, MAF, and percent of fixed alleles. While they did not make specific recommendations for the number of animals to sample, they did acknowledge the superiority of OCS for selecting animals. Based on the findings from the simulated, Suffolk, and Simmental populations studied here, OCS 100 is the recommended selection strategy for capturing the genetically important animals in a population using pedigree-based methods.

CHAPTER V

ASSESS THE ROBUSTNESS OF SAMPLING STRATEGIES ACROSS SPECIES AND BREEDS

5.1 Introduction

Developing a sampling strategy that captures the genetic diversity available in a breed is important for a variety of purposes, including research, gene banking, building a reference population, and determining which animals to keep or cull. Sampling becomes even more challenging when considering the wide range of livestock species and breeds throughout the world. Species and breeds have been shaped by their original formation, including how many founder animals there were and if any additional animals have been added over time. Bottlenecks throughout the history of the population will narrow the genetic base. The direction of selection pressure and the selection intensity will shape the population and will influence the extent of linkage disequilibrium in the population (Gibbs et al., 2009; Kijas et al., 2012). The physical distribution of the animals, the male to female ratio, and the exchange of genetics among breeders also impacts the population structure. For example, Wilson et al. (2022) found divergence among Suffolk breeders even in close physical proximity to each other, likely due to differing breeding objectives.

Sampling strategies to capture genetic diversity have been developed primarily to enhance genetic selection programs and minimize inbreeding, particularly for conservation of small populations or for gene banking (Gourdine et al., 2012b; Windig and Oldenbroek, 2015; van Breukelen et al., 2019). Avendaño et al. (2003) successfully used OCS to constrain inbreeding while increasing genetic gain for both a sheep and beef breed with different breed histories. van Breukelen et al. (2019) used OCS to develop core collections for cattle breeds within the Dutch gene bank. Machová et al. (2021) used GCI to assess the genetic diversity of two sheep breeds with different breed histories. Comparison of selection

strategies within a breed can be compared by evaluating the kinship, MAF, and percent of fixed alleles (Engelsma et al., 2011).

For a selection strategy to be effective, it must meet a variety of uses across species and breeds while capturing the genetic diversity of each population. Using pedigree-based methods to select animals and then quantifying the overall genetic diversity captured using both quantitative and molecular tools was used to validate the selection strategies. The three populations included a simulated population where the genotypes of every animal in the population was known and included a moderate number of markers, a sheep population with a small population size with many markers, and a beef cattle population with a large population with a moderate number of markers. The objectives of this study were to 1) summarize the population structure of the selection candidates for the three populations, and 2) since OCS 100 was the most effective at capturing the available alleles in each of the three populations, describe the ability of this selection strategy to capture the genetic diversity of each population.

5.2 Materials and Methods

For the simulated population, the full pedigree included 204,930 animals with 13,662 animals from the 15th year considered as the selection candidates, and 53,901 markers used to assess genetic diversity. For the Suffolk sheep population, the full pedigree included 1,565 animals and included the ancestors of the 244 genotyped animals that were considered the selection candidates. From the OvineHD BeadChip, 577,401 autosomal markers were included in the analyses. For the Simmental population, 5,613 animals with overlapping markers from a variety of SNP chips were included as selection candidates. The full pedigree included the ancestors of these animals and was comprised of 54,790 animals. Because of the grading up allowed in the Simmental registry, non-Simmental animals

were included in the computation of relationships. All other quantitative comparisons included animals that were at least 84 percent Simmental (n = 34,462). There were 29,449 autosomal markers included to assess genetic diversity.

The population structure for the selection candidates was described for each population. This included the average and range of pedigree-based inbreeding, heterozygosity measures, and molecular inbreeding. MAF categories, ROH, and N_e were also reported. Model-based population structure was plotted.

Diversity assessments of OCS 100 across the three populations included capturing the range of a phenotypic trait, weaning weight, and a breeding value, weaning weight EBV or EPD. Heterozygosity, MAF categories, and ROH were compared. Model-based population structure plots were compared. Finally, the percentages of alleles captured for each population were summarized.

5.3 Results

Population structure. The average and range of inbreeding for the selection candidates for each population are presented in Table 5.1. The average inbreeding of each population was low with some individual animals with high inbreeding coefficients. Heterozygosity and molecular inbreeding were summarized in Table 5.2 for each population. High levels of heterozygosity and low levels of molecular inbreeding exist in each population. Direct comparisons of heterozygosity across the three populations should not be made since different SNP chips were used for each population.

MAF categories were summarized for the selection candidates in each population as fixed, rare, moderate, and high (Table 5.3). When comparing the population structure for the three populations, those with a higher percentage of rare alleles are more challenging to capture the available alleles. The simulated population followed by the Suffolk population have a higher percentage of low frequency

alleles in the population when compared to Simmental. Percentage of ROH by class size are summarized in Table 5.4. As with heterozygosity measures, direct comparisons of ROH should not be made due to differing SNP chips used for each population; however, some general conclusions can be made. Most importantly, the majority of ROH for the three populations are in the smallest class size, which is generally considered to be from historical events such as breed founder events. Maintenance of these ROH in the sampled populations is desirable as they reflect the history of the population. The Simmental population includes a higher percentage of ROH from the intermediate size class than the other populations; this may be indicative of more recent inbreeding.

Current N_e for the three populations was 289.8, 58.3, and 153.8 for the simulated, Suffolk, and Simmental populations, respectively. Model-based population structure was summarized as 15, 14, and 15 subpopulations for the simulated, Suffolk, and Simmental population, respectively (Figure 5.1). The Suffolk population (b) showed more distinct population substructure than the simulated (a) and Simmental (c) populations.

Table 5.1 Mean, minimum (Min), and maximum (Max) inbreeding (F) for the selection candidates for the Simulated, Suffolk, and Simmental populations

Inbreeding category	Simulated	Suffolk	Simmental
Mean F	0.003	0.011	0.050
Min F	0.000	0.000	0.000
Max F	0.257	0.257	0.306

Table 5.2 Average expected heterozygosity (H_E), observed heterozygosity (H_O), and molecular inbreeding (F_{IS}) for the selection candidates for the Simulated, Suffolk, and Simmental populations

	Simulated	Suffolk	Simmental
H_E	0.259	0.318	0.416
H_O	0.259	0.308	0.408
F_{IS}	-0.001	0.030	0.019

Table 5.3 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for the selection candidates for the Simulated, Suffolk, and Simmental populations

MAF Category: % of SNP	Simulated	Suffolk	Simmental
Fixed (0)	14.2	7.2	0.05
Rare (< 0.01)	9.9	3.4	0.6
Moderate (0.01 – 0.3)	52.1	55.2	34.1
High (0.3 - 0.5)	23.8	34.3	65.3

Table 5.4 Runs of homozygosity (ROH) by size class and total ROH percentage for the selection candidates for the Simulated, Suffolk, and Simmental populations

ROH Class Category (Mbps)	Percent Total ROH in each Class Size		
	Simulated	Suffolk	Simmental
1 – 5	79.0	95.9	68.4
5 – 10	16.4	3.8	30.5
10 – 20	3.9	0.3	1.1
20 – 40	0.7	0.0	0.0
> 40	0.1	0.0	0.0

