Effects of Sample Size and Single- vs. Multiple-Breed Analyses on Characterizing Runs of Homozygosity in Dairy Cattle

by

Calista Louise Vogelzang

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Animal Biosciences

Guelph, Ontario, Canada

ABSTRACT

EFFECTS OF SAMPLE SIZE AND SINGLE- VS. MULTIPLE-BREED ANALYSES ON CHARACTERIZING RUNS OF HOMOZYGOSITY IN DAIRY CATTLE

Calista Louise Vogelzang
University of Guelph, 2018

Advisor:
Dr. Christine Baes

The understanding and management of inbreeding is an ongoing challenge in the dairy industry. Runs of homozygosity (**ROH**) can provide more accurate estimates of the realized proportion of the genome shared between individuals. The objective of this thesis was to characterize ROH in dairy cattle breeds using different ROH detection methods.

Data was prepared using PLINK and R, and BCFtools was the used to detect ROH. The R packages *detectRUNS* and *pedigree* were then used to characterize ROH and to calculate pedigree-based inbreeding respectively. SNP inclusion in single- and multiple-breed analyses impacted the degree of breed differences observed. Sample size was shown to affect the accuracy of ROH characterization. The results from this study indicate ROH could potentially be used manage inbreeding levels. Future research will aim to develop a method to minimize breed bias in multiple-breed analyses and to understand breed differences.

ACKNOWLEDGEMENTS

As this chapter in life comes to a close, I would like to take the opportunity to acknowledge the widespread team I have had to pleasure of working with over the past two years.

First and foremost, I would like to sincerely thank my advisor, Dr. Christine Baes. Christine, you have pushed me to be an engaged student, a critical thinker, and a responsible researcher. Through your guidance, I have achieved more than I could have imagined at the start of this journey. Thank you for investing in me, believing in me, and providing me with such incredible opportunities to grow.

Thank you to Dr. Nina Melzer and the Leibniz Institute for Farm Animal Biology in Dummerstorf, Germany, for hosting my research stay in the fall of 2017. Nina, I cannot express the enormity of my gratitude for the mentorship you gave me while I was at FBN. I would not have been able to climb this mountain if you had not taken the time to help me learn and practice coding in R. Thank you for encouraging me to appreciate the importance of truly understanding my data at every stage of analyses and instilling in me the value of organization and reflection when conducting my research.

To my remaining advisory committee members, Dr. Filippo Miglior, Dr. Mehdi Sargolzaei, and Dr. Christian Maltecca, I am truly appreciative of your support and time your have provided me over the past two years.

This work would not have been possible without the support of our funding partners, the Canadian Dairy Network and the Natural Sciences and Engineering Research Council. Thank you for supporting our thirst of knowledge and our love of science.

I have been privileged to work with Dr. Gabriele Marras and Bayode Makanjuola on this project. Gabriele, I cannot express how grateful I am for your time and your patience as I developed my coding skills and gained a better understanding of bioinformatics. Bayode, thank you for always being willing to sit with me as I worked to develop a deeper understanding of different topics in genetics. Gentlemen, this journey would not have been the same without you, and your guidance will not be forgotten.

To my peers, thank you for your support and encouragement through all the ups and downs. A special thank you to Stephanie Lam, for being my rock, my cheerleader, and my role model. To my Diamond family, thank you for the laughs, the love, and the dogs. You are the reason I still have (most) of my sanity.

I am extremely fortunate to have amazing friends and family who have walked this path with me. Although they may not understand why I do the things I do (and really, I question it myself sometimes), they are always willing to listen and offer advice when I need it.

And finally, thank you to my parents. Mom, Dad, Chris, and Sandra, you have taught me the value of hard work, perseverance, flexibility, and humility. Thank you for giving me deep roots and allowing me to grow. This thesis would not have been possible without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

AI: Artificial insemination

AY: Ayrshire

BS: Brown Swiss

F_{PED}: Pedigree-based inbreeding

F_{ROH}: Runs of homozygosity-based inbreeding

GU: Guernsey

h^2: heritability

HD: High density

HO: Holstein

IBD: Identical by descent

IBS: Identical by state

JE: Jersey

MAF: Minor allele frequency

QTL: Quantitative trait loci

ROH: Runs of homozygosity

SE: Standard error

SNP: Single nucleotide polymorphism

VCF: Variant call format

WGS: Whole genome sequencing

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION AND OBJECTIVES

INTRODUCTION

Improved inbreeding management has been an ongoing goal in the dairy industry. Traditional methods of calculating inbreeding, such as using pedigree information, have been found to underestimate the true level of homozygosity that is identical-by-descent (autozygosity) of an individual, as pedigree-based inbreeding **(F_{PED})** is limited by factors such as pedigree depth and correct record keeping (McQuillan *et al.*, 2008; Ferenčaković *et al.*, 2013a). F_{PED} assumes that founder animals are not related, and assumes Mendelian sampling, which cannot predict which 50% of the maternal and 50% of the paternal genome will be passed to offspring. The sequencing of the bovine genome, along with decreasing genotyping costs, led to the implementation of genomic selection in dairy cattle in 2009 (Keller *et al.*, 2011; VanRaden *et al.*, 2011; Silva *et al.*, 2014).

Genomics has allowed for a more accurate estimation of the realized proportion of the genome that is shared between individuals. The increased reliability (up to 80%) of genomic estimated breeding values allowed young, unproven bulls to be considered as artificial insemination (**AI**) candidates, decreasing the cost of progeny testing by 92% as well as shortening the generation interval (Schaeffer, 2006; Hayes *et al.*, 2009; VanRaden *et al.*, 2011; Blasco and Toro, 2014; Weller *et al.*, 2017). This shortened generation interval resulted in a dramatic increase in the rate of genetic change for key traits of interest such as milk yield, percent fat, and percent protein. Traits with low heritabilities (**h^2**, i.e. health and fertility traits) have also benefited (Weller *et al.*, 2017). However, the dairy industry has seen an increased loss in genetic diversity along with an increase in the rate of inbreeding per year (Canadian Dairy Network, 2018a). Higher levels of inbreeding have resulted in reduced levels of fitness in individuals, and loss of genetic diversity

negatively effects the ability of the dairy industry to adjust to changes is market demands (Keller *et al.* 2011). Therefore, there is a need for alternative methods to manage inbreeding levels without compromising genetic gain in the dairy industry (Keller *et al.*, 2011).

Runs of homozygosity (**ROH**), which are homozygous regions in the genome, have been shown to be a more accurate method of calculating the level of autozygosity of an individual when compared to F_{PED} (Stachowicz *et al*., 2011; Ferenčaković *et al.*, 2011; Ferenčaković *et al.*, 2013a). In addition to being used to calculate ROH-based inbreeding (**F_{ROH}**), ROH have been used to investigate population histories, as well as to find and explore breed differences within a species (McQuillan *et al.*, 2008; Kirin *et al.*, 2010; Szmatoła *et al.*, 2016). ROH have been used in human disease and population history studies for over a decade; however, as ROH would most likely be used to improve mating decisions in the dairy industry, it is important to develop a standard definition and method of identification of ROH (Howrigan *et al.*, 2011; Curik *et al.*, 2014). It is important to understand how to calculate ROH, and to investigate how different factors influence their identification and analysis (Kirin *et al*., 2010; Purfield *et al.*, 2012; Ferenčaković *et al.*, 2013b; Curik *et al.*, 2014). Therefore, the purpose of this research was to investigate how different methods of ROH detection and analysis effect ROH identification and characterization in different breeds of dairy cattle. This was done by examining differences between single- and multiple-breed analyses with similar and dissimilar population sizes, as well as investigating the impact of sample size on the accuracy of ROH detection and characterization using the program BCFtools.

OBJECTIVES

The main objective of this research was to observe how different methods of analysis influenced ROH identification and characterization in in five dairy cattle breeds as this information could be used by breeders in the future to make more informed mating decisions. Specific objectives were to:

1. Evaluate how BCFtools can be used to identify ROH;
2. Evaluate the effect of single- and multiple-breed analyses on ROH detection;
3. Evaluate the impact of small, medium, and large sample sizes on ROH characterization.

LITERATURE REVIEW

Genomic selection and inbreeding in the Canadian Dairy Industry

The dairy industry represents a major economic sector in Canada and is comprised of approximately 945,000 milking cows on 10,951 farms across the country. (Canadian Dairy Information Centre, 2017a). Canadian dairy production mainly uses seven dairy breeds, including Holstein, Jersey, Ayrshire, Brown Swiss, Guernsey, Canadienne, and Milking Shorthorn (Canadian Dairy Information Centre, 2017b). Holstein is the most commonly used breed in Canada, making up over 90%of the Canadian dairy cow population in 2017 due to the breed's high level of production, followed by Jersey (~4%), and Ayrshire (~2%) (Canadian Dairy Information Centre, 2017b).

Canadian genetics are known to be of select quality due to the industry's genetic evaluation and improvement programs, and thus are valued worldwide. The export of Canadian dairy genetics in the form of embryos, semen, and live animals totalled $148.9 million in 2017 alone (Canadian Dairy Information Centre, 2018b). In the same year, Canada imported $65.3 million worth of genetics, the majority from the United States of America (99.4%), in addition to Spain, Australia, Denmark, and Germany (Canadian Dairy Information Centre, 2018b).

Genomic selection was first described by Meuwissen *et al.* in 2001 but was not implemented in Canada until 2009 which was when the bovine genome was sequenced, and the cost of genotyping was declining (Silva *et al.*, 2014). In 2006, Schaeffer stated that genomic selection would double the rate of genetic gain, decrease the cost of progeny-testing bulls by 92%, and allow for young animals to be selected using their genomic estimated breeding values (reliability up to 80%) (Schaeffer, 2006; Hayes *et al.*, 2009). It cost roughly $400 to genotype an animal in 2006; as of 2014, it was estimated to cost $50 to genotype an animal with a medium

(50K) density array, and genotyping costs are expected to decrease further in the coming years (Schaeffer, 2006, Blasco and Toro, 2014).

Since 2009, the dairy industry has seen a dramatic increase in genetic gain not only for production traits, which have moderate to high heritablities (**h^2**), but also in traits with low h^2, such as health and fertility traits (Blasco and Toro, 2014; Weller *et al*., 2017). With this increase in genetic gain, however, the dairy industry also saw an increased loss in genetic variability and an increased rate of inbreeding (Canadian Dairy Network, 2018a).

According to the Canadian Dairy Network, the average inbreeding percentage per breed (based on pedigree information) as of 2017 ranged from 2.75% (Milking Short Horn) to 8.75% (Canadienne) (Canadian Dairy Network, 2018a). Apart from the Guernsey and Canadienne breeds, in which the level of inbreeding decreased 0.09% and 0.02% between 2010-2017, the level of inbreeding within each breed increased between 2010-2017, with the largest increase being in the Holstein breed (+0.23%) (Canadian Dairy Network, 2018a). Additionally, the rate of inbreeding in the Canadian Brown Swiss population has been decreasing over the past two years, most likely due to an increased use of European bulls (Canadian Dairy Network, 2018a). Increases in overall inbreeding levels observed across breeds has been greatly influenced largely by genomic selection, specifically through the shortening of the generation interval of replacement animals.

Prior to genomic selection, animals were selected for breeding programs based on their estimated breeding values, which were calculated based on measured phenotypes and a weighted selection index. Sires were required to be progeny tested, meaning the performance of 100 daughters across multiple herds was measured to estimate the estimated breeding value (**EBV**) of a candidate bull for various key traits (Boichard *et al.*, 2016). This method yielded superior bulls with EBV that had high reliabilities, however it was a lengthy and costly process (Boichard *et al.*,

2016). When genomic selection started to be used, the process of choosing bulls for artificial insemination (**AI**) programs shifted from progeny testing to selecting unproven bulls based on their genomic estimated breeding values. Genomic selection revolutionized the dairy industry. As of 2017, herds employing AI were predominately using unproven bulls between the ages of 2 to 4 years old (44.1%), followed by proven bulls aged 5 to 8 years (27.6%), unproven bulls under 2 years of age (23.3%), and proven bulls 9+ years of age (4.9%) (Canadian Dairy Information Centre, 2018a). In comparison, herds employing AI in 2009 were predominately using proven bulls 5 to 8 years of age (47.6%), followed by unproven bulls under 2 years of age (30.7%), proven bulls 9+ years of age (11.2%), and unproven bulls 2 to 4 years of age (10.5%) (Canadian Dairy Information Centre, 2018a)

Defining Inbreeding

The management of inbreeding has been an important goal in the dairy industry. In order to manage inbreeding, the mechanisms behind inbreeding must be understood. Inbreeding is a complex topic, and several methods have been developed to calculate the level of inbreeding of an individual (Howard *et al*., 2017).

Traditionally, inbreeding has been calculated using pedigree information and Mendelian sampling probabilities. Pedigree-based inbreeding (**F_{PED}**), does, however, have several limitations that affect its accuracy (Purfield *et al*., 2012; Ferenčaković *et al.*, 2013a; Purfield *et al*., 2017). Pedigree-based inbreeding assumes that all founders of a population are unrelated, which may not be necessarily true (Purfield *et al*., 2012; Purfield *et al*., 2017). Additionally, F_{PED} relies on the depth of the pedigree and correct record keeping practices and does not consider the random nature of recombination. (Purfield *et al*., 2012; Ferenčaković *et al.*, 2013a; Purfield *et al*., 2017).

Therefore, the use of genomic data is becoming a popular method to calculate inbreeding more accurately.

Genomic data uses single nucleotide polymorphisms (**SNP**) on genotype arrays called SNP chips. Several SNP chip densities are available for dairy cattle and include low (~3,000 SNP), medium (50K, ~54,000 SNP), and high (**HD**, ~777,000 SNP) density panels; higher density evaluation may include whole-genome sequencing (**WGS**) (Zhang *et al*., 2015a). Some studies have noted that the 50K panel lacked the sensitivity required for certain analyses (e.g. breed history and genome structure studies) when compared to HD and WGS data (Purfield *et al*., 2012; Marras *et al*., 2015). Imputation can be used when analyses require higher density information than the density of the genotypes that are available (Forutan *et al.*, 2018).

Genomic information from SNP chips can be used to estimate the realized proportion of the genome that is shared between individuals to provide a more accurate estimation of inbreeding (Howard *et al*., 2017). It should be noted that inbreeding is not necessarily detrimental, however the accumulation of deleterious alleles has been shown to cause inbreeding depression, where the health, survivability, and fertility of an animal is compromised (Miglior *et al*., 2005; González-Recio *et al.*, 2007; Vergeer *et al*., 2012; Leroy, 2014; Venney *et al*., 2016; Howard *et al*., 2015a). Inbreeding depression has been shown to be expressed differently between males and females (Fortes *et al.*, 2013; Berry *et al.*, 2014; Ebel and Phillips, 2016). There are three hypotheses to explain inbreeding depression, which are overdominance, partial dominance, and epistasis (Vergeer *et al*., 2012; Leroy, 2014; Venney *et al*., 2016). Leroy (2014) has summarized each hypothesis in his 2014 meta-analysis of inbreeding depression in livestock. The general consensus favours partial dominance as the main cause of inbreeding depression, where recessive deleterious alleles have increased in expression due to increased genome autozygosity (Vergeer *et al*., 2012).

However, two studies, one in plants and one in Chinook salmon, have found strong evidence to support that epigenetic processes, such as methylation, can impact inbreeding depression (Vergeer *et al*., 2012; Venney *et al*., 2016). Methylation and de-methylation can be influenced by several factors including age, environmental stress, and inbreeding level of an individual (Venney *et al*., 2016). This is an interesting area of research that needs to be explored further.

Inbreeding depression may lead to a reduction in fertility, which can be observed in a heifer or cow at any stage in her reproductive life (Berry *et al*., 2014). Reproductive biotechnology such as AI, heat synchronization, and embryonic transfer have become increasingly popular among dairy farmers when managing their herd to counteract the effects of inbreeding depression (Canadian Dairy Network, 2018a). These technologies are potentially only a temporary solution to compromised fertility, and thus, undesirable alleles continue to exist within a population for a longer period of time, potentially accumulating over generations unknowingly. The Canadian Dairy Network has reported an increasing trend in the number of herds being inseminated on only two days a week, providing strong evidence that farmers are increasingly using reproductive biotechnologies to control the estrus cycles of their cows, specifically heat synchronization in this case (Canadian Dairy Network, 2018b). Consumers have had mixed reactions to reproductive biotechnologies, particularly with hormone use and the safety of the dairy products they consume, therefore it is important for the dairy industry to consider alternative methods of improving fertility to avoid an adverse reaction from consumers in the future (Canadian Dairy Network, 2018b). Pryce *et al.* (2012) have estimated that the value of genotyping with the intent of controlling inbreeding is a profit between $5 to $10/cow. This may provide farmers the incentive to address inbreeding depression in their herds at the genomic level, an approach that consumers may be more receptive of compared to certain reproductive biotechnologies.

Runs of Homozygosity

Runs of homozygosity (**ROH**) are regions of homozygous loci in the genome that can be used to calculate the realized level of autozygosity of an individual and have been used in population history and disease studies in humans for over a decade (Kirin *et al*., 2010; Purfield *et al*., 2012; Curik *et al.*, 2014; Iacolina *et al*., 2016). Typically, long ROH indicate recent inbreeding, as there has not been sufficient time for recombination events to occur and break up the ROH (Gibson *et al*., 2006). In contrast, short ROH allude to ancient inbreeding (Kirin *et al*., 2010). Both short and long ROH are useful in evaluating inbreeding in populations under selection (i.e. livestock production).

Runs of homozygosity started to be used as a measure of inbreeding in dairy cattle in 2011 (Ferenčaković *et al*., 2011; Marras *et al*., 2015). Studies have shown that ROH-based inbreeding (**F_{ROH}**) provided a better estimate of inbreeding when compared to F_{PED} (Ferenčaković *et al*., 2011; Ferenčaković *et al*., 2013a, Forutan *et al.*, 2018), and both Gurgul *et al*. (2016) and Forutan *et al.* (2018) found that F_{ROH} provided better estimates of inbreeding compared to inbreeding calculated using the genomic relationship matrix. F_{ROH} has been found to be influenced by the genotype density used to detect ROH. Medium density chips lack the sensitivity to accurately identify short ROH (<4 Mb) (Purfield *et al*., 2012; Zhang *et al.*, 2015a; Zhang *et al.*, 2015b), however Zhang *et al.* (2015b) found a high correlation between 50K and WGS data, the highest correlation being associated with long ROH. Their study found that WGS data was best used to detect ancient inbreeding (i.e. for breed formation and genome structure studies), whereas 50K genotypes could be used for studies exploring recent inbreeding was recommended, as medium density genotypes estimate long ROH with high accuracy and are less costly and time consuming to run computational analyses.

Some studies state that ROH allow differentiation between regions of the genome that are identical by state (**IBS**) and identical by descent (**IBD**) (Leroy, 2014; Purfield *et al.,* 2017). The latter occurs when individuals inherit identical alleles from both parents that originated from a common ancestor, whereas the former occurs in non-related individuals due to areas of high linkage disequilibrium and low rates of recombination in the genome (Browning and Browning, 2010; Purfield *et al.*, 2017). The ability to make this distinction is one of the reasons why F_{ROH} is considered to provide a more accurate estimate of inbreeding compared to F_{PED} (Purfield *et al.,* 2017). Caution must be taken when analysing short ROH detected using 50K data, as it is more likely to mistake IBS loci for IBD loci due to the lack of sensitivity of the medium density array (Marras *et al*., 2015). This is more of a concern for research studies investigating breed histories and genome structure. In the case of the dairy industry, which is interested in identifying long ROH as indication of recent inbreeding events, 50K data has been shown to be sufficient. (Ferenčaković *et al.*, 2013b) An additional consideration that may affect the ability to differentiate between IBS and IBD loci are hot and cold ROH regions. In these regions, the recombination rate has been found to be either high (ROH cold spot) or low (ROH hot spot) because of low or high linkage disequilibrium respectively (Curik *et al.*,2014; Purfield *et al.,* 2017). In areas of high linkage disequilibrium, identical haplotypes that originate from unrelated ancestors could be passed on to offspring, making it appear that the haplotypes were IBD when they were in fact IBS (Curik *et al.*,2014; Purfield *et al.,* 2017). Furthermore, ROH hot and cold spots have been found to differ across breeds. For example, Ferenčaković *et al.* (2013b) located two ROH hot spots on BTA 6 in Brown Swiss (**BS**), Pinzgauer, and Tyrol Grey breeds. It would be interesting to explore how ROH hot and cold spots could be applied in breeding programs as research in this area continues.

Runs of homozygosity have been linked with inbreeding depression. Howard *et al*. (2017) discussed how ROH are rich with deleterious alleles, and these alleles have been found to accumulate at a faster rate than if they were found outside an ROH. It should be noted that deleterious alleles can either be lethal or loss of function alleles (Pryce *et al*., 2014). Howard *et al.* (2015a) found that the effect of the presence of ROH on an individual, whether it be positive or negative, depended on the location of ROH in the genome. For example, Bjelland *et al.* (2013) found that ROH affected heifer calf mortality rate in BS increased by 4.0% when F_{PED} was > 0.10. Bjelland *et al.* (2013) also observed an increase of 1.72 days of days open when F_{ROH} increased by 1.0%. Long ROH have been found to affect calving interval in BS, and the interval from first to last insemination for heifers and multiparous cows has been shown to be negatively impacted by high levels of inbreeding in Ayrshire (**AY**) (Fuerst-Waltl and Fuerst, 2012; Martikainen *et al.*, 2017). Calf mortality and days open have costly economic implications, and thus the dairy industry has much to gain with a clearer understanding of how ROH affects these and other fertility traits.

Despite the benefits of using ROH as a measure of inbreeding, there have been some challenges with adapting ROH to the dairy industry. As Ferenčaković *et al.* (2013b) state, there is a lack of a standardized definition of ROH, which can bias the results of ROH detection. There are several reasons why this standardization does not exist. One reason is that there are multiple ROH-analysing software available, and each software approaches the identification of ROH differently. Forutan *et al*. (2018) compared three of these software – PLINK (Chang *et al*., 2015), BCFtools (Narasimhan *et al.*, 2016), and SNP1101 (Sargolzaei, 2014) - and applied them on simulated data. PLINK is one of the most commonly used software as it is computationally fast (Forutan *et al*., 2018). It uses a fixed sliding window approach and allows the user to specify multiple parameters to determine what the program will identify as an ROH (Chang *et al*., 2015). PLINK was found to

underestimate F_{ROH} using simulated data, whereas in the same study BCFtools and SNP1101 were found to calculate F_{ROH} closest to the true level of inbreeding using simulated data (Forutan *et al*., 2018). SNP1101 uses an overlapping sliding window approach to detect ROH and is computationally faster when compared to BCFtools, thus Forutan *et al.* (2018) suggested that of the three software considered, SNP1101 was the best software to use to detect ROH. This program, however, is not currently open access. While BCFtools requires the most time to run analyses of the three programs compared here, it is open access and was found to estimate F_{ROH} with a high level of accuracy (Forutan *et al.,* 2018). BCFtools uses a Hidden Markov Model and parameter optimization (Viterbi training) based on allele frequencies and recombination rates to detect ROH (Narasimhan *et al.,* 2016; Forutan *et al*., 2018; BCFtools, 2018). In the case of the latter, a constant recombination rate can be set, or a genetic map can be provided to account for recombination hot spots (Narasimhan *et al.,* 2016). It is important to understand the costs and benefits of each program when making the decision of which to use to detect ROH, as ROH detection and characterization will change between each program.

Another challenge using ROH is which SNP are included in analyses. Several factors influence SNP inclusion in ROH analyses, including genotyping errors, quality control parameters and thresholds, and breed influences (Ferenčaković *et al.,* 2013b, Curik *et al.,* 2014; Hay and Rekaya, 2015; van den Berg *et al.,* 2016). Each of these factors could change the number and/or length of ROH detected, and the programs that detect ROH are not equipped to adjust for these factors. An assumption of multiple-breed analyses is that the effect of each SNP is uniform across breeds, which is not necessarily true (Hay and Rekaya, 2015). Large populations such as Holstein (**HO**) can bias allele frequencies, and in multiple-breed analyses, SNP that are found at high frequency in HO but not in the smaller breeds included in the analyses will pass filtering

thresholds. These alleles could be heterozygous and located within a long ROH in another breed, causing the ROH to be broken into smaller ROH. This could lead to an underestimation of recent inbreeding. Future research should focus on each of these challenges with the goal of developing a standardized definition and method of ROH identification and characterization to ensure the dairy industry can use F_{ROH} to both maximize genetic gain and manage inbreeding effectively (Curik *et al.*, 2014, Howrigan *et al.*, 2011).

REFERENCES

Berry, D. P., E. Wall, and J. E. Pryce. 2014. Genetics and genomics of reproductive performance in dairy and beef cattle. Animal. 8:105-121. https://dx.doi.org/10.1017/S1751731114000743.

Bjelland, D. W., K. A. Weigel, N. Vukasinovic, and J. D. Nkrumah. 2013. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. J. Dairy Sci. 96:4697-4706. https://dx.doi.org/10.3168/jds.2012-6435.