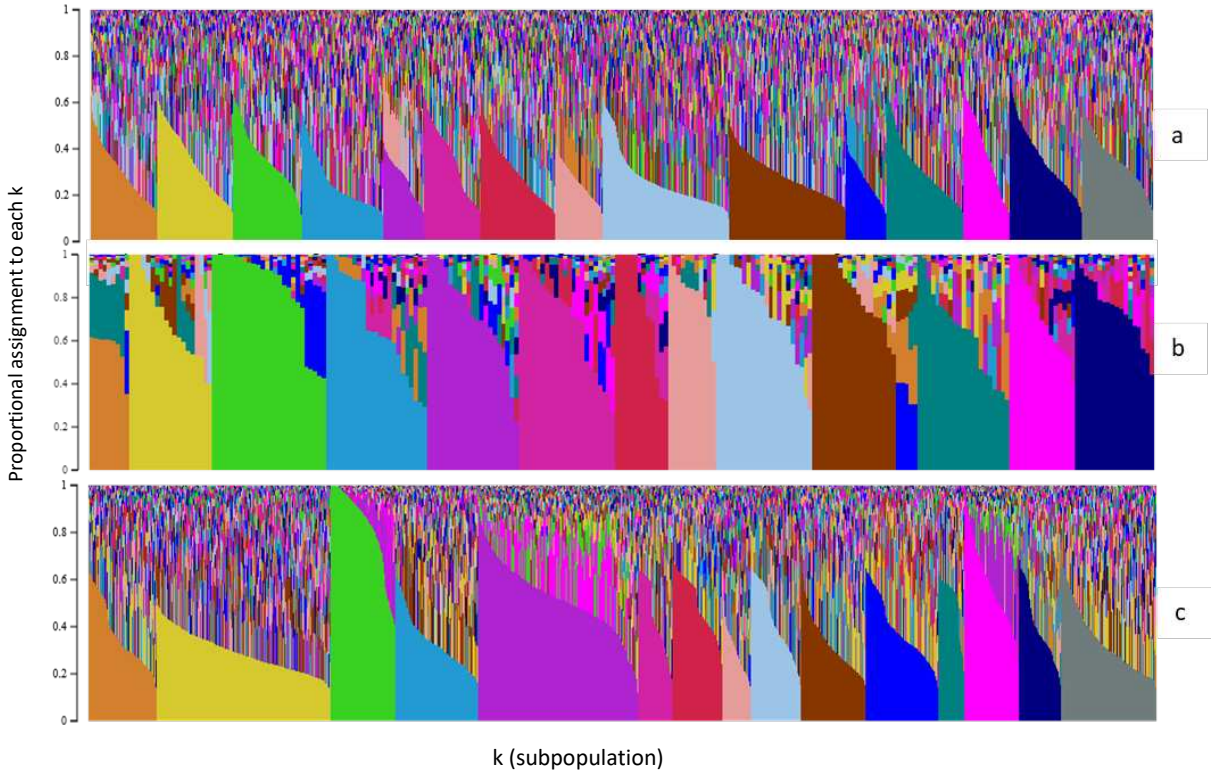

Figure 5.1 Model-based population structure for the selection candidates for the Simulated (a), Suffolk (b), and Simmental (c) populations, displaying the proportional assignment of each animal as a column and sorted by highest proportional assignment to a subpopulation

OCS 100. When comparing the means of OCS 100 to the full simulated population, the means were significantly different for both phenotypic weaning weight and weaning weight EBV, indicating the sampled population did not completely reflect the full population. This was also observed for the Simmental population where the means were different between OCS 100 and the full population for phenotypic weaning weight, weaning weight EPD, and the All-Purpose Index. When comparing the means for the Suffolk population and the OCS 100 sampled population, they were not different for phenotypic weaning weight, weaning weight EBV, or the Carcass Plus Index. This suggests the OCS 100 samples reflect these traits as they exist in the full Suffolk population.

Heterozygosity for the selection candidates and the OCS 100 sampled populations are presented for the three populations in Table 5.5. Heterozygosity of the OCS 100 sampled populations had as much

heterozygosity or more for each of the three populations as was available in the selection candidates. Representation in MAF categories is compared for the selection candidates and the OCS 100 sampled populations for the three populations (Table 5.6). The number of fixed alleles increased in all OCS 100 sampled populations. This was most evident in the simulated population as the number of rare alleles decreased and the number of fixed alleles increased, indicating a loss of rare alleles.

Table 5.5 Average expected heterozygosity (H_E) and observed heterozygosity (H_O) for the selection candidates and optimal contribution selection (OCS) OCS 100 population for the Simulated, Suffolk, and Simmental populations

	Simulated		Suffolk		Simmental	
	Selection candidates	OCS 100	Selection candidates	OCS 100	Selection candidates	OCS 100
H_E	0.259	0.259	0.318	0.321	0.416	0.422
H_O	0.259	0.260	0.308	0.315	0.408	0.413

Table 5.6 Percentage of single nucleotide polymorphisms (SNP) in each minor allele frequency (MAF) category for the selection candidates and optimal contribution selection (OCS) OCS 100 population for the Simulated, Suffolk, and Simmental populations

	Simulated		Suffolk		Simmental	
MAF Category: % of SNP	Selection candidates	OCS 100	Selection candidates	OCS 100	Selection candidates	OCS 100
Fixed (0)	14.2	20.8	7.2	8.5	0.05	0.2
Rare (< 0.01)	9.9	2.7	3.4	1.8	0.6	0.1
Moderate (0.01 – 0.3)	52.1	52.5	55.2	55.3	34.1	32.3
High (0.3 - 0.5)	23.8	24.0	34.3	34.5	65.3	67.4

Total percentage of ROH by class size were summarized for the selection candidates and OCS 100 sampled populations for the three populations (Table 5.7). The percentage of ROH assigned to each class category varied by population. Within each population, the OCS 100 sampled population closely matched the selection candidates. For recent N_e, OCS 100 had the highest values when compared to the other sampled populations. Model-based population structure was compared for the selection candidates and the OCS 100 sampled population for the simulated population (Figure 5.2), the Suffolk

population (Figure 5.3), and the Simmental population (Figure 5.4). All subpopulations were represented in the OCS sampled populations for the three populations. The simulated and Suffolk populations had a more representative sampling of the subpopulations than the Simmental population. This may be due to the highly admixed overall Simmental population.

Table 5.7 Runs of homozygosity (ROH) by size class and total ROH percentage for the selection candidates and optimal contribution selection (OCS) OCS 100 population for the Simulated, Suffolk, and Simmental populations

	Simulated		Suffolk		Simmental	
ROH Class Category (Mbps)	Selection candidates	OCS 100	Selection candidates	OCS 100	Selection candidates	OCS 100
1 - 5	79.0	79.6	95.9	96.2	68.4	69.2
5 - 10	16.4	15.7	3.8	3.6	30.5	29.7
10 - 20	3.9	3.8	0.3	0.2	1.1	1.1
20 - 40	0.7	0.8	0.0	0.0	0.0	0.0
> 40	0.1	0.1	0.0	0.0	0.0	0.0

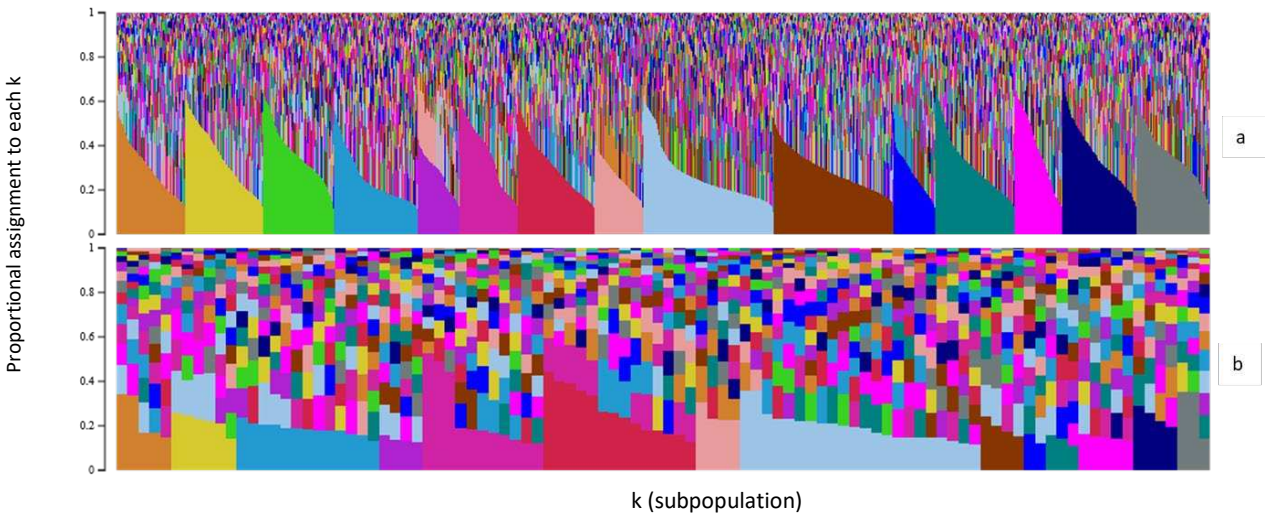

Figure 5.2 Model-based population structure for the Simulated population for selection candidates (a) and optimal contribution selection (OCS) OCS 100 (b) populations, displaying the proportional assignment of each animal as a column and sorted by highest proportional assignment to a subpopulation