Blasco, A., and M. A. Toro. 2014. A short critical history of the application of genomics to animal breeding. Livest. Sci. 166:4-9. https://dx.doi.org/10.1016/j.livsci.2014.03.015.

Boichard, D., V. Durocq, P. Croiseau, and S. Fritz. 2016. Genomic selection in domestic animals: principles, applications and perspectives. C. R. Biologies. 339:274-277. https://dx.doi.org/10.1016.j.crvi.2016.04.007.

Browning and Browning. 2010. High-resolution detection of identity by descent in unrelated individuals. Am. J. Hum. Genet. 86:526-539. https://dx.doi.org/10.1016/j.ajhg.2010.02.021.

Canadian Dairy Information Centre. 2017a. Number of farms, dairy cows and heifers. Accessed July 14, 2018. http://www.dairyinfo.gc.ca/index_e.php?s1=dff-fcil&s2=farm-ferme&s3=nb.

Canadian Dairy Information Centre. 2017b. Dairy animal registrations. Accessed July 14, 2018. http://www.dairyinfo.gc.ca/index_e.php?s1=dff-fcil&s2=mrr-pcle&s3=dcr-ebl.

Canadian Dairy Information Centre. 2018a. Artificial insemination sire usage by age at insemination. Accessed July 14, 2018a. http://www.dairyinfo.gc.ca/index_e.php?s1=dff-fcil&s2=mrr-pcle&s3=yb-jt

Canadian Dairy Information Centre. 2018b. Trade of dairy genetics. Accessed July 14, 2018. http://dairyinfo.gc.ca/index_e.php?s1=dff-fcil&s2=imp-exp&s3=gen&menupos=01.01.14.

Canadian Dairy Network. 2018a. Inbreeding update. Accessed August 9, 2018. https://www.cdn.ca/document.php?id=506.

Canadian Dairy Network. 2018b. The fertility challenge. Accessed April 11, 2018. https://www.cdn.ca/document.php?id=493.

Curik, I., M. Ferenčaković, and J. Sölkner. 2014. Inbreeding and runs of homozygosity: a possible solution to an old problem. Livest. Sci. 166:26-34. http://dx.doi.org/10.1016/j.livsci.2014.05.034.

Ebel, E. R., and P. C. Phillips. 2016. Intrinsic differences between males and females determine sex-specific consequences of inbreeding. BMC Evol. Biol. 16:36. https://dx/doi.org/10.1186/s12862-016-0604-5.

Ferenčaković, M., E. Hamzić, B. Gredler, I. Curik, and J. Sölkner. 2011. Runs of homozygosity reveal genome-wide autozygosity in the Austrian Fleckvieh cattle. Agric. Conspec. Sci. 76:325–328. https://dx.doi.org/10.1111/age.12634.

Ferenčaković, M., E. Hamzić, B. Gredler, T. R. Solberg, G. Klemetsdal, I. Curik, and J. Sölkner. 2013a. Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. J. Anim. Breed. Genet. 130:286-293. http://dx.doi.org/10.1111/jbg/12012.

Ferenčaković, M., J. Sölkner, and I. Curik. 2013b. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genet. Sel. Evol. 45:42. http://dx.doi.org/10.1186/1297-9686-45-42.

Fortes, M. R. S., K. L. DeAtley, S. A. Lehnert, B. M. Burns, A. Reverter, R. J. Hawken, G. Boe-Hansen, S. S. Moore, and M. G. Thomas. 2013. Genomic regions associated with fertility traits in male and female cattle: advances from microsatellites to high-density chips and

beyond. Anim. Reprod. Sci. 141:1-19. https://dx.doi.org/10.1016/j.anireprosci.2013.07.002.

Forutan, M., S. A. Mahyari, C. Baes, N. Melzer, F. S. Schenkel, and M. Sargolzaei. 2018. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. BMC Genomics 19:98. https://dx.doi.org/10.1186/s12864-018-4453-z.

Fuerst-Waltl, B., and C. Fuerst. 2012. Effect of inbreeding depression on survival of Austrian Brown Swiss calves and heifers. J. Dairy Sci. 95:6086-6092. https://dx.doi.org/10.3168/jds.2011-4684.

Gibson, J., N. E. Morton, and A. Collins. 2006. Extended tracts of homozygosity in outbred human populations. Hum. Mol. Genet. 15:789-795. https://dx.doi.org/10.1093/hmg/ddi493

Gurgul, A., T. Szmatoła, P. Topolski, I. Jasielczuk, K. Żukowski, and M. Bugno-Poniewierska. 2016. The use of runs of homozygosity for estimation of recent inbreeding in Holstein cattle. J. Appl. Genet. 57:527–530. https://dx.doi.org/10.1007/s13353-016-0337-6.

González-Recio. O., E. L. de Maturana, and J. P. Gutierrez. 2007. Inbreeding depression on female fertility and calving ease in Spanish dairy cattle. J. Dairy Sci. 90:5744-5752. https://dx.doi.org/10.3168/jds.2007-0203.

Hay, E. H., and R. Rekaya. 2015. A multi-compartment model for genomic selection in multi-breed populations. Livest. Sci. 177:1-7. https://dx.doi.org/10.1016/j.livsci.2015.03.027.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:433-443. https://dx.doi.org/10.3168/jds.2008-1646.

Howard, J. T., C. Maltecca, M. Haile-Mariam, B. J. Hayes, and J. E. Pryce. 2015a.

Characterizing homozygosity across United States, New Zealand and Australian Jersey cow and bull populations. BMC Genomics 16:187. https://dx.doi.org/10.1186/s12864-015-1352-4.

Howard, J. T., M. Haile-Mariam, J. E. Pryce, and C. Maltecca. 2015b. Investigation of regions impacting inbreeding depression and their association with the additive genetic effect for United States and Australia Jersey dairy cattle. BMC Genomics 16:813. https://dx.doi.org/10.1186/s12864-015-2001-7.

Howard, J. T., J. E. Pryce, C. Baes, and C. Maltecca. 2017. Invited review: inbreeding in the genomics era: inbreeding, inbreeding depression, and management of genomic variability. J. Dairy Sci.100:1-16. https://dx.doi.org/10.3168/jds.2017-12787.

Howrigan, D. P., M. A. Simonsom, and M. C. Keller. 2011. Detecting autozygosity through runs of homozygosity: a comparison of three autozygosity detecting algorithms. BMC Genomics 12:460. https://dx.doi.org/10.1186/1471-2164-12-460.

Iacolina, L., A. V. Stronen, C. Pertoldi, M. Tokarska, L. S. Nørgaard, J. Muñoz, A. Kjærsgaard, A. Ruiz-Gonzalez, S. Kamiński, and D. C. Purfield. 2016. Novel graphical analyses of runs of homozygosity among species and livestock breeds. Int. J. Genomics 2016:2152847 https://dx.doi.org/10.1155/2016/2152847.

Keller, M. C., P. M. Visscher, and M. E. Goddard. 2011. Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics 189:237-249. https://dx.doi.org/10.1534/genetics.111.130922.

Kim, E.-S., T. S. Sonstegard, C. P. Van Tassell, G. Wiggans, and M. F. Rothschild. 2015. The relationship between runs of homozygosity and inbreeding in Jersey cattle under selection. PLoS One 10:e0129967. https://dx.doi.org/10.1371/journal.pone.0129967.

Kirin, M., R. McQuillan, C. S. Franklin, H. Campbell, P. M. McKeigue, and J. F. Wilson. 2010. Genomic runs of homozygosity record population history and consanguinity. PLoS One 5:e13996. https://dx.doi.org/10.1371/journal.pone.0013996.

Leroy, G. 2014. Inbreeding depression in livestock species: review and meta-analysis. Anim. Genet. 45:618–628. https://dx.doi.org/10.1111/age.12178.

Marras, G., G. Gaspa, S. Sorbolini, C. Dimauro, P. Ajmone-Marsan, A. Valentini, J. L. Williams, and N. P. P. Macciotta. 2015. Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. Anim. Genet. 46:110-21. http://dx.doi.org.subzero.lib.uoguelph.ca/10.1111/age.12259.

Martikainen, K., A. M. Tyrisevä, K. Matilainen, J. Pösö, and P. Uimari. 2017. Estimation of inbreeding depression on female fertility in the Finnish Ayrshire population. J. Anim. Breed. Genet. 134:383-392. https://dx.doi.org/10.1111/jbg.12285.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819-1829.

Miglior F., B. L. Muir, and B. J. VanDoormaal. 2005. Selection indices in Holstein cattle of various countries. J. Dairy Sci. 2005. 88:1255-1263. https://dx.doi.org/10.3168/jds.S0022-0302(05)72792-2.

Narasimhan, V., P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, and R. Durbin. 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. BCFtools version 1.5. Bioinformatics 32:1749-1751. https://dx.doi.org/10.1038/hgv.2016.25.

Pryce, J. E., B. J. Hayes, and M. E. Goddard. 2012. Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. J. Dairy Sci. 95:377–

388. https://dx.doi.org/10.3168/jds.2011-4254.

Pryce, J. E., M. Haile-Mariam, M. E. Goddard, and B. J. Hayes. 2014. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genet. Sel. Evol. 46.71 https://dx.doi.org/10.1186/s12711-014-0071-7.

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

Purfield, D. C., S. McParland, E. Wall, and D. P. Berry. 2017. The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. PLoS One 12:e0176780. https://dx.doi.org/10.1371/journal.pone.0176780.

Sargolzaei, M. 2014. SNP1101 User's Guide. Version 1. HiggsGene Solutions Inc.

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

Silva, M. V. B., D. J. A. dos Santos, S. A. Boison, A. T. H. Utsunomiya, A. S. Carmo, T. S. Sonstegard, J. B. Cole, and C. P. Van Tassell. 2014. The development of genomics applied to dairy breeding. Livest. Sci. 166:66-75. https://dx.doi.org/10.1016/j.livsci.2014.05.017.

Stachowicz, K., M. Sargolzaei, F. Miglior, and F. S. Schenkel. 2011. Rates of inbreeding and genetic diversity in Canadian Holstein and Jersey cattle. J. Dairy Sci. 94:5160-5175. https://dx.doi.org/10.3168/jds.2010-3308.

van den Berg, I., D. Boichard, and M. S. Lund. 2016. Comparing power and precision of within-breeding and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. J. Dairy Sci. 99:8932-8945. https://dx.doi.org/10.3168/jds.2016-11073.

VanRaden, P. M., K. M. Olson, G. R. Wiggans, J. B. Cole, and M. E. Tooker. 2011. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 94:5673-5682. http://dx.doi.org./10.3168/jds.2011-4500.

Venney, C. J., M. L. Johansson, and D. D. Heath. 2016. Inbreeding effects on gene-specific DNA methylation among tissues of Chinook salmon. Mol. Ecol. 25:4521-433. http://dx.doi.org/10.1111/mec.13777.

Vergeer, P., N. C. A. M. Wagemaker, and N. J. Ouborg. 2012. Evidence for an epigenetic role in inbreeding depression. Biol. Letters 8:798-801. http://dx.doi.org./10.1098/rsbl.2012.0494.

Weller, J. I., E. Ezra, and M. Ron. 2017. Invited review: a perspective on the future of genomic selection in dairy cattle. J. Dairy Sci. 100:8633-8644. https://dx.doi.org/10.3168/jds.2017-12879.

Zhang, Q., B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. 2015a. Runs of homozygosity and distribution of functional variants in the cattle genome. BMC Genomics 16:542. https://dx.doi.org/10.1186/s12864-015-1715-x.

Zhang, Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015b. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics 16:88. http://dx.doi.org/10.1186/s12863-015-0227-7.

CHAPTER 2: Characterizing runs of homozygosity in single- and multiple-breed analyses using five dairy cattle breeds and varying sample sizes

ABSTRACT

The effect of different methods of runs of homozygosity (**ROH**) detection and analysis on ROH characterization was investigated. An in-depth investigation of the use of the ROH-detecting program BCFtools was conducted. Additionally, single- and multiple-breed analyses were compared. The latter were divided into two scenarios; a three-breed analysis with similar population sizes (Ayrshire (n=6,560), Brown Swiss (n=5,667), and Guernsey (n=1,074), and a four-breed analysis with dissimilar population sizes (Jersey (n=77,581), Holstein (n=132,155, AY (n=6,560), and BS (n=5,667)). The effect of sample size on ROH characterization was also explored. The single nucleotide polymorphisms (**SNP**) included in filtered data were found to change between single- and multiple-breed analyses as well as across breeds. BCFtools was found to remove breed-specific mono-allelic SNP when they were present in multiple-breed analyses, and the program did not converge if breed-specific missing SNP were included in the filtered data. Additional differences between single- and multiple-breed analyses were observed in the identification and characterization of number and length of ROH across breeds. Accuracy of ROH characterization increased when sample size increased. The results of this study suggested that ROH characterization can change depending on different methods of detection and analysis, indicating that calculating genomic inbreeding using ROH in dairy cattle is complex and requires further study.

Key words: runs of homozygosity, inbreeding, dairy cattle

INTRODUCTION

Since the introduction of genomic selection in dairy production, there has been an increased rate of genetic gain, selection intensity, and loss of genetic variation, due in part to a shortened generation interval (Schaeffer, 2006). Additionally, the rate of inbreeding has increased in dairy breeds, compromising the health and fertility of individuals due to the accumulation of deleterious alleles (Leroy, 2014; Howard *et al.*, 2017; Reiner-Benaim *et al.*, 2017). Therefore, the dairy industry is investing in gaining a more comprehensive understanding of the mechanisms of inbreeding and the methods that can be used to manage it.

Runs of homozygosity (**ROH**) are regions of homozygous loci in the genome. They occur when parents transmit identical haplotypes to their offspring, and they can be used to estimate genomic inbreeding (Ferenčaković *et al.*, 2011; Ferenčaković *et al.*, 2013a; Szmatoła *et al.*, 2016). Long ROH indicate recent inbreeding, as there has not been time for recombination events to occur to break the ROH (Gibson *et al.*, 2006). Conversely, short ROH allude to ancient inbreeding, and have been used in studies investigating breed history and genome structure (Kirin *et al.*, 2010; Purfield *et al.*, 2012; Curik *et al.*, 2014).

Runs of homozygosity allow geneticists to measure the level of ROH-based inbreeding (**F_{ROH}**) of an individual more accurately when compared to pedigree-based inbreeding (**F_{PED}**) and genomic relationship-based inbreeding (Ferenčaković *et al.*, 2011; Ferenčaković *et al.*, 2013a, Forutan *et al.*, 2018). Nevertheless, there have been challenges in implementing F_{ROH} in the dairy industry. Some researchers analyse single breeds individually, whereas others analyse multiple breeds together (VanRaden *et al.*, 2011; Marras *et al*., 2015; Hay and Rekaya, 2015; van den Berg *et al.*, 2016). Studies have also found that a breed with a large population size impacted ROH detection and characterization in multiple-breed studies that included breeds with comparatively

smaller populations sizes (van den Berg *et al.*, 2016). Furthermore, different programs, such as PLINK (Chang *et al*., 2015), BCFtools (Narasimhan *et al.*, 2016), and SNP1101 (Sargolzaei, 2014), detect ROH using a variety of different approaches, and users are able to set varying parameters for each such as sliding window size, number of heterozygotes allowed in an ROH, and parameter optimization (Viterbi training) (Ferenčaković *et al.*, 2013b; Forutan *et al*., 2018). This availability of alternative approaches and parameter settings has the potential to introduce bias in ROH detection, and by extension, the detection and characterization of ROH. This in turn lead to an over or underestimation of F_{ROH} of an individual animal and effects the efforts of the dairy industry to use F_{ROH} to manage inbreeding levels in their populations (Ferenčaković *et al.*, 2013b).

Based on the findings of Forutan *et al.* (2018), BCFtools was chosen as the program to detect ROH. The purpose of this research was to identify and characterize ROH using different methods in five dairy breeds and observe how different methods changed results. This was accomplished by: 1) evaluating BCFtools as a program to identify ROH; 2) evaluating the effect of single- and multiple-breed analyses on ROH; and 3) evaluating the impact of various sample sizes on ROH characterization.

MATERIALS AND METHODS

Data and Data Preparation

Low density genotype data of 222,997 animals (birth year 1950-2017) were provided by the Canadian Dairy Network and imputed to 50K density (Illumina, 2016). The five dairy breeds included in the analyses were Ayrshire (**AY**), Brown Swiss (**BS**), Guernsey (**GU**), Jersey (**JE**), and Holstein (**HO**). The population sizes of each breed are summarized in Table 1. The number of

records in each pedigree, the average number of generations, and the maximum number of generations have been summarized in Table 2.

Animals were randomly selected to create three different sample sizes ($n_1 = 100$, $n_2 = 500$, and $n_3 = 1{,}000$). This process was repeated five times (Round 1 – Round 5). This method of sampling was done to avoid sampling a subset that did not represent the overall population well. Sample sizes were then combined into a single dataset for scenarios 2 and 3 (e.g., $n_{1AY} = 100$, $n_{1BS} = 100$, $n_{1JE} = 100$, $n_{1HO} = 100$; $n_{1AYBSJEHO} = 400$).

Three different scenarios were considered in this study (Figure 1). The first scenario was the set of single-breed analyses. The second was a three-breed set of analyses that included AY, BS, and GU, and the third was a four-breed set of analyses that included AY, BS, JE, and HO. The framework for each scenario is shown in Figure 1.

Genetic maps were created chromosome-wise using an R script programmed by Dr. Nina Melzer. This script was used to calculate recombination rates using the marker positions from the National Center for Biotechnology Information Annotation Release 104 along with the chromosome lengths from UMD3.1 (Table 3) (Kersey *et al.*, 2004; National Center for Biotechnology Information, 2017). It should be noted that when creating a genetic map chromosome-by-chromosome, the first SNP of each chromosome was lost. A few SNP were located beyond the end of the chromosome (according to the UMD3.1 map) and were manually removed (2, 65, 2, and 2 SNP on BTA 2, 6, 14, and 27 respectively). One SNP (Hapmap38362-BTA-94562) was identified at position 0 bp on Chromosome 6 and was removed.

R v3.3.2 was used to create a script to prepare the data for BCFtools analysis (R Core Team, 2016). The data originally included 45,187 SNP. PLINK v1.90b4.1 (Chang *et al.*, 2015)

was used to filter breeds separately and together, following the recommendations of Wiggans *et al.* (2009). For single-breed analyses, single nucleotide polymorphisms (**SNP**) that had a missing genotype rate ≥10.0% and a minor allele frequency (**MAF**) ≤2.5% for AY, BE, JE, and HO were removed. For GU animals, a missing genotype rate of ≥10.0% and a MAF ≤5.0% were applied as the population was less than 5,000 animals (Wiggans *et al.*, 2009). For multiple-breed analyses, SNP were filtered based on a missing genotype call rate of ≥10.0% and a MAF ≤2.5%. The number of SNP remaining after filtering are summarized in Table 1. After filtering, variant call format (**VCF**) files were created from the PLINK output binary files of the single-breed or multiple-breed datasets and compressed using PLINK (Appendix I). These VCF files, along with the genetic map, were then used in BCFtools analyses (Appendix I).

BCFtools

BCFtools v1.5 is a software that can be used to identify ROH (Narasimhan *et al.*, 2016; BCFtools, 2018). BCFtools uses a Hidden Markov Model (**HMM**) and optimizes allele frequencies and recombination rates through implementing Viterbi training (Narasimhan *et al.*, 2016; Forutan *et al*., 2018, BCFtools, 2018). The convergence threshold for Viberbi training, which is the only parameter that can be manipulated in BCFtools, was set to $1x10^{-10}$ (BCFtools, 2018). While including a genetic map is not necessary to run BCFtools analyses, it was included to account for recombination hot spots, rather than specifying a constant recombination rate for each base (Narasimhan *et al.*, 2016). .

BCFtools analyses were then run chromosome-wise using the VCF files and chromosomal genetic maps (Appendix I). The output was converted using a python script to resemble PLINK ROH output format (Figure 4). This was done in order to use the R package *detectRUNS* (Biscarini *et al*., 2018)to read the results of the BCFtools ROH analysis into R and to characterize ROH.

ROH Characterization

The R package *detectRUNS* was used to characterize ROH for each scenario, breed, sample size, and round. ROH were sorted into five length classes: <2 Mb, 2-4 Mb, 4-8 Mb, 8-16 Mb, and >16 Mb. The distributions of ROH across the genome as well as length classes were calculated using the function *summaryRuns*. The average number and length of ROH for each round was determined, and these values were then averaged across all five rounds. F_{ROH} coefficients were calculated genome-wide and averaged across the five rounds in the same manner as outlined previously. The formula to calculate F_{ROH} was:

$$F_{ROH} = \frac{\sum L_{ROH}}{L_{AUTOSOMES}}$$

where $\sum L_{ROH}$ is the sum of the length of ROH in the genome, and $L_{AUTOSOMES}$ is the total length of autosomes (McQuillan *et al.*, 2008; Curik *et al.*, 2014). Pedigree-based inbreeding was calculated using the R package *pedigree* (Coster, 2013) and averaged across rounds. Standard errors (**SE**) were calculated for each average of the five rounds using the following equation:

$$SE = \frac{s}{\sqrt{n}}$$

where SE is the standard error of the mean, s is the standard deviation of the rounds (samples), and n is the sample size.

RESULTS AND DISCUSSION

Filtering Data

Data was filtered based on the recommendations of Wiggans *et al*. (2009), where SNP with missing genotype rate of ≤10.0% and MAF ≤2.5% (when n ≥5,000) or ≤5.0% (when n ≤5,000) thresholds were applied.

The SNP remaining after filtering differed between single- and multiple-breed analyses (Table 1). The inclusion of SNP depended on the quality control parameters applied (i.e. MAF, missing genotype rate), and allele frequencies within populations, which in turn depended on the size of the population being filtered. The different scenarios considered in this study changed the frequency of SNP in each filtered dataset, and these differences directly corresponded to differences observed across breeds (Table 4 to Table 7, Figure 6 to Figure 9). In the multiple-breed analyses, the change detected most likely was due to the influence of large breeds on the number of SNP that passed the quality control thresholds. For example, in Scenario 3, almost 95% of animals were either HO or JE. If a SNP found in either of these breeds passed quality control, it would be included in the 'clean' data, regardless of if that SNP met the quality control thresholds in the smaller breeds. The inclusion of such SNP could be found to impact ROH characterization (i.e. number and length of ROH detected), leading to inaccurate calculations of F_{ROH}. Therefore, it is important for the dairy industry to consider the impact of dissimilar population size on F_{ROH} and by extension inbreeding management in multiple-breed analyses.

It is worth noting that while the SNP included in each single-breed analysis were breed and population specific, single-breed analyses compromise the sensitivity of SNP-specific comparisons. These types of comparisons would benefit studies of rare haplotypes in small breeds and would be beneficial in disease studies.

BCFtools

Figures 2, 3, and 5 are samples of the first few lines of BCFtools ROH output with Viterbi training, without Viterbi training, and statistics output respectively.

In this study, a statistics file was created for each BCFtools analysis. This file served as an informal check point to ensure the data inputted into BCFtools was being processed as expected.

Of particular interest was the summary section, as seen in Figure 5. Important information about the number of samples and the number of SNP being analysed could be found in this section. Additionally, the number of no-ALTs (mono-allelic SNP) was reported in this section.

During preliminary analyses, it was discovered that when multiple breeds were filtered together and then analysed separately using BCFtools, additional SNP were removed during ROH detection when compared to the number of SNP removed during filtering. Upon closer inspection, these additional SNP were found to be breed-specific and mono-allelic. In general, the larger the population was, the fewer mono-allelic SNP were removed. When the single-breed analyses were investigated, it was noted that no additional SNP were removed during ROH detection as mono-allelic SNP had been removed during filtering. Also of note was that missing SNP in small populations masked by SNP in large populations. For example, in the current dataset, 16 SNP were found to be missing in the GU population. In single-breed analyses, these were removed during data filtering. However, when GU was filtered with JE and HO in multiple-breed analyses, these SNP remained in the data. It was discovered that BCFtools is not able to converge when SNP were missing. These results regarding mono-allelic SNP and missing genotypes not only emphasized the importance of data filtration, but also defined limitations to both BCFtools and multiple-breed analyses.

BCFtools allows users to specify whether to use Viterbi training or not in analyses. Viterbi training provides parameter optimization based on allele frequencies and recombination rates and uses these observations to determine the likelihood of a SNP being found within an ROH for each individual animal. In this study, BCFtools analyses were performed with Viterbi training as it was thought to provide the most accurate results. Computational time dramatically increased with the inclusion of Viterbi training, as optimized parameters are estimated for each animal individually.

Additionally, the format of the BCFtools output file changed when Viterbi training was used and was not (Figures 2 and 3). The difference in output files meant that the R package *detectRUNS* could not read in BCFtools output directly and required a python script (created by Dr. Gabriele Marras) to convert the file into a format *detectRUNS* could read (Figure 4). Future research would benefit from a comparison study to determine if there is a significant difference in ROH detection when Viterbi training is applied or not. Additionally, these findings stress the importance of being aware what the data being analyzed looks like at each stage of analysis.