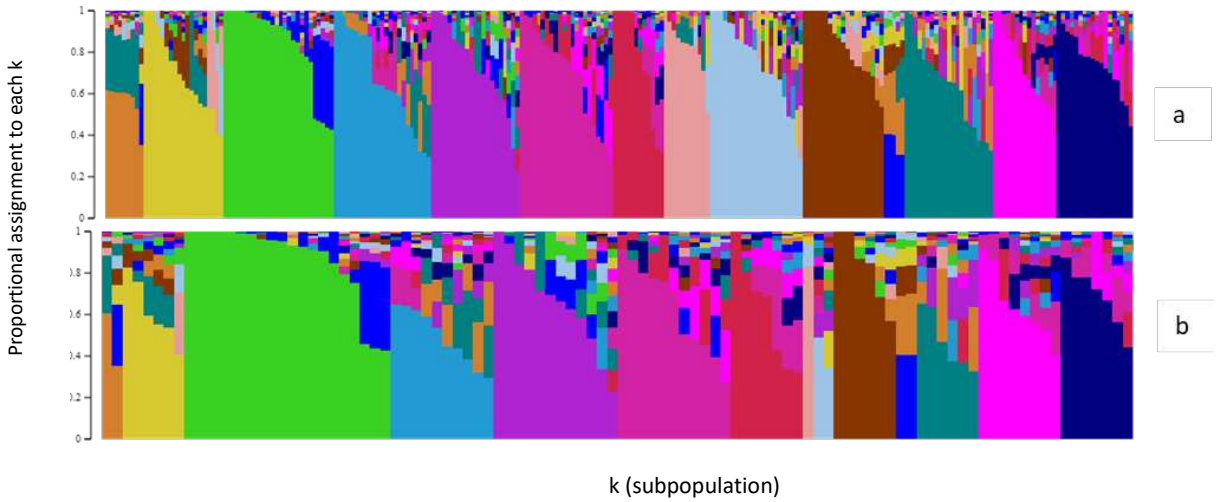

Figure 5.3 Model-based population structure for the Suffolk population for selection candidates (a) and optimal contribution selection (OCS) OCS 100 (b) populations, displaying the proportional assignment of each animal as a column and sorted by highest proportional assignment to a subpopulation

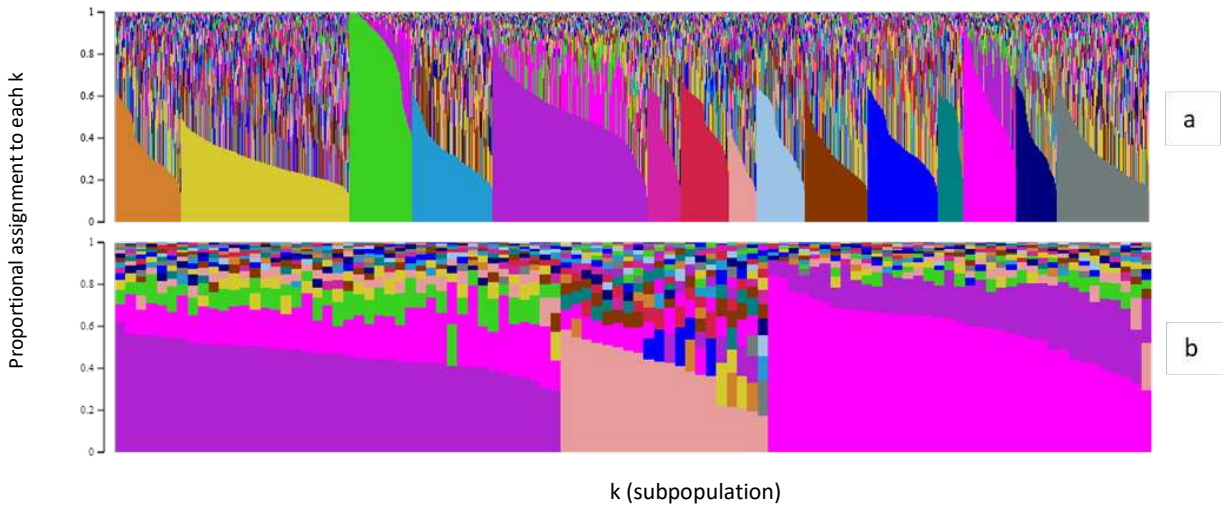

Figure 5.4 Model-based population structure for the Simmental population for selection candidates (a) and optimal contribution selection (OCS) OCS 100 (b) populations, displaying the proportional assignment of each animal as a column and sorted by highest proportional assignment to a subpopulation

Capturing the available alleles in the population is the most important measure of the success of the sampling strategies. Across the three populations, OCS 100 captured almost all available alleles and

only missed alleles at a low frequency in the population (Table 5.8). OCS 100 captured more than 99 percent of the available alleles in the population except for the simulated population, which had a high frequency of rare alleles in the population.

Table 5.8 Percentage of Simulated, Suffolk, and Simmental population alleles captured, alleles captured when minor allele frequency (MAF) > 0.01, and maximum frequency of missing alleles from the optimal contribution selection (OCS) OCS 100 sampling strategy

	Simulated population	Suffolk	Simmental
% Alleles Captured	96.45	99.33	99.93
% Alleles Captured MAF > 0.01	99.60	99.93	99.99
Max Frequency of Missing Alleles	0.05	0.05	0.02

5.4 Discussion

The three populations were characterized by 1) many purebred animals with a moderate size SNP chip in the simulated population, 2) few animals with a large SNP chip in the Suffolk population, and 3) many percentage animals with a moderate size SNP chip in the Simmental population. The populations had low levels of inbreeding, high levels of heterozygosity, and large N_e. The simulated population had a larger percentage of low frequency alleles, which were challenging to capture when sampling. Model-based population structure had distinct subpopulations for Suffolk with more admixture for the simulated and Simmental populations.

The OCS 100 failed to capture the phenotypic variation available for quantitative traits for both the simulated population and the Simmental population. The variation was, however, captured for the Suffolk population. The Suffolk population was small, resulting in 100 of the 244 animals from the full population being sampled for OCS 100. The OCS 100 sampling resulted in at least as much heterozygosity as in the full populations. However, the number of fixed alleles was higher for the OCS 100 samples than the full populations. This tended to result in a loss of rare alleles and an increase in

high frequency alleles. The percentages of ROH per class size were similar between each full population and the OCS 100 populations. For model-based population structure, the OCS 100 population reflected the subpopulations of the full population for the simulated and Suffolk populations, but not for the Simmental population.