ROH Characterization

Tables 4 to 6 summarize the distribution of ROH across length classes for each scenario, breed, and subset. The values presented in Tables 4 to 6 are averages of rounds 1 to 5, which in turn are averages of the individuals found within each round. The standard errors calculated are based on the averages of rounds, as the results of individual animals were not considered in this study.

Table 4 summarizes the results for each single-breed analysis. The majority of ROH in AY were found to be between 2 to 8 Mb in length (34% were 2 to 4 Mb in length, and 29% were 4 to 8 Mb). Very few ROH were found to be longer than 16 MB (0.082%). Similar distributions across class lengths were found in GU, JE, and HO. Brown Swiss was anomalous compared to the other four breeds in that few short ROH were found between ≤2 Mb (0.094%), and that long ROH (>16 Mb) accounted for 12% of ROH found. It is possible that this difference is due to cross breeding with other breeds, however when investigated, BS did not appear to have an open herd book in Canada (The Canadian Brown Swiss and Braunvieh Association, 2017). This finding will require further investigation.

In Scenario 2, the subsets of AY had a higher proportion of ROH between ≤2 Mb in length, as seen in Table 5. In most length classes, the SE of AY (n=500) was found to be smaller than that of AY (n=1,000). Brown Swiss subsets had a higher proportion of short ROH as well, and the SE of the extreme class lengths (≤2 Mb and >16 Mb) were found to be smallest in the 500 animal subset. In both cases, smaller SE were calculated for the sample size of 500 animals in different breeds, suggesting that the medium-sized subset characterized ROH more precisely than the 1,000 animal subset. It is most likely this was due to the way animals were randomly sampled in this study and may change if animals were sampled more than five times. Additionally, it should be noted that the SE calculated in most cases were extremely small ($\pm SEx10^{-3}$). This may be due to a bias imposed by animals that were sampled more than once across rounds. In this study, the effect of repeated animals was not analysed. Future research should consider performing the same analyses using a bootstrap approach to account for these animals.

The proportion of short ROH in Scenario 3 was lower than expected in AY and BS (5% and 6% respectively) when comparing the 1,000 animal subset of each breed to the whole population (Table 6). The SE of BS (n=1,000) was larger in the extreme class lengths (≤2 Mb, ≥16 Mb) compared to BS (n=500), suggesting the ROH distribution in these classes is more precise in a subset of 500 animals. Jersey (n=1,000) was found to have the most similar ROH distribution to what was expected in the population, with the exception of the ≤2 Mb length class. Holstein subsets had a higher proportion of short ROH (5% more than expected based on the overall population), and consequently a lower proportion of the medium to long ROH class lengths.

Generally, it was found that in single-breed analyses, it is possible to use a smaller subset of animals (n=1,000) to detect and characterize ROH and obtain similar results to that found in the

population (Table 4). Using smaller datasets will drastically reduce computational time, which is useful when using BCFtools to identify ROH.

Due to the nature of the sample sizes considered in each scenario, it is not possible to make across-scenario comparisons as the sample sizes considered were not the same in each scenario and any comparisons made would be the consequence of confounding results. However, it can be said that in general, a greater proportion of shorter ROH appeared to be found in multiple-breed analyses than in single-breed analyses (0.094% to 15% in Scenario 1, 13% to 26% in Scenario 2, and 19% to 25% in Scenario 3). This may be due in part to the SNP that are included in single- and multiple-breed analyses. With more SNP included in the multiple-breed analyses (Table 1), there is a chance that SNP may break ROH that are found to be continuous in single-breed analyses. As well, the percentage of long ROH (≥16 Mb) remained relatively constant within breeds across scenarios and sample sizes. The exception to this observation was the BS breed, in which both medium (4 to 8 Mb, 8 to 16 Mb) and long (≥16 Mb) ROH distributions decreased. In each scenario, BS was found to have the highest percentage of ROH ≥16 Mb, and HO had the smallest change in distribution across scenarios and sample sizes. The latter suggested a breed bias in the SNP included in each analysis that favoured HO (van den Berg *et al.,* 2016), whereas the former suggested that Canadian and American BS animals were inbred more recently. This would need to be further investigated by looking at annual changes in ROH for each class length.

The inclusion of SNP strongly depended on t allele frequencies within populations (Ferenčaković *et al.,* 2013b). VanRaden *et al.* (2011) have suggested that 50K genotypes were sensitive enough for within breed analyses, however suggested a higher density SNP chip was necessary to perform across-breed analyses. To feasibly do this in the dairy industry, genotype data would need to be imputed from low- and medium-density to high-density. Another factor to

consider is which SNP are present in the SNP chip being used and the breeds used to validate the SNP, as this can influence the number and length of ROH identified (VanRaden *et al.,* 2011; Ferenčaković *et al.,* 2013b; Illumina, 2016).

The distribution of ROH across the genome (corrected for chromosome length in Mb) was also considered in this study (Figure 6 to 8). Overall, similar trends were seen between each breed within both single- and multiple-breed analyses, although breed differences were observed in location and number of ROH (Figures 6 to 8). Chromosome 10 had the highest percentage of ROH in each scenario, breed, and sample size. Each of the five breeds analysed in this study have been under intense selection for milk production traits, and numerous QTL associated with udder traits have been found on this chromosome (Schrooten *et al.*, 2000; Hiendleder *et al.*, 2003; Ashwell *et al.,* 2005). This higher percentage of autozygosity on BTA 10 is most likely a signature of selection, however has not been found in other studies (Kim *et al*., 2013; Howard *et al.*, 2015). Additionally, the proportion of ROH found on BTA 14 was less than expected considering the intense selection for milk production and the gene DGAT1 being found on this chromosome. It would be interesting to examine both these regions further and investigate the number, length, and location of ROH on an individual level in the future.

Table 7 summarizes the F_{ROH}, F_{PED}, average ROH lengths, and average number of ROH detected for each scenario, breed, and sample size. Again, comparisons across scenarios cannot be performed, however in general, the results of this study indicate that F_{ROH} increased from single- to multiple-breed analyses in all breeds. F_{ROH} in HO was found to be in agreement with Szmatoła *et al*. (2016), and F_{ROH} in BS in agreement with Ferenčaković *et al.* (2013b) and Ferenčaković *et al.* (2013a). Regarding F_{PED}, the values calculated in this study were found to be lower than in the

literature (Stachowicz *et al.*, 2011; Canadian Dairy Network, 2018a). This may be due to a smaller pedigree used to calculated F_{PED}, and should be investigated further in the future (Table 2)

The average length of ROH decreased across scenarios whereas the average number of ROH increased, similar to results present in Tables 4 to 6. It should be noted, however, that the SE for the average number of ROH varied greatly (SE as large as ±0.55), indicating that the precision of determining the average number of ROH per individual was low. The average length observed for HO in the current study was found to be longer than what was reported by Forutan *et al*. in 2018. On average, JE was found to have more ROH than the other four breeds, which was shown in Table 7 and Figure 9. This is in agreement with the results presented by Zhang *et al*. (2015a).

Figure 9 illustrates the differences in average ROH length and average number of ROH between breeds for Scenario 1. Guernsey animals were found to be more tightly clustered within breed than the other four breeds, indicating less variation in number of ROH and average ROH length. Ayrshire and BS had similar numbers and lengths of ROH, and the lengths of ROH detected were slightly more variable that GU. Jersey animals were found to have the most ROH per individual on average and had a larger variation in both number and length of ROH. Holstein individuals had less variation in the number of ROH per individual, but were more variable in ROH length, and had animals with ROH ≥16 Mb than the other four breeds.

For Scenario 2, Figure 9 shows an even stronger clustering of the GU population, with more ROH identified, and less variation in average ROH length. Ayrshire and BS also lost variation in average ROH length, however the number of ROH detected remained similar to that in Scenario 1. Scenario 3 shows again a loss in variation in the average ROH length for AY, BS, and JE (Figure 9), however ROH length in HO remained similar to that in Scenario 1.

The apparent loss of variation in the average length of ROH is possibly due to the short ROH that are found more abundantly in multiple-breed analyses. Short ROH found in a population are attributed to regions of the genome that have been under selection for a longer time and are therefore likely associated with favourable traits (Zhang *et al.*, 2015a). Short ROH have also been found to become fixed in a population, resulting in a loss of genetic diversity (Zhang *et al.*, 2015a), thus depending on the trait and the associated gene(s), it would be reasonable to observe less variation in ROH length as selection for desirable traits is a major breeding goal. The loss of genetic variation may mean less opportunity for genetic gain in the future, a common concern associated with inbreeding. If this rate of loss continues, it will become more difficult for the dairy industry to adapt to future market or environmental changes.

There are a number of areas for future study based on the findings of this research. Firstly. repeated animals need to be accounted for, as the can create a dependence between samples (rounds), thus decreasing the SE of the means. This dependence increases as sample size increases or as the number of rounds sampled increases. Secondly, principle component analysis should be performed to determine the degree of relatedness between breeds and countries, as this information may help explain breed differences observed in this study. The results of this study cannot report which scenario was best as the sample sizes considered in each scenario were different. To avoid confounding results and to be able to compare scenarios, sample sizes should be the same in each scenario in future analyses. Finally, the trends of ROH per year should be investigated as this information will allow for more direct inferences on the levels of homozygosity in an individual that are due to IBD segments.

CONCLUSION

In conclusion, the characterization of ROH is influenced by many factors. Results from single- and multiple-breed analyses were found to differ, especially in smaller breeds. In general, multiple-breed analyses appeared to identify a larger number of short ROH compared to single-breed analyses. The distribution of ROH across the genome and across class lengths varied between breeds, most likely due to different selection pressures and population histories and structures. The number of ROH detected varied greatly across scenarios, breeds, and sample sizes, and overall, the precision of ROH detection increased as sample size increased. More variation in average ROH length was observed in some breeds (HO, JE), whereas others had less variation in length (AY, BS, GU).

The results of this study indicate that quantifying ROH is a difficult task that is influenced by a variety of factors, some known, and others not yet fully understood. Further work is needed to create a more robust and uniform method to identify and characterize ROH. Regardless of the challenges, ROH have the potential to improve mating decisions, which will allow the dairy industry to optimize genetic gain while maintaining inbreeding levels and genetic variation.

REFERENCES

Ashwell, M. S., D. W. Heyen, J. I. Weller, M. Ron, T. S. Sonstegard, C. P. Van Tassell, and H. A. Lewin. 2005. Detection of quantitative trait loci influencing conformation traits and calving ease in Holstein-Friesian cattle. J. Dairy Sci. 88:4111-4119.

BCFtools. 2018. Manual page. Accessed November 14, 2017. https://samtools.github.io/bcftools/bcftools.html

Biscarini, F., P. Cozzi, G. Gaspa, and G. Marras. 2018. detectRUNS: Detect runs of homozygosity and runs of heterozygosity in diploid genomes. R package version 0.9.5. https://CRAN.R-project.org/package=detectRUNS

Chang, C. C., C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. PLINK version 1.90b4.1. Gigascience 4:7. https://dx.doi.org/10.1186/s13742-015-0047-8.

Coster, A. 2013. pedigree: Pedigree functions. R package version 1.4. https://CRAN.R-project.org/package=pedigree

Curik, I., M. Ferenčaković, and J. Sölkner. 2014. Inbreeding and runs of homozygosity: a possible solution to an old problem. Livest. Sci. 166:26-34. https://dx.doi.org/10.1016/j.livsci.2014.05.034.

Ferenčaković, M., E. Hamzić, B. Gredler, I. Curik, and J. Sölkner. 2011. Runs of homozygosity reveal genome-wide autozygosity in the Austrian Fleckvieh cattle. Agric. Conspec. Sci. 76:325-328.

Ferenčaković, M., E. Hamzić, B. Gredler, T. R. Solberg, G. Klemetsdal, I. Curik, and J. Sölkner. 2013a. Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. J. Anim. Breed. Genet. 130:286-293. https://dx.doi.org/10.1111/jbg/12012.

Ferenčaković, M., J. Sölkner, and I. Curik. 2013b. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genet. Sel. Evol. 45:42. https://dx.doi.org/10.1186/1297-9686-45-42.

Forutan, M., S. A. Mahyari, C. Baes, N. Melzer, F. S. Schenkel, and M. Sargolzaei. 2018. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. BMC Genomics 19:98. https://dx.doi.org/10.1186/s12864-018-4453-z.

Hay, E. H., and R. Rekaya. 2015. A multi-compartment model for genomic selection in multi-breed populations. Livest. Sci. 177:1-7. https://dx.doi.org/10.1016/j.livsci.2015.03.027.

Hiendleder, S., H. Thomsen, N. Reinsch, J. Bennewitz, B. Leyhe-Horn, C. Looft, N. Xu, I. Medjugorac, I. Russ, C. Kuhn, G. A. Brockmann, J. Blumel, B. Brenig, F. Reinhardt, R. Reents, G. Averdunk, M. Schwerin, M. Forster, E. Kalm, and G. Erhardt. 2003. Mapping of QTL for body conformation and behavior in cattle. J. Hered. 94:496-506. https://dx.doi.org/10.1093/jhered/esg090.

Howard, J. T., C. Maltecca, M. Haile-Mariam, B. J. Hayes, and J. E. Pryce. 2015. Characterizing homozygosity across United States, New Zealand and Australian Jersey cow and bull populations. BMC Genomics 16:187. https://dx.doi.org/10.1186/s12864-015-1352-4.

Howard, J. T., J. E. Pryce, C. Baes, and C. Maltecca. 2017. Invited review: inbreeding in the genomics era: inbreeding, inbreeding depression, and management of genomic variability. J. Dairy Sci.100:1-16. https://dx.doi.org/10.3168/jds.2017-12787.

Illumina. 2016. BovineSNP50 Genotyping BeadChip. Accessed November 23, 2017. https://www.illumina.com/Documents/products/datasheets/datasheet_bovine_snp5O.pdf

Kersey, P. J., J. E. Allen, I., Armean, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, P.

Davis, L. J. Falin, C. Grabmueller, J. Humphrey, A. Kerhornou, J. Khobova, N. K. Aranganathan, N. Langridge, E. Lowy, M. D. McDowall, U. Maheswari, M. Nuhm, C. K. Ong, B. Overduin, M. Paulini, H. Pedro, E. Perry, G. Spudich, E. Tapanari, B. Walts, G. Williams, M. Tello-Ruiz, J. Stein, S. Wei, D. Ware, D. M. Bolser, K. L. Howe, E. Kulesha, D. Lawson, G. Maslen, and D. M. Staines. 2015. Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res. 4:574-580. https://dx.doi.org/10.1093/nar/gkv1209.

Kim, E. -S., J. B. Cole, H. Hudson, G. R. Wiggans, C. P. Van Tassell, B. A. Crooker, G. Liu, Y. Da., and T. S. Sonstegard. 2013. Effect of artificial selection on runs of homozygosity in U.S. Holstein Cattle. PLoS ONE. 8:1-14. https://dx.doi.org/10.1371/journal.pone.0080813.

Kirin, M., R. McQuillan, C. S. Franklin, H. Campbell, P. M. McKeigue, and J. F. Wilson. 2010. Genomic runs of homozygosity record population history and consanguinity. PLoS ONE. 5:e13996. https://dx.doi.org/10.1371/journal.pone.0013996.

Leroy, G. 2014. Inbreeding depression in livestock species: review and meta-analysis. Anim. Genet. 45:618-628. https://dx.doi.org/10.1111/age.12178.

Marras, G., G. Gaspa, S. Sorbolini, C. Dimauro, P. Ajmone-Marsan, A. Valentini, J. L. Williams, and N. P. P. Macciotta. 2015. Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. Anim. Genet 46:110-121. https://dx.doi.org.subzero.lib.uoguelph.ca/10.1111/age.12259.

McQuillan, R., A. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, L. Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, P. Ozren, A. Tenesa, A. K. Macleod, S. M. Farrington, P. Rudan, C. Hayward, V. Vitart, I. Rudan, S. H. Wild, M. G. Dunlop, A. F. Wright, h. Campbell, and J. F. Wilson. 2008. Runs of homozygosity in European populations. Am. J. Hum. Genet. 83:359-372. https://dx.doi.org/10.1016/j.ajhg.2008.08.007.

Narasimhan, V., P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith, and R. Durbin. 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. BCFtools version 1.5. Bioinformatics 32:1749-1751. https://dx.doi.org/10.1038/hgv.2016.25.

National Center for Biotechnology Information. 2014. Annotation Release 104. Accessed October 14, 2017. https://www.ncbi.nlm.nih.gov/projects/mapview/map_search.cgi?taxid=9913

Purfield, D. C., S. McParland, E. Wall, and D. P. Berry. 2017. The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. PLoS ONE 12:e0176780. https://dx.doi.org/10.1371/journal.pone.0176780.

R Core Team. 2016. R: A language and environment for statistical computing. R version 3.3.2. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org

Reiner-Benaim, A., E. Ezra, and J. I. Weller. 2017. Optimization of a genomic breeding program for a moderately sized dairy cattle population. J. Dairy Sci. 100:2892-2904. https://dx.doi.org/10.3168/jds.2016-11748.

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

Schrooten, C., H. Bovenhuis, W. Coppieters, and J. A. M. Van Arendonk. 2000. Whole genome scan to detect quantitative trait loci for conformation and functional traits in dairy cattle. J. Dairy Sci. 83:795-806. https://dx.doi.org/10.3168.jds.S0022-0302(00)74942-3.

Stachowicz, K., M Sargolzaei, F. Miglior, and F. S. Schenkel. 2011. Rates of inbreeding and genetic diversity in Canadian Holstein and Jersey cattle. J. Dairy Sci. 94:5160-5175. https://dx.doi.org/10.3168/jds.2010-3308. Szmatoła, T., A. Gurgul, K. Ropka-Molik, I.

Jasielczuk, T. Ząbek, and M. Bugno-Poniewierska. 2016. Characteristics of runs of homozygosity in selected cattle breeds maintained in Poland. Livest. Sci. 188:72–80. https://dx.doi.org/10.1016/j.livsci.2016.04.006.

The Canadian Brown Swiss and Braunvieh Association. 2017. By-laws. Accessed August 19, 2018. http://www.browncow.ca/Content/Static/Files/bylaws-en.pdf

Wiggans, G. R., T. S. Sonstegard, P. M. VanRaden, L. K. Matukumalli, R. D. Schnabel, J. F. Taylor, F. S. Schenkel, and C. P. Van Tassell. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92:3431-3436. https://dx.doi.org/10.3168/jds.2008-1758.

van den Berg, I., D. Boichard, and M. S. Lund. 2016. Comparing power and precision of within-breeding and multibreed genome-wide association studies of production traits using whole-genome sequence data for 5 French and Danish dairy cattle breeds. J. Dairy Sci. 99:8932-8945. https://dx.doi.org/10.3168/jds.2016-11073.

VanRaden, P. M., K. M. Olson, G. R. Wiggans, J. B. Cole, and M. E. Tooker. 2011. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 94:5673–5682. https://dx.doi.org/10.3168/jds.2011-4500.

Zhang, Q., B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. 2015a. Runs of homozygosity and distribution of functional variants in the cattle genome. BMC Genomics 16:542. https://dx.doi.org/10.1186/s12864-015-1715-x.

Table 1: Number of animals in each scenario and breed, and the number of SNP remaining after filtering

Scenario	Breed	Number of Animals	Number of SNP
1	Ayrshire (**AY**)	6,560	40,022
	Brown Swiss (**BS**)	5,667	37,261
	Guernsey (**GU**)	1,074	37,603
	Jersey (**JE**)	77,581	36,971
	Holstein (**HO**)	132,115	42,245
2	Ayrshire, Brown Swiss, Guernsey (**AY/BS/GU**)	13,301	41,841
3	Ayrshire, Brown Swiss, Jersey, Holstein (**AY/BS/JE/HO**)	221,923	43,417

Table 2: Number of records, average number of generations, and maximum number of generations for each pedigree

Breed	Number of Records	Average Number of Generations	Maximum Number of Generations
Ayrshire	35393	8.48 (3.26)[a]	22
Brown Swiss	138077	9.75 (1.57)[a]	23
Guernsey	25895	7.14 (2.35)[b]	24
Jersey	464192	14.22 (9.60)[c]	25
Holstein	3055874	13.27 (1.00)[a]	25

[a]±SE x10^{-2}, [b]±*SE x*10^{-4}, [c]±*SE x*10^{-3}

Table 3: Length of autosomes in bp and cM.

Chromosome	Chromosome Length (bp)[1]	Chromosome Length (cM)[2]
1	158337067	142.1
2	137060424	120.4
3	121430405	125.2
4	120829699	101.5
5	121191424	122.1
6	119458736	125.6
7	112638659	134.1
8	113384836	116.1
9	105708250	108.4
10	14305016	101.4
11	107310763	123.5
12	91163125	105.8
13	84240350	87.1
14	84648390	85.7
15	85296676	93.4
16	81724687	96.5
17	75158596	98.6
18	66004023	84.7
19	64057457	99.5
20	72042655	75.0
21	71599096	87.6
22	61435874	81.1
23	52530062	67.1
24	62714930	62.5
25	42904170	64.9
26	51681464	72.6
27	45407902	64.1
28	46312546	52.4
29	51505224	65.0

[1]Kersey *et al.* (2015); [2]National Center for Biotechnology Information Annotation (2017)