The OCS 100 successfully captured most available alleles for each population. The sampling strategy captured more than 99 percent of the alleles for the Suffolk and Simmental population. It was less successful for the simulated population, which was characterized by high levels of low frequency alleles. The simulated population had 13,662 selection candidates and selecting 100 of the animals represents 0.7 percent of the total population. In a large population with many low frequency alleles, either more animals will need to be sampled or some loss of alleles will need to be accepted. Even though OCS 100 did not capture all available alleles in the population, the maximum MAF of the missing alleles was 0.05.

Information about the low frequency alleles in the population is needed to make informed decisions about sampling. If the alleles are decreasing in frequency in the population because they are associated with decreased fitness, the loss of those alleles may not be genetically important. This study evaluated capturing the genetic diversity of a population by selecting relatively few animals, which would be relevant in research or gene banking. For other purposes, the ability to select more animals would increase the opportunity to capture every allele.

Based on these results, a process for sampling animals is presented. First, the objective of the sampling needs to be defined as research, gene banking, reference population, or cull/keep animals. The objectives of the research will also determine the next step, which is to define the number of animals needed to sample. For gene banking, a recommendation of 50 to 100 males per breed is standard practice under current models (FAO, 2012; Blackburn, 2018). In contrast, development of a reference population for genomic selection may include a minimum of several thousand animals (Larroque et al.,

2014). The number of animals needed for research or keep/cull will vary. Then, the number of available animals needs to be determined and their pedigree records obtained. While the optiSel package in R was used with this research to compute optimal contributions, other software packages are available that employ the same methodology, including EVA software (Berg et al., 2006) and Gencont2 (Dagnachew and Meuwissen, 2016). The optimal contributions then provide a list of the most genetically unique combination of animals for use in research, gene banking, reference populations, or to meet other selection needs.

5.5 Conclusion

The three populations represented varying population sizes, number of markers, number of low frequency alleles, and breed history. Evidence of different breed histories can be seen by different N_e, ROH percentages by class size, and model-based population substructure. OCS 100 successfully sampled animals that are representative of the entire population using pedigree data for each of the three populations. OCS provides the opportunity to use pedigree data to effectively capture the genetic diversity available in a population and continues to be the “gold standard” for sampling animals.

LITERATURE CITED

Aberle, K. S., H. Hamann, C. Drögemüller , and O. Distl. 2004. Genetic diversity in German draught horse breeds compared with a group of primitive, riding and wild horses by means of microsatellite DNA markers. Animal Genetics doi: 10.1111/j.1365-2052.2004.01166.x

Ahbara, A., H. Bahbahani, F. Almathen, M. Al Abri, M. O. Agoub, A. Abeba, A. Kebede, H. H. Musa, S. Mastrangelo, F. Pilla, E. Ciani, O. Hanotte, and J. M. Mwacharo. 2018. Genome-wide variation, candidate regions and genes associated with fat deposition and tail morphology in Ethiopian indigenous sheep. Frontiers in genetics 9:699-699. doi: 10.3389/fgene.2018.00699

Al-Mamun, H. A., S. A Clark, P. Kwan, and C. Gondro. 2015. Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. Genetics Selection Evolution 47(1):90-90. doi: 10.1186/s12711-015-0169-6

Albrechtsen, A., F. C. Nielsen, and R. Nielsen. 2010. Ascertainment biases in SNP chips affect measures of population divergence. Molecular Biology and Evolution 27(11):2534-2547. doi: 10.1093/molbev/msq148

Alderson, G. L. H. 1992. A system to maximize the maintenance of genetic variability in small populations. In: G. L. H. Alderson and I. Bodo, editors, Genetic conservation of domestic livestock No. 2. CAB International, United Kindom. p. 18-29.

Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome research 19(9):1655-1664. doi: 10.1101/gr.094052.109

ASA. 2021. American Simmental Association. https://www.simmental.org/site/.

ASA Annual Report. 2020. American Simmental Association. https://simmental.org/site/userimages/minutes/Annual%20Report.pdf.

ASA Beef Briefs. 2021. American Simmental Association. https://www.simmental.org/site/userimages/History%20of%20the%20Simmental%20Breed.pdf .

Avendaño, S., B. Villanueva, and J. A. Woolliams. 2003. Expected increases in genetic merit from using optimized contributions in two livestock populations of beef cattle and sheep. Journal of Animal Science 81(12):2964-2975. doi: 10.2527/2003.81122964x

Banner Sheep Magazine. 2021. Past reports of registrations by purebred sheep associations Banner Sheep Magazine. Greg Deakin.

Berg, P., J. Nielsen, and M. K. Sørensen. 2006. EVA: Realized and predicted optimal genetic contributions. In: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Minas Gerais, Brazil. p 27-09.

Beynon, S. E., G. T. Slavov, M. Farré, B. Sunduimijid, K. Waddams, B. Davies, W. Haresign, J. Kijas, I. M. MacLeod, C. J. Newbold, L. Davies, and D. M. Larkin. 2015. Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. BMC Genetics 16(1):65-65. doi: 10.1186/s12863-015-0216-x

Biscarini, F., P. Cozzi, G. Gaspa, and G. Marras. 2019. detectRUNS: an R package to detect runs of homozygosity heterozygosity in diploid genomes. (Accessed February 1, 2023.

Blackburn, H. D. 2009. Genebank development for the conservation of livestock genetic resources in the United States of America. Livestock Science doi: 10.1016/j.livsci.2008.07.004

Blackburn, H. D. 2018. Biobanking genetic material for agricultural animal species. Annual Review of Animal Biosciences 6(1):69-82. doi: 10.1146/annurev-animal-030117-014603

Blackburn, H. D., S. R. Paiva, S. Wildeus, W. Getz, D. Waldron, R. Stobart, D. Bixby, P. H. Purdy, C. Welsh, S. Spiller, and M. Brown. 2011. Genetic structure and diversity among sheep breeds in the

united states: Identification of the major gene pools. Journal of Animal Science 89(8):2336-2348. doi: 10.2527/jas.2010-3354

Brito, L. F., M. Jafarikia, D. A. Grossi, J. W. Kijas, L. R. Porto-Neto, R. V. Ventura, M. Salgorzaei, and F. S. Schenkel. 2015. Characterization of linkage disequilibrium, consistency of gametic phase and admixture in Australian and Canadian goats. BMC Genetics 16(1)doi: 10.1186/s12863-015-0220-1

Brito, L. F., J. W. Kijas, R. V. Ventura, M. Sargolzaei, L. R. Porto-Neto, A. Cánovas, Z. Feng, M. Jafarikia, and F. S. Schenkel. 2017. Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers. BMC Genomics 18(1)doi: 10.1186/s12864-017-3610-0

Campbell, N. A., and J. B. Reece. 2008. Biology. 8th ed. Pearson Benjamin Cummings, San Francisco, CA.

Chen, G. K., P. Marjoram, and J. D. Wall. 2009. Fast and flexible simulation of DNA sequence data. Genome Research doi: 10.1101/gr.083634.108

Conner, J. K., and D. L. Hartl. 2004. A primer of ecological genetics.

Curik, I., M. Ferencakovic, B. Gredler, and J. Sölkner. 2010. Genome-wide heterozygosity and pedigree inbreeding coefficients in Simmental cattle population. In: 9th World Congress on Genetics Applied to Livestock Production, Leipzig

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. De Los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. doi: 10.1534/genetics.112.147983

Dagnachew, B. S., and T. H. Meuwissen. 2016. A fast Newton–Raphson based iterative algorithm for large scale optimal contribution selection. Genetics Selection Evolution 48:1-10.