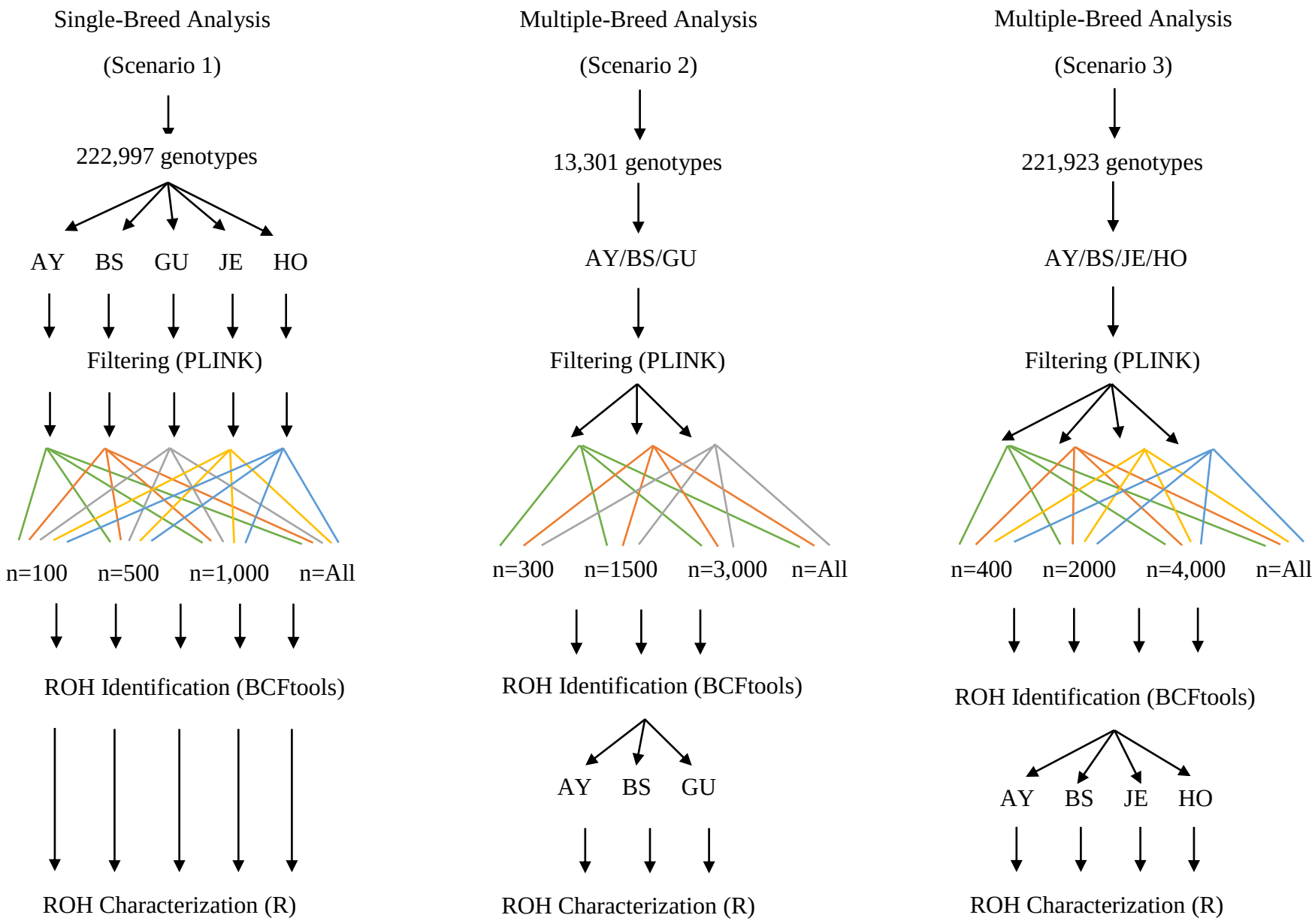


Figure 1: Workflow of study outlining each step taken for each analysis: single-breed – Scenario 1; three-breeds – Scenario 2; four-breeds – Scenario 3. (AY – Ayrshire; BS – Brown Swiss; GU – Guernsey; JE – Jersey; HO – Holstein)

```
# This file was produced by: bcftools roh(1.5+htslib-1.5)
# The command line was: bcftools roh -m Final_Maps/map1.txt -e - -G30 -V 1e-10 VCF_Files/All/AY/AY_All-chr1.new.vcf.gz
#
# RG    [2]Sample       [3]Chromosome   [4]Start        [5]End  [6]Length (bp)  [7]Number of markers    [8]Quality (average fwd-bwd p
# ST    [2]Sample       [3]Chromosome   [4]Position     [5]State (0:HW, 1:AZ)   [6]Quality (fwd-bwd phred score)
# VT, Viterbi Training  [2]Sample       [3]Iteration    [4]dAZ  [5]dHW  [6]1 - P(HW|HW) [7]P(AZ|HW)     [8]1 - P(AZ|AZ) [9]P(HW|AZ)
VT      AY_AY840F3005646800     1       1.545807e-05    7.771275e-05    1.552507e-05    1.552507e-05    7.771775e-05    7.771775e-05
VT      AY_AY840F3005646800     2       1.413402e-04    3.034243e-04    1.568652e-04    1.568652e-04    3.811421e-04    3.811421e-04
VT      AY_AY840F3005646800     3       7.365163e-04    1.781080e-03    8.933815e-04    8.933815e-04    2.162222e-03    2.162222e-03
VT      AY_AY840F3005646800     4       4.162988e-03    1.001683e-02    5.056370e-03    5.056370e-03    1.217906e-02    1.217906e-02
VT      AY_AY840F3005646800     5       2.292444e-02    5.346114e-02    2.798081e-02    2.798081e-02    6.564019e-02    6.564019e-02
VT      AY_AY840F3005646800     6       1.099988e-01    2.218561e-01    1.379796e-01    1.379796e-01    2.874963e-01    2.874963e-01
VT      AY_AY840F3005646800     7       3.127411e-01    4.189476e-01    4.507207e-01    4.507207e-01    7.064439e-01    7.064439e-01
VT      AY_AY840F3005646800     8       3.272948e-01    2.344088e-01    7.780156e-01    7.780156e-01    9.408527e-01    9.408527e-01
VT      AY_AY840F3005646800     9       1.499740e-01    5.069734e-02    9.279896e-01    9.279896e-01    9.915500e-01    9.915500e-01
VT      AY_AY840F3005646800     10      4.996807e-02    7.361051e-03    9.779576e-01    9.779576e-01    9.989110e-01    9.989110e-01
VT      AY_AY840F3005646800     11      1.538786e-02    9.525623e-04    9.933455e-01    9.933455e-01    9.998636e-01    9.998636e-01
VT      AY_AY840F3005646800     12      4.652680e-03    1.194513e-04    9.979982e-01    9.979982e-01    9.999831e-01    9.999831e-01
VT      AY_AY840F3005646800     13      1.400232e-03    1.484617e-05    9.993984e-01    9.993984e-01    9.999979e-01    9.999979e-01
VT      AY_AY840F3005646800     14      4.208531e-04    1.840368e-06    9.998192e-01    9.998192e-01    9.999997e-01    9.999997e-01
VT      AY_AY840F3005646800     15      1.264943e-04    2.279364e-07    9.999457e-01    9.999457e-01    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     16      3.859125e-05    2.804593e-08    9.999843e-01    9.999843e-01    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     17      1.142495e-05    3.573696e-09    9.999958e-01    9.999958e-01    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     18      3.237086e-06    5.250127e-10    9.999990e-01    9.999990e-01    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     19      7.995561e-07    8.267653e-11    9.999998e-01    9.999998e-01    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     20      1.651313e-07    1.324674e-11    1.000000e+00    1.000000e+00    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     21      3.170474e-08    2.130296e-12    1.000000e+00    1.000000e+00    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     22      5.983417e-09    3.426148e-13    1.000000e+00    1.000000e+00    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     23      1.125381e-09    5.528911e-14    1.000000e+00    1.000000e+00    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     24      2.115290e-10    8.770762e-15    1.000000e+00    1.000000e+00    1.000000e+00    1.000000e+00
VT      AY_AY840F3005646800     25      3.975476e-11    1.554312e-15    1.000000e+00    1.000000e+00    1.000000e+00    1.000000e+00
ROH     AY_AY840F3005646800     1       267940  0       3.0
ROH     AY_AY840F3005646800     1       393248  0       30.1
ROH     AY_AY840F3005646800     1       471078  0       33.6
ROH     AY_AY840F3005646800     1       516404  0       33.3
ROH     AY_AY840F3005646800     1       533815  0       59.2
ROH     AY_AY840F3005646800     1       571340  0       81.1
ROH     AY_AY840F3005646800     1       883895  0       59.9
ROH     AY_AY840F3005646800     1       929617  0       99.0
ROH     AY_AY840F3005646800     1       950841  0       99.0
```

Figure 2: BCFtools output using Viterbi training.

```
# This file was produced by: bcftools roh(1.5+htslib-1.5)
# The command line was: bcftools roh -m Final_Maps/map1.txt -e - -G30 VCF_Files/All/AY/AY_All-chr1.new.vcf.gz
#
# RG    [2]Sample       [3]Chromosome   [4]Start        [5]End  [6]Length (bp)  [7]Number of markers    [8]Quality
# ST    [2]Sample       [3]Chromosome   [4]Position     [5]State (0:HW, 1:AZ)   [6]Quality (fwd-bwd phred score)
ST      AY_AY840F3005646800     1       267940  0       3.0
ST      AY_AY840F3005646800     1       393248  0       54.7
ST      AY_AY840F3005646800     1       471078  0       60.6
ST      AY_AY840F3005646800     1       516404  0       65.5
ST      AY_AY840F3005646800     1       533815  0       99.0
ST      AY_AY840F3005646800     1       571340  0       99.0
ST      AY_AY840F3005646800     1       883895  0       98.9
ST      AY_AY840F3005646800     1       929617  0       99.0
ST      AY_AY840F3005646800     1       950841  0       99.0
ST      AY_AY840F3005646800     1       974586  0       99.0
ST      AY_AY840F3005646800     1       1009504 0       99.0
ST      AY_AY840F3005646800     1       1039814 0       99.0
ST      AY_AY840F3005646800     1       1114422 0       99.0
ST      AY_AY840F3005646800     1       1209308 0       99.0
ST      AY_AY840F3005646800     1       1234172 0       99.0
ST      AY_AY840F3005646800     1       1288510 0       99.0
ST      AY_AY840F3005646800     1       1336351 0       99.0
ST      AY_AY840F3005646800     1       1359951 0       99.0
ST      AY_AY840F3005646800     1       1385605 0       99.0
ST      AY_AY840F3005646800     1       1436503 0       99.0
ST      AY_AY840F3005646800     1       1513277 0       99.0
ST      AY_AY840F3005646800     1       1582828 0       99.0
ST      AY_AY840F3005646800     1       1668494 0       99.0
ST      AY_AY840F3005646800     1       1673525 0       99.0
ST      AY_AY840F3005646800     1       1760113 0       99.0
ST      AY_AY840F3005646800     1       1782724 0       99.0
ST      AY_AY840F3005646800     1       1851399 0       99.0
ST      AY_AY840F3005646800     1       1896112 0       99.0
ST      AY_AY840F3005646800     1       1921756 0       99.0
ST      AY_AY840F3005646800     1       1983902 0       99.0
ST      AY_AY840F3005646800     1       2049400 0       99.0
ST      AY_AY840F3005646800     1       2128924 0       99.0
ST      AY_AY840F3005646800     1       2211567 0       99.0
ST      AY_AY840F3005646800     1       2313042 0       99.0
ST      AY_AY840F3005646800     1       2393763 0       99.0
```

Figure 3: BCFtools output without using Viterbi training

FID	IID	PHE	CHR	SNP1	SNP2	POS1	POS2	KB	NSNP	DENSITY	PHOM	PHET
AY	AY840F3005646800	-9	1	ARS-BFGL-NGS-103348	Hapmap41389-BTA-95970	67055698	99642111	32586	473			
AY	AY840F3005646986	-9	1	BTB-00012128	BTA-109574-no-rs	28599381	38811666	10212	165	0		
AY	AY840F3007398804	-9	1	ARS-BFGL-NGS-36767	Hapmap50624-BTA-22932	10776728	16145053	5368	50			
AY	AY840F3007398804	-9	1	ARS-BFGL-NGS-107257	ARS-BFGL-NGS-70523	148756389	151125563	2369	43			
AY	AY840F3007398834	-9	1	ARS-BFGL-NGS-16466	BTB-01084177	267940	16277524	16009	238	0	0	
AY	AY840F3007398834	-9	1	ARS-BFGL-NGS-98138	Hapmap51743-BTA-54330	136443961	140835240	4391	70			
AY	AY840F3007398834	-9	1	BTA-15333-no-rs	ARS-BFGL-NGS-85061	157656358	158229218	572	38	0		

Figure 4: BCFtools output (with Viterbi training, as seen in Figure 2) converted using python to resemble PLINK ROH output file.

```
# This file was produced by bcftools stats (1.5+htslib-1.5) and can be plotted using plot-vcfstats.
# The command line was: bcftools stats  VCF_Files/All/AY/AY_All-chr1.new.vcf.gz
#
# Definition of sets:
# ID    [2]id   [3]tab-separated file names
ID      0       VCF_Files/All/AY/AY_All-chr1.new.vcf.gz
# SN, Summary numbers:
# SN    [2]id   [3]key  [4]value
SN      0       number of samples:      6560
SN      0       number of records:      2571
SN      0       number of no-ALTs:      0
SN      0       number of SNPs: 2571
SN      0       number of MNPs: 0
SN      0       number of indels:       0
SN      0       number of others:       0
SN      0       number of multiallelic sites:   0
SN      0       number of multiallelic SNP sites:       0
```

Figure 5: Statistic log for BCFtools output (as seen in Figure 2).

Table 4: Distribution of class length for Scenario 1. Each breed has been divided into three sample sizes, and the means of Round 1 to Round 5 (random samples) for each sample size are reported, with standard errors presented in brackets (±SEx10^{-3}).

		Class Length (Mb)				
		Short ROH		Medium ROH		Long ROH
Breed	**Sample Size**	**≤2**	**2-4**	**4-8**	**8-16**	**≥16**
Ayrshire	100	0.13 (2.20)	0.34 (2.52)	0.29 (2.72)	0.16 (1.93)	0.079 (2.95)
	500	0.14 (1.53)	0.34 (1.24)	0.29 (2.38)	0.16 (0.64)	0.083 (0.77)
	1000	0.14 (0.58)	0.34 (1.05)	0.29 (0.66)	0.16 (0.72)	0.082 (0.99)
	All	0.14	0.34	0.29	0.16	0.082
Brown Swiss	100	0.090 (1.82)	0.26 (2.75)	0.31 (2.43)	0.22 (3.62)	0.12 (1.50)
	500	0.093 (0.69)	0.27 (1.24)	0.30 (1.28)	0.22 (0.99)	0.12 (0.95)
	1000	0.093 (0.38)	0.27 (0.72)	0.30 (0.51)	0.22 (0.38)	0.12 (0.89)
	All	0.094	0.27	0.30	0.22	0.12
Guernsey	100	0.12 (2.03)	0.34 (3.85)	0.30 (3.84)	0.17 (2.70)	0.081 (2.36)
	500	0.13 (0.63)	0.34 (0.61)	0.29 (1.03)	0.16 (0.86)	0.077 (0.53)
	1000	0.13 (0.23)	0.34 (0.24)	0.29 (0.22)	0.16 (0.27)	0.078 (0.11)
	All	0.13	0.34	0.29	0.16	0.078
Jersey	100	0.11 (2.47)	0.33 (4.19)	0.30 (2.70)	0.18 (2.86)	0.082 (2.30)
	500	0.12 (0.71)	0.32 (0.84)	0.31 (1.12)	0.17 (1.16)	0.080 (0.81)
	1000	0.12 (0.46)	0.32 (1.97)	0.31 (0.93)	0.17 (0.53)	0.080 (0.52)
	All	0.12	0.32	0.31	0.17	0.080
Holstein	100	0.14 (1.76)	0.30 (2.30)	0.28 (1.38)	0.18 (2.49)	0.094 (1.95)
	500	0.15 (1.80)	0.30 (1.53)	0.28 (0.72)	0.18 (1.35)	0.094 (1.07)
	1000	0.15 (0.83)	0.30 (1.19)	0.28 (1.46)	0.18 (0.48)	0.094 (0.59)
	All	0.15	0.30	0.28	0.18	0.094

Table 5: Distribution of class length for Scenario 2. Each breed has been divided into three sample sizes, and the means of Round 1 to Round 5 (random samples) for each sample size are reported, with standard errors presented in brackets ($SEx10^{-3}$).