Davenport, K. M., C. Hiemke, S. D. McKay, J. W. Thorne, R. M. Lewis, T. Taylor, and B. M. Murdoch. 2020. Genetic structure and admixture in sheep from terminal breeds in the United States. Animal Genetics 51(2):284-291. doi: 10.1111/AGE.12905

David, C. M. G., C. R. Quirino, W. H. O. Vega, A. Bartholazzi Junior, A. d. F. Madella-Oliveira, and R. L. D. Costa. 2018. Diversity of indigenous sheep of an isolated population. BMC Veterinary Research 14(1):350-350. doi: 10.1186/s12917-018-1682-y

Do, C., R. S. Waples, D. Peel, G. M. Macbeth, B. J. Tillett, and J. R. Ovenden. 2014. NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size from genetic data. Molecular Ecology Resources 14(1):209-214. doi: 10.1111/1755-0998.12157

Doekes, H. P., R. F. Veerkamp, P. Bijma, S. J. Hiemstra, and J. Windig. 2018. Value of the Dutch Holstein Friesian germplasm collection to increase genetic variability and improve genetic merit. Journal of Dairy Science doi: 10.3168/jds.2018-15217

Dolebo, A. T., N. Khayatzadeh, A. Melesse, D. Wragg, M. Rekik, A. Haile, B. Rischkowsky, M. F. Rothschild, and J. M. Mwacharo. 2019. Genome-wide scans identify known and novel regions associated with prolificacy and reproduction traits in a sub-Saharan African indigenous sheep (Ovis aries). Mammalian Genome doi: 10.1007/s00335-019-09820-5

Edea, Z., T. Dessie, H. Dadi, K. T. Do, and K. S. Kim. 2017. Genetic diversity and population structure of Ethiopian sheep populations revealed by high-density SNP markers. Frontiers in Genetics 8(DEC)doi: 10.3389/fgene.2017.00218

Eding, J. H., and G. Laval. 1999. Measuring the genetic uniqueness in livestock. In: J. K. Oldenbroek, editor, Genebanks and the management of farm animal genetic resources. DLO Institute for Animal Science and Health. p. 33-58.

Emenheiser, J., and D. Notter. 2011. LAMBPLAN terminal indexes for NSIP breeders.

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus, P. Bijma, and J. J. Windig. 2012. Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. Journal of Animal Breeding and Genetics 129(3):195-205. doi: 10.1111/j.1439-0388.2012.00987.x

Engelsma, K. A., R. F. Veerkamp, M. P. L. Calus, and J. J. Windig. 2011. Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. Journal of Animal Breeding and Genetics 128(6):473-481. doi: 10.1111/j.1439-0388.2011.00936.x

Eynard, S. E., P. Croiseau, D. Laloë, S. Fritz, M. P. L. Calus, and G. Restoux. 2018a. Which Individuals To Choose To Update the Reference Population? Minimizing the Loss of Genetic Diversity in Animal Genomic Selection Programs. G3: Genes, Genomes, Genetics 8(1):113-121. doi: 10.1534/G3.117.1117

Eynard, S. E., J. J. Windig, I. Hulsegge, S. J. Hiemstra, and M. P. L. Calus. 2018b. The impact of using old germplasm on genetic merit and diversity—A cattle breed case study. Journal of Animal Breeding and Genetics 135(4):311-322. doi: 10.1111/jbg.12333

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longman Group Ltd., Essex, UK.

FAO. 1998. FAO Secondary Guidelines for development of national farm animal genetic resources management plans.

FAO. 2012. Cryoconservation of Animal Genetic Resources.

FAO. 2015. The second report on the state of the world's FAO commission on genetic resources for food and agriculture assessments. 9789251088203.

FAS. 2021. USDA-Foreign Agricultural Service. https://apps.fas.usda.gov/psdonline/app/index.html#/app/advQuery.

Faux, A.-M., G. Gorjanc, R. C. Gaynor, M. Battagin, S. M. Edwards, D. L. Wilson, S. J. Hearne, S. Gonen, and J. M. Hickey. 2016. AlphaSim: software for breeding program simulation. The Plant Genome 9(3):0-0. doi: 10.3835/plantgenome2016.02.0013

Fernández, J., M. A. Toro, and A. Caballero. 2004. Managing individuals' contributions to maximize the allelic diversity maintained in small, conserved populations. Conservation Biology 18(5):1358-1367. doi: 10.1111/j.1523-1739.2004.00341.x

Fernández, J., M. A. Toro, F. Gómez-Romano, and B. Villanueva. 2016. The use of genomic information can enhance the efficiency of conservation programs. Animal Frontiers 6(1):59-64. doi: 10.2527/af.2016-0009

Francis, R. M. 2017. <scp>pophelper</scp>: an R package and web app to analyse and visualize population structure. Molecular Ecology Resources 17(1):27-32. doi: 10.1111/1755-0998.12509

Fröhlich, J., M. Vozdova, S. Kubickova, H. Cernohorska, H. Sebestova, and J. Rubes. 2015. Variation of meiotic recombination rates and MLH1 foci distribution in spermatocytes of cattle, sheep and goats. Cytogenetic and genome research 146(3):211-221.

Funk, D. A. 2006. Major advances in globalization and consolidation of the artificial insemination industry. Journal of Dairy Science 89(4):1362-1368. doi: 10.3168/jds.S0022-0302(06)72203-2

Gaynor, R. C., G. Gorjanc, and J. M. Hickey. 2021. AlphaSimR: an R package for breeding program simulations. G3 Genes|Genomes|Genetics 11(2)doi: 10.1093/G3JOURNAL/JKAA017

Gibbs, R. A., J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole, C. A. Gill, R. D. Green, D. L. Hamernik, S. M. Kappes, S. Lien, L. K. Matukumalli, J. C. McEwan, L. V. Nazareth, R. D. Schnabel, G. M. Weinstock, D. A. Wheeler, P. Ajmone-Marsan, P. J. Boettcher, A. R. Caetano, J. F. Garcia, O. Hanotte, P. Mariani, L. C. Skow, T. S. Sonstegard, J. L. Williams, B. Diallo, L. Hailemariam, M. L. Martinez, C. A. Morris, L. O. C. Silva, R. J. Spelman, W. Mulatu, K. Zhao, C. A. Abbey, M. Agaba, F. R. Araujo, R. J. Bunch, J. Burton, C. Gorni, H. Olivier, B. E. Harrison, B. Luff, M. A. Machado, J. Mwakaya, G. Plastow, W. Sim, T. Smith, M. B. Thomas, A. Valentini, P. Williams, J. Womack, J. A. Woolliams, Y. Liu, X. Qin, K. C. Worley, C. Gao, H. Jiang, S. S. Moore, Y. Ren, X.-Z. Song, C. D. Bustamante, R. D. Hernandez, D. M. Muzny, S. Patil, A. San Lucas, Q. Fu, M. P. Kent, R. Vega, A. Matukumalli, S. McWilliam, G. Sclep, K. Bryc, J. Choi, H. Gao, J. J. Grefenstette, B. Murdoch, A.

Stella, R. Villa-Angulo, M. Wright, J. Aerts, O. Jann, R. Negrini, M. E. Goddard, B. J. Hayes, D. G. Bradley, M. Barbosa Da Silva, L. P. L. Lau, G. E. Liu, D. J. Lynn, F. Panzitta, and K. G. Dodds. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324(5926):528-532. doi: 10.1126/science.1167936

Giontella, A., C. Pieramati, M. Silvestrelli, and F. M. Sarti. 2019. Analysis of founders and performance test effects on an autochthonous horse population through pedigree analysis: structure, genetic variability and inbreeding. Animal 13(1):15-24. doi: 10.1017/s1751731118001180

Glowatzki-Mullis, M. L., J. Muntwyler, W. Pfister, E. Marti, S. Rieder, P. A. Poncet, and C. Gaillard. 2006. Genetic diversity among horse populations with a special focus on the Franches-Montagnes breed. Animal Genetics 37(1):33-39. doi: 10.1111/j.1365-2052.2005.01376.x

Goddard, M. E. 2012. Uses of genomics in livestock agriculture. Animal Production Science 52(3):73. doi: 10.1071/an11180

Golden, B. L., W. M. Snelling, and C. H. Mallinckrodt. 1992. Animal breeders tool-kit: user's guide. Colorado State University, Exp. Sta. Tech. Bull. LTB92-2

Gorjanc, G., and J. M. Hickey. 2018. AlphaMate: a program for optimizing selection, maintenance of diversity and mate allocation in breeding programs. Bioinformatics 34(19):3408-3411. doi: 10.1093/bioinformatics/bty375

Gourdine, J. L., A. C. Sørensen, and L. Rydhmer. 2012a. There is room for selection in a small local pig breed when using optimum contribution selection: A simulation study. Journal of Animal Science 90(1):76-84. doi: 10.2527/jas.2011-3898

Gourdine, J. L., A. C. Sørensen, and L. Rydhmer. 2012b. There is room for selection in a small local pig breed when using optimum contribution selection: A simulation study1,2. Journal of Animal Science 90(1):76-84. doi: 10.2527/jas.2011-3898

Grasso, A. N., V. Goldberg, E. A. Navajas, W. Iriarte, D. Gimeno, I. Aguilar, J. F. Medrano, G. Rincón, and G. Ciappesoni. 2014. Genomic variation and population structure detected by single nucleotide polymorphism arrays in Corriedale, Merino and Creole sheep. Genetics and Molecular Biology 37(2):389-395. doi: 10.1590/S1415-47572014000300011

Greenbaum, G., A. R. Templeton, Y. Zarmi, and S. Bar-David. 2014. Allelic richness following population founding events - A stochastic modeling framework incorporating gene flow and genetic drift. PLoS ONE 9(12)doi: 10.1371/journal.pone.0115203

Gutiérrez, J. P., and F. Goyache. 2005. A note on ENDOG: a computer program for analysing pedigree information. Journal of Animal Breeding and Genetics 122(3):172-176. doi: 10.1111/J.1439-0388.2005.00512.X

Halliburton, R., and R. Halliburton. 2004. *Introduction to population genetics*. Pearson/Prentice Hall, Upper Saddle River, NJ.

Handley, L. J. L., K. Byrne, F. Santucci, T. S., T. M, B. M.W., and H. J.M. 2007. Genetic structure of European sheep breeds. Heredity 99:620-631.

Hickey, J. M., and G. Gorjanc. 2012. Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. doi: 10.1534/g3.111.001297

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38(6):226-231. doi: 10.1007/BF01245622

Hoggart, C. J., M. Chadeau-Hyam, T. G. Clark, R. Lampariello, J. C. Whittaker, M. De Iorio, and D. J. Balding. 2007. Sequence-level population simulations over large genomic regions. Genetics 177(3):1725-1731. doi: 10.1534/genetics.106.069088

Howard, D. M., R. Pong-Wong, P. W. Knap, V. D. Kremer, and J. A. Woolliams. 2014. The structural impact of implementing optimal contribution selection in a commercial pig breeding population. Proceedings of the World Congress on Genetics Applied to Livestock Production:024-024.

Howe, K. L., B. Contreras-Moreira, N. De Silva, G. Maslen, W. Akanni, J. Allen, J. Alvarez-Jarreta, M. Barba, D. M. Bolser, L. Cambell, M. Carbajo, M. Chakiachvili, M. Christensen, C. Cummins, A. Cuzick, P. Davis, S. Fexova, A. Gall, N. George, L. Gil, P. Gupta, K. E. Hammond-Kosack, E. Haskell, S. E. Hunt, P. Jaiswal, S. H. Janacek, P. J. Kersey, N. Langridge, U. Maheswari, T. Maurel, M. D. McDowall, B. Moore, M. Muffato, G. Naamati, S. Naithani, A. Olson, I. Papatheodorou, M. Patricio, M. Paulini, H. Pedro, E. Perry, J. Preece, M. Rosello, M. Russell, V. Sitnik, D. M. Staines, J. Stein, M. K. Tello-Ruiz, S. J. Trevanion, M. Urban, S. Wei, D. Ware, G. Williams, A. D. Yates, and P. Flicek. 2020. Ensembl Genomes 2020—enabling non-vertebrate genomic research. Nucleic Acids Research 48(D1):D689-D695. doi: 10.1093/NAR/GKZ890

Hozé, C., M. N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genetics Selection Evolution 45(1)doi: 10.1186/1297-9686-45-33

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

Illumina. 2015. Illumina Data Sheet: Agrigenomics. https://www.illumina.com/documents/products/datasheets/datasheet_bovineHD.pdf.

Jakobsson, M., and N. A. Rosenberg. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23(14):1801-1806. doi: 10.1093/bioinformatics/btm233

Karimi, K., A. Esmailizadeh Koshkoiyeh, M. Asadi Fozi, L. R. Porto-Neto, and C. Gondro. 2016. Prioritization for conservation of Iranian native cattle breeds based on genome-wide SNP data. Conservation Genetics 17(1):77-89. doi: 10.1007/s10592-015-0762-9

Kelleher, M. M., D. P. Berry, J. F. Kearney, S. McParland, F. Buckley, and D. C. Purfield. 2017. Inference of population structure of purebred dairy and beef cattle using high-density genotype data. Animal 11(1):15-23. doi: 10.1017/S1751731116001099

Kijas, J. W., T. Hadfield, M. Naval Sanchez, and N. Cockett. 2016. Genome-wide association reveals the locus responsible for four-horned ruminant. Animal Genetics 47(2):258-262. doi: 10.1111/age.12409

Kijas, J. W., J. A. Lenstra, B. Hayes, S. Boitard, L. R. Neto, M. S. Cristobal, B. Servin, R. McCulloch, V. Whan, K. Gietzen, S. Paiva, W. Barendse, E. Ciani, H. Raadsma, J. McEwan, and B. Dalrymple. 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biology 10(2)doi: 10.1371/journal.pbio.1001258

Kijas, J. W., L. Porto-Neto, S. Dominik, A. Reverter, R. Bunch, R. McCulloch, B. J. Hayes, R. Brauning, and J. McEwan. 2014. Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. Animal Genetics 45(5):754-757. doi: 10.1111/age.12197

Kirschten, D. P., D. R. Notter, T. D. Leeds, M. R. Mousel, J. B. Taylor, and G. S. Lewis. 2013. Evaluation of Columbia, USMARC-Composite, Suffolk, and Texel rams as terminal sires in an extensive rangeland production system: V. Postweaning growth, feed intake, and feed efficiency1. Journal of Animal Science 91(5):2021-2033. doi: 10.2527/jas.2012-5152

Kopelman, N. M., J. Mayzel, M. Jakobsson, N. A. Rosenberg, and I. Mayrose. 2015. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Molecular ecology resources 15(5):1179-1191. doi: 10.1111/1755-0998.12387

Kristensen, T. N., A. A. Hoffmann, C. Pertoldi, and A. V. Stronen. 2015. What can livestock breeders learn from conservation genetics and vice versa? Frontiers in Genetics No. 5. Frontiers Research Foundation.

Kuehn, L. A., J. W. Keele, G. L. Bennett, T. G. McDaneld, T. P. L. Smith, W. M. Snelling, T. S. Sonstegard, and R. M. Thallman. 2011. Predicting breed composition using breed frequencies of 50,000

markers from the US Meat Animal Research Center 2,000 bull project. Journal of Animal Science doi: 10.2527/jas.2010-3530
Lacy, R. C. 2000. Should we select genetic alleles in our conservation breeding programs? Zoo Biology 19(4):279-282. doi: 10.1002/1098-2361(2000)19:4<279::AID-ZOO5>3.0.CO;2-V
Larroque, H., F. Barillet, G. Baloche, J. M. Astruc, D. Buisson, F. Shumbusho, V. Clément, G. Lagriffoul, I. Palhière, R. Rupp, C. Carillier, C. Robert-Granié, and A. Legarra. 2014. Toward genomic breeding programs in French dairy sheep and goats. Proceedings, 10th World Congress of Genetics Applied to Livestock Production Vancouver
Leroy, G., T. Mary-Huard, E. Verrier, S. Danvy, E. Charvolin, and C. Danchin-Burge. 2013. Methods to estimate effective population size using pedigree data: Examples in dog, sheep, cattle and horse. Genetics Selection Evolution 45(1):1-1. doi: 10.1186/1297-9686-45-1
Leroy, G., and X. Rognon. 2012. Assessing the impact of breeding strategies on inherited disorders and genetic diversity in dogs. Veterinary Journal doi: 10.1016/j.tvjl.2012.06.025
Machová, K., M. Milerski, J. Rychtářová, B. Hofmanová, H. Vostrá-Vydrová, N. Moravčíková, R. Kasarda, and L. Vostrý. 2021. Assessment of the genetic diversity of Two Czech autochthonous sheep breeds. Small Ruminant Research 195
Makina, S. O., J. F. Taylor, E. Van Marle-Köster, F. C. Muchadeyi, M. L. Makgahlela, M. D. MacNeil, and A. Maiwashe. 2015. Extent of linkage disequilibrium and effective population size in four South African sanga cattle breeds. Frontiers in Genetics 6(DEC)doi: 10.3389/fgene.2015.00337
Martinez, A. M., J. L. Vega-Pla, J. M. Leon, M. E. Camacho, J. V. Delgado, and M. N. Ribeiro. 2010. Is the Murciano-Granadina a single goat breed? A molecular genetics approach. Arquivo Brasileiro de Medicina Veterinária e Zootecnia 62:1191-1198.
Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. L. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4(4)doi: 10.1371/journal.pone.0005350
McManus, C., O. Facó, L. Shiotsuki, J. L. J. P. Rolo, and V. Peripolli. 2019. Pedigree analysis of Brazilian Morada Nova hair sheep. Small Ruminant Research 170:37-42.
McManus, C., S. A. Santos, B. S. L. Dallago, S. R. Paiva, R. F. S. Martins, J. Braccini Neto, P. R. Marques, and U. G. P. D. Abreu. 2013. Evaluation of conservation program for the Pantaneiro horse in Brazil. Revista Brasileira de Zootecnia 42:404-413.
McPeek, M. S., and T. P. Speed. 1995. Modeling Interference in Genetic Recombination.
Melka, M. G., K. Stachowicz, F. Miglior, and F. S. Schenkel. 2013. Analyses of genetic diversity in five Canadian dairy breeds using pedigree data. Journal of Animal Breeding and Genetics 130(6):476-486. doi: 10.1111/jbg.12050
Menezes, L. M., W. H. Sousa, E. P. C. Filho, F. Q. Cartaxo, J. A. Viana, and L. T. Gama. 2015. Genetic variability in a nucleus herd of Boer goats in Brazil assessed by pedigree analysis. Small Ruminant Research 131:85-92.
Meuwissen, T. 2009. Genetic management of small populations: A review. Acta Agriculturae Scandinavica A: Animal Sciences 59(2):71-79. doi: 10.1080/09064700903118148
Meuwissen, T. H. 1997. Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science 75(4):934-934. doi: 10.2527/1997.754934x
Meuwissen, T. H. E., and J. A. Woolliams. 1994. Effective sizes of livestock populations to prevent a decline in fitness. Theoretical and Applied Genetics 89(7-8):1019-1026. doi: 10.1007/BF00224533
Miller, L. R., J. Stepanek Shiflett, D. J. Marsh, and P. Rodgers. 2016. U.S. Sheep Industry Research, Development, and Education Priorities.

NASS. 2021. National Agriculture Statistics Service (NASS). https://www.nass.usda.gov/Data_and_Statistics/ (Accessed December 1 2021).

Neuditschko, M., H. W. Raadsma, M. S. Khatkar, E. Jonas, E. J. Steinig, C. Flury, H. Signer-Hasler, M. Frischknecht, R. Von Niederhäusern, T. Leeb, and S. Rieder. 2017. Identification of key contributors in complex population structures. PLoS ONE 12(5)doi: 10.1371/journal.pone.0177638

Nicoloso, L., L. Bomba, L. Colli, R. Negrini, M. Milanesi, R. Mazza, T. Sechi, S. Frattini, A. Talenti, B. Coizet, S. Chessa, D. Marletta, M. D'Andrea, S. Bordonaro, G. Ptak, A. Carta, G. Pagnacco, A. Valentini, F. Pilla, P. Ajmone-Marsan, P. Crepaldi, and The Italian Goat Consortium. 2015. Genetic diversity of Italian goat breeds assessed with a medium-density SNP chip. Genetics Selection Evolution 47(1):62-62. doi: 10.1186/s12711-015-0140-6

NIH. 2004. NIH News Advisory. https://www.genome.gov/12512874/2004-release-bovine-genome-assembled.

NSIP. 2021. National Sheep Improvement Program. http://nsip.org/.

Oldenbroek, J. K. 1999. Genebanks and the management of farm animal genetic resources. isbn9075124066.

Oliveira, R. R., L. H. A. Brasil, J. V. Delgado, J. Peguezuelos, J. M. León, D. G. P. Guedes, J. K. G. Arandas, and M. N. Ribeiro. 2016. Genetic diversity and population structure of the Spanish Murciano–Granadina goat breed according to pedigree data. Small Ruminant Research 144:170-175.

Ott, R. L., and M. T. Longnecker. 2015. An introduction to statistical methods and data analysis.

Paiva, S. R., A. D. S. Mariante, and H. D. Blackburn. 2011. Combining US and Brazilian microsatellite data for a meta-analysis of sheep (Ovis aries) breed diversity: Facilitating the FAO global plan of action for conserving animal genetic resources. Journal of Heredity 102(6):697-704. doi: 10.1093/jhered/esr101

Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li. 2012. Genomic patterns of homozygosity in worldwide human populations. American Journal of Human Genetics doi: 10.1016/j.ajhg.2012.06.014

Pfaff, B. 2014. The R package cccp: design for solving cone constrained convex programs. R Finance:16-17.

Plante, Y., J. L. Vega-Pla, Z. Lucas, D. Colling, B. De March, and F. Buchanan. 2007. Genetic diversity in a feral horse population from Sable Island, Canada. Journal of Heredity 98(6):594-602.

Porto-Neto, L. R., J. W. Kijas, and A. Reverter. 2014. The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genetics Selection Evolution 46(1):1-5. doi: 10.1186/1297-9686-46-22/FIGURES/2

Posbergh, C. J., M. L. Thonney, and H. J. Huson. 2019. Genomic approaches identify novel gene associations with out of season lambing in sheep. Journal of Heredity 110(5):577-586. doi: 10.1093/jhered/esz014

Prystupa, J. M., R. Juras, E. G. Cothran, F. C. Buchanan, and Y. Plante. 2012. Genetic diversity and admixture among Canadian, Mountain and Moorland and Nordic pony populations. Animal 6(1):19-30. doi: 10.1017/S1751731111001212

Pszczola, M., T. Strabel, and M. P. L. Calus. 2014. Proceedings, 10th World Congress of Genetics Applied to Livestock Production Size of required reference population updates to achieve constant genomic prediction accuracy across generations. 91:44144423.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. De Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81(3):559-575. doi: 10.1086/519795

Purfield, D. C., D. P. Berry, S. McParland, and D. G. Bradley. 2012. Runs of homozygosity and population history in cattle. BMC Genetics 13doi: 10.1186/1471-2156-13-70

Ramasamy, R., S. Ramasamy, B. Bindroo, and V. Naik. 2014. STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. SpringerPlus 3(1):431-431. doi: 10.1186/2193-1801-3-431

Rosegrant, M. W., M. Fernández, A. Sinha, J. Alder, H. Ahammad, C. d. Fraiture, B. Eickhout, J. Fonseca, J. Huang, O. Koyama, A. M. Omezzine, P. L. Pingali, R. Ramírez, C. Ringler, S. Robinson, P. K. Thornton, D. v. Vuuren, and H. Yana-Shapiro. 2009. Looking into the future for agriculture and AKST.

Ruane, J. 1999. Selecting breeds for conservation. In: J. K. Oldenbroek, editor, Genebanks and the Conservation of Farm Animal Genetic Resources. DLO Institute for Animal Science and Health. p. 59-73.

Saad, H. M., M. G. Thomas, S. E. Speidel, R. K. Peel, W. M. Frasier, and R. M. Enns. 2020. Differential response from selection for high calving ease vs. low birth weight in American Simmental beef cattle. Journal of Animal Science 98(7):skaa162.

Sánchez-Molano, E., D. Tsiokos, D. Chatziplis, H. Jorjani, L. Degano, C. Diaz, A. Rossoni, H. Schwarzenbacher, F. Seefried, L. Varona, D. Vicario, E. L. Nicolazzi, and G. Banos. 2016. A practical approach to detect ancestral haplotypes in livestock populations. BMC Genetics 17(1):91-91. doi: 10.1186/s12863-016-0405-2

Saura, M., A. PÉRez-Figueroa, J. FernÁNdez, M. A. Toro, and A. Caballero. 2008. Preserving Population Allele Frequencies in Ex Situ Conservation Programs. Conservation Biology 22(5):1277-1287. doi: 10.1111/j.1523-1739.2008.00992.x

Skaarud, A., J. A. Woolliams, and H. M. Gjøen. 2011. Strategies for controlling inbreeding in fish breeding programs; an applied approach using optimum contribution (OC) procedures. Aquaculture 311(1-4):110-114. doi: 10.1016/J.AQUACULTURE.2010.11.023

Spangler, G. L., B. D. Rosen, M. B. Ilori, O. Hanotte, E. S. Kim, T. S. Sonstegard, J. M. Burke, J. L. M. Morgan, D. R. Notter, and C. P. Van Tassell. 2017. Whole genome structural analysis of Caribbean hair sheep reveals quantitative link to West African ancestry. PLoS ONE 12(6)doi: 10.1371/journal.pone.0179021

Stepanek Shiflett, J. 2017. American Sheep Industry Association U.S. sheep industry economic impact study.

Stronen, A. V., C. Pertoldi, L. Iacolina, H. N. Kadarmideen, and T. N. Kristensen. 2019. Genomic analyses suggest adaptive differentiation of northern European native cattle breeds. Evolutionary Applications 12(6):1096-1113. doi: 10.1111/eva.12783

Templeton, A. R. 2021. Population genetics and microevolutionary theory. John Wiley & Sons.

Thaon D'arnoldi, C., J.-L. Foulley, and L. Ollivier. 1998. An overview of the Weitzman approach to diversity. Genetics Selection Evolution 30:149-161.

Thornton, P. K. 2010. Livestock production: Recent trends, future prospects Philosophical Transactions of the Royal Society B: Biological Sciences No. 365. p 2853-2867. Royal Society.

Toro, M., and A. Mäki-Tanila. 1999. Establishing a conservation scheme. In: J. K. Oldenbroek, editor, Genebanks and the management of farm animal genetic resources. DLO Institute for Animal Science and Health. p. 75-90.

USSA. 2021. United Suffolk Sheep Association. https://suffolks.org/.

van Breukelen, A. E., H. P. Doekes, J. J. Windig, and K. Oldenbroek. 2019. Characterization of Genetic Diversity Conserved in the Gene Bank for Dutch Cattle Breeds. Diversity 2019, Vol. 11, Page 229 11(12):229-229. doi: 10.3390/D11120229

van der Werf, J. H. J., R. G. Banks, S. A. Clark, S. J. Lee, H. D. Daetwyler, B. J. Hayes, and A. A. Swan. 2014. Genomic Selection in Sheep Breeding Programs. Proceedings, 10th World Congress of Genetics Applied to Livestock Production Vancouver, Canada

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 Genes|Genomes|Genetics 11(11)doi: 10.1093/G3JOURNAL/JKAB231

Visser, C., S. F. Lashmar, E. Van Marle-Köster, M. A. Poli, and D. Allain. 2016. Genetic diversity and population structure in South African, French and Argentinian Angora goats from genome-wide SNP data. PLoS ONE 11(5)doi: 10.1371/journal.pone.0154353

Waples, R. S., and C. Do. 2008. ldne: a program for estimating effective population size from data on linkage disequilibrium. Molecular Ecology Resources 8(4):753-756. doi: 10.1111/j.1755-0998.2007.02061.x

Weigel, K. A. 2001. Controlling inbreeding in modern breeding programs. Journal of Dairy Science doi: 10.3168/jds.s0022-0302(01)70213-5

Weitzman, M. L. 1992. On Diversity. The quarterly journal of economics. 107(2):363-405. doi: 10.2307/2118476

Wellmann, R. 2019. Optimum contribution selection for animal breeding and conservation: the R package optiSel. BMC Bioinformatics 20(1)doi: 10.1186/s12859-018-2450-5

Wellmann, R., S. Hartwig, and J. Bennewitz. 2012. Optimum contribution selection for conserved populations with historic migration. Genetics Selection Evolution 44(1):34-34. doi: 10.1186/1297-9686-44-34

Whitacre, L. K., and M. L. Spangler. 2012. The Simmental breed: population structure and generation interval trends. Nebraska Beef Cattle Report.

Wiggans, G. R., T. A. Cooper, P. M. Vanraden, C. P. Van Tassell, D. M. Bickhart, and T. S. Sonstegard. 2016. Increasing the number of single nucleotide polymorphisms used in genomic evaluation of dairy cattle. Journal of Dairy Science 99(6):4504-4511. doi: 10.3168/jds.2015-10456

Wilson, C. S., J. L. Petersen, H. D. Blackburn, and R. M. Lewis. 2022. Assessing population structure and genetic diversity in US Suffolk sheep to define a framework for genomic selection. Journal of Heredity 113(4):431-443. doi: 10.1093/jhered/esac026

Windig, J. J., and H. P. Doekes. 2018. Limits to genetic rescue by outcross in pedigree dogs. Journal of Animal Breeding and Genetics 135(3):238-248. doi: 10.1111/jbg.12330

Windig, J. J., H. Eding, L. Moll, and L. Kaal. 2004. Effects on inbreeding of different strategies aimed at eliminating scrapie sensitivity alleles in rare sheep breeds in The Netherlands.

Windig, J. J., and K. A. Engelsma. 2010. Perspectives of genomics for genetic conservation of livestock. Conservation Genetics 11(2):635-641. doi: 10.1007/s10592-009-0007-x

Windig, J. J., and K. Oldenbroek. 2015. Genetic management of Dutch golden retriever dogs with a simulation tool. Journal of Animal Breeding and Genetics 132(6):428-440. doi: 10.1111/jbg.12149

Wright, S. 1931. Evolution in Mendelian populations.97-159.

Wright, S. 1951. The genetical structure of populations. Annals of Eugenics 15(1):323-354. doi: 10.1111/j.1469-1809.1949.tb02451.x

Wyoming Livestock Roundup. 2021. Sheep Industry Focuses on Genetic Improvement. https://www.wylr.net/2013/12/21/sheep-industry-focuses-on-genetic-improvement/.

Zhang, L., M. R. Mousel, X. Wu, J. J. Michal, X. Zhou, B. Ding, M. V. Dodson, N. K. El-Halawany, G. S. Lewis, and Z. Jiang. 2013. Genome-wide genetic diversity and differentially selected regions among Suffolk, Rambouillet, Columbia, Polypay, and Targhee sheep. PLoS ONE 8(6):e65942-e65942. doi: 10.1371/journal.pone.0065942