		Class Length (Mb)				
		Short ROH		Medium ROH		Long ROH
Breed	**Sample Size**	**≤2**	**2-4**	**4-8**	**8-16**	**≥16**
Ayrshire	100	0.20 (3.27)	0.35 (4.19)	0.25 (1.94)	0.13 (1.94)	0.065 (0.87)
	500	0.20 (1.04)	0.35 (1.12)	0.25 (0.98)	0.13 (0.64)	0.070 (0.81)
	1000	0.20 (2.18)	0.35 (4.88)	0.25 (0.78)	0.13 (1.25)	0.066 (1.62)
	All	0.18	0.35	0.26	0.14	0.072
Brown Swiss	100	0.14 (3.09)	0.30 (2.71)	0.27 (1.52)	0.19 (1.60)	0.10 (1.42)
	500	0.14 (0.81)	0.31 (0.12)	0.27 (0.81)	0.19 (0.49)	0.10 (0.96)
	1000	0.15 (0.31)	0.30 (0.22)	0.27 (0.83)	0.19 (0.69)	0.10 (0.52)
	All	0.13	0.30	0.28	0.19	0.10
Guernsey	100	0.19 (2.45)	0.37 (1.25)	0.25 (0.70)	0.13 (2.07)	0.064 (1.76)
	500	0.19 (0.54)	0.37 (0.64)	0.25 (0.42)	0.13 (0.59)	0.060 (0.36)
	1000	0.19 (0.27)	0.37 (0.23)	0.25 (0.29)	0.13 (0.19)	0.060 (0.19)
	All	0.26	0.35	0.22	0.12	0.053

Table 6: Distribution of class length for Scenario 3. Each breed has been divided into three sample sizes, and the means of Round 1 to Round 5 (random samples) for each sample size are reported, with standard errors presented in brackets ($SEx10^{-3}$).

		Class Length (Mb)				
		Short ROH		Medium ROH		Long ROH
Breed	**Sample Size**	**≤2**	**2-4**	**4-8**	**8-16**	**≥16**
Ayrshire	100	0.20 (1.29)	0.36 (4.10)	0.25 (2.09)	0.13 (2.10)	0.065 (1.60)
	500	0.20 (1.18)	0.35 (1.16)	0.25 (1.09)	0.13 (0.86)	0.068 (0.87)
	1000	0.20 (0.60)	0.35 (1.00)	0.24 (0.60)	0.13 (0.47)	0.069 (0.41)
	All	0.25	0.34	0.22	0.12	0.062
Brown Swiss	100	0.15 (1.45)	0.30 (1.29)	0.26 (2.31)	0.18 (1.99)	0.10 (2.76)
	500	0.15 (0.46)	0.30 (1.23)	0.26 (1.68)	0.18 (1.06)	0.098 (0.59)
	1000	0.15 (0.85)	0.31 (0.60)	0.26 (0.62)	0.18 (0.48)	0.096 (0.81)
	All	0.21	0.29	0.24	0.17	0.088
Jersey	100	0.19 (3.15)	0.35 (2.48)	0.26 (3.20)	0.14 (1.69)	0.063 (0.76)
	500	0.20 (0.75)	0.35 (0.94)	0.26 (0.68)	0.14 (0.81)	0.062 (0.53)
	1000	0.20 (0.81)	0.35 (0.52)	0.26 (0.60)	0.14 (0.75)	0.062 (0.50)
	All	0.19	0.35	0.26	0.14	0.062
Holstein	100	0.24 (2.01)	0.29 (1.81)	0.24 (2.68)	0.15 (2.08)	0.080 (1.17)
	500	0.24 (0.57)	0.29 (1.09)	0.24 (1.69)	0.15 (1.24)	0.080 (1.00)
	1000	0.24 (0.91)	0.29 (0.61)	0.24 (0.70)	0.15 (0.73)	0.079 (0.63)
	All	0.19	0.30	0.26	0.16	0.086

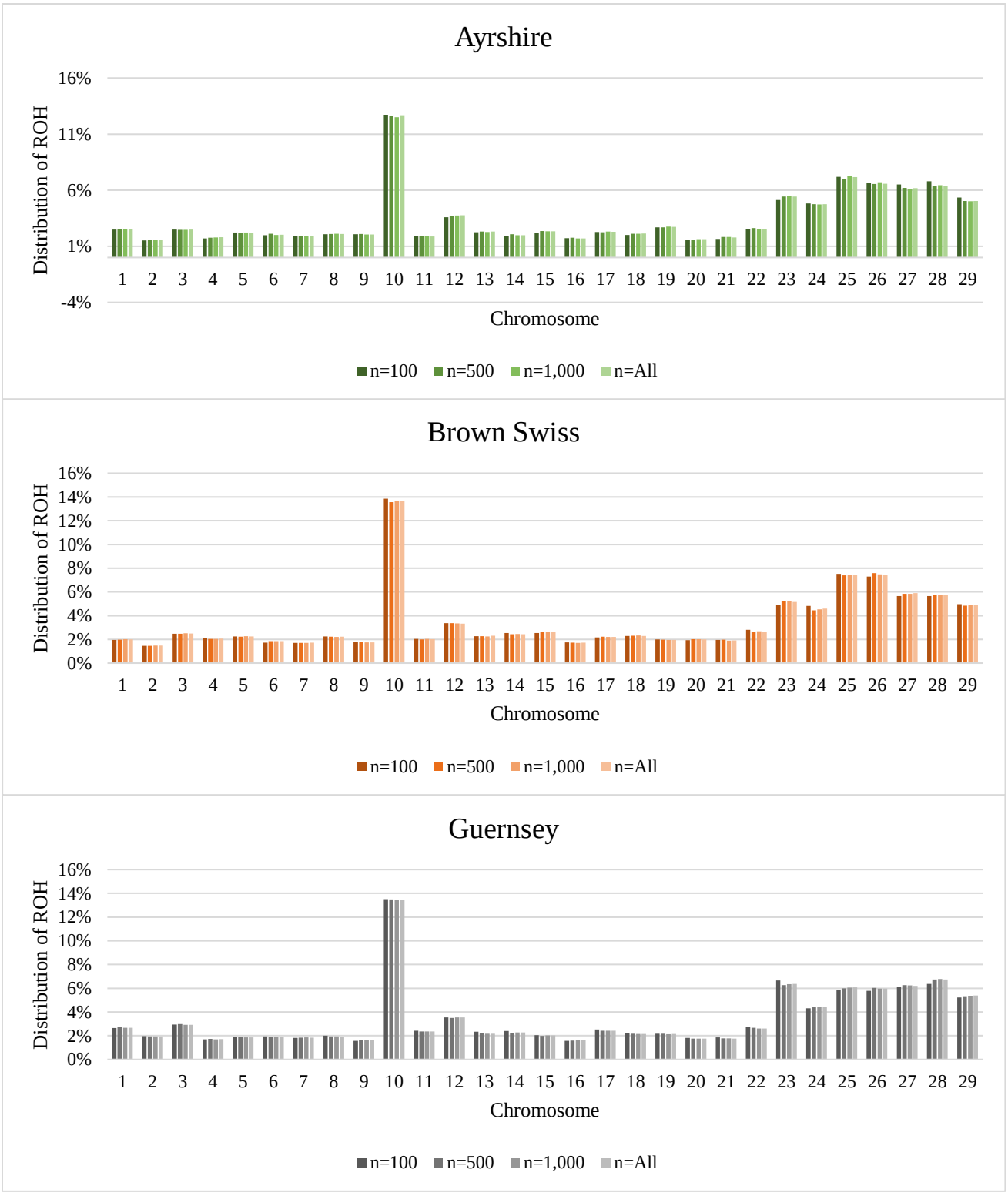
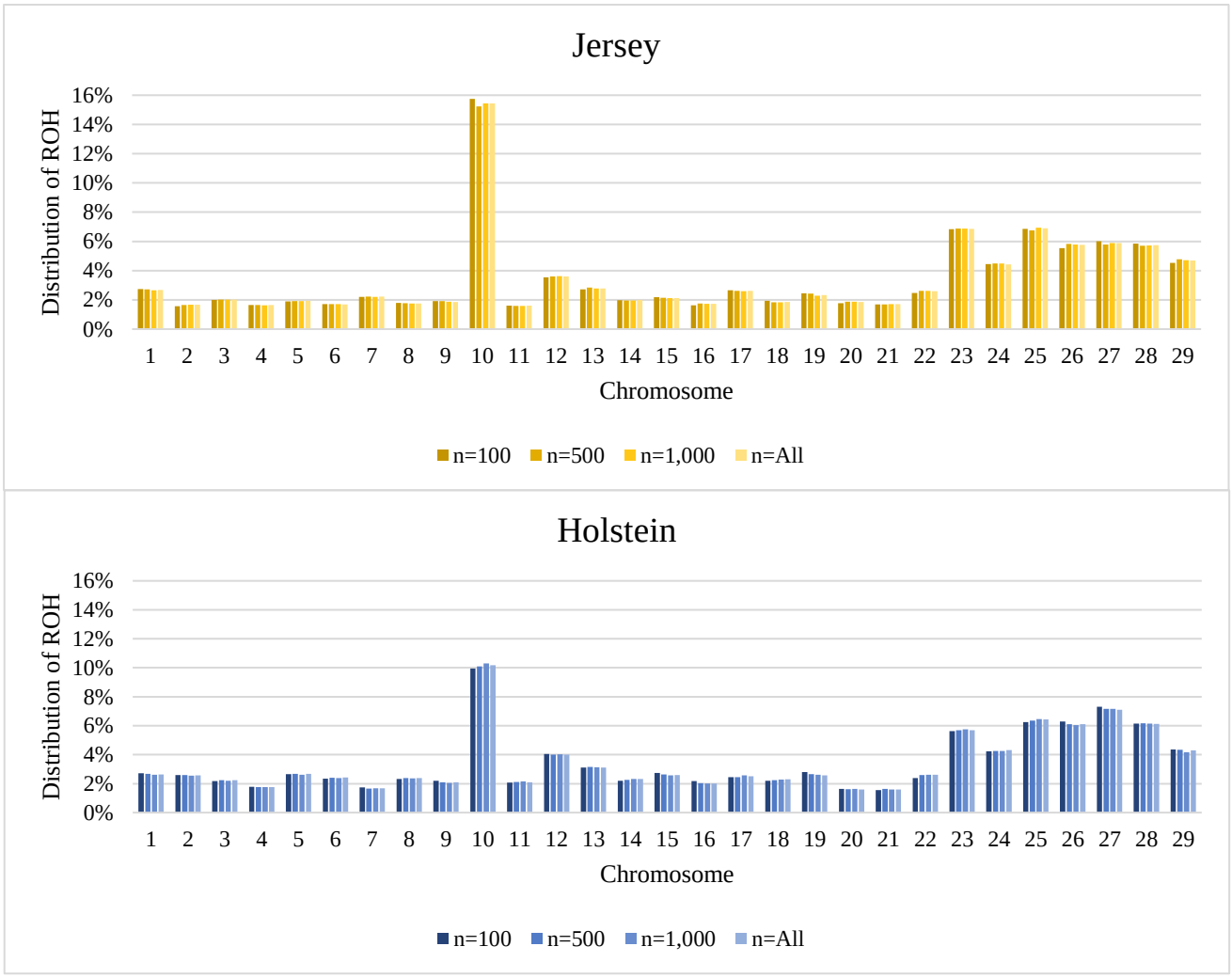


Figure 6: Distribution of ROH across the genome for samples sizes (n=100, n=500, n=1,000) and for the population of each breed in Scenario 1. Ayrshire (SE $\pm 6.78 \times 10^{-5}$ to $\pm 1.70 \times 10^{-3}$); Brown Swiss (SE $\pm 3.76 \times 10^{-5}$ to $\pm 1.77 \times 10^{-3}$); Guernsey (SE $\pm 2.65 \times 10^{-5}$ to $\pm 2.26 \times 10^{-3}$); Jersey (SE $\pm 2.5 \times 10^{-5}$ to $\pm 1.78 \times 10^{-3}$); Holstein (SE $\pm 1.36 \times 10^{-4}$ to $\pm 2.48 \times 10^{-3}$).


Jersey
Distribution of ROH
16%
14%
12%
10%
8%
6%
4%
2%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Chromosome
n=100 n=500 n=1,000 n=All
Holstein
Distribution of ROH
16%
14%
12%
10%
8%
6%
4%
2%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Chromosome
n=100 n=500 n=1,000 n=All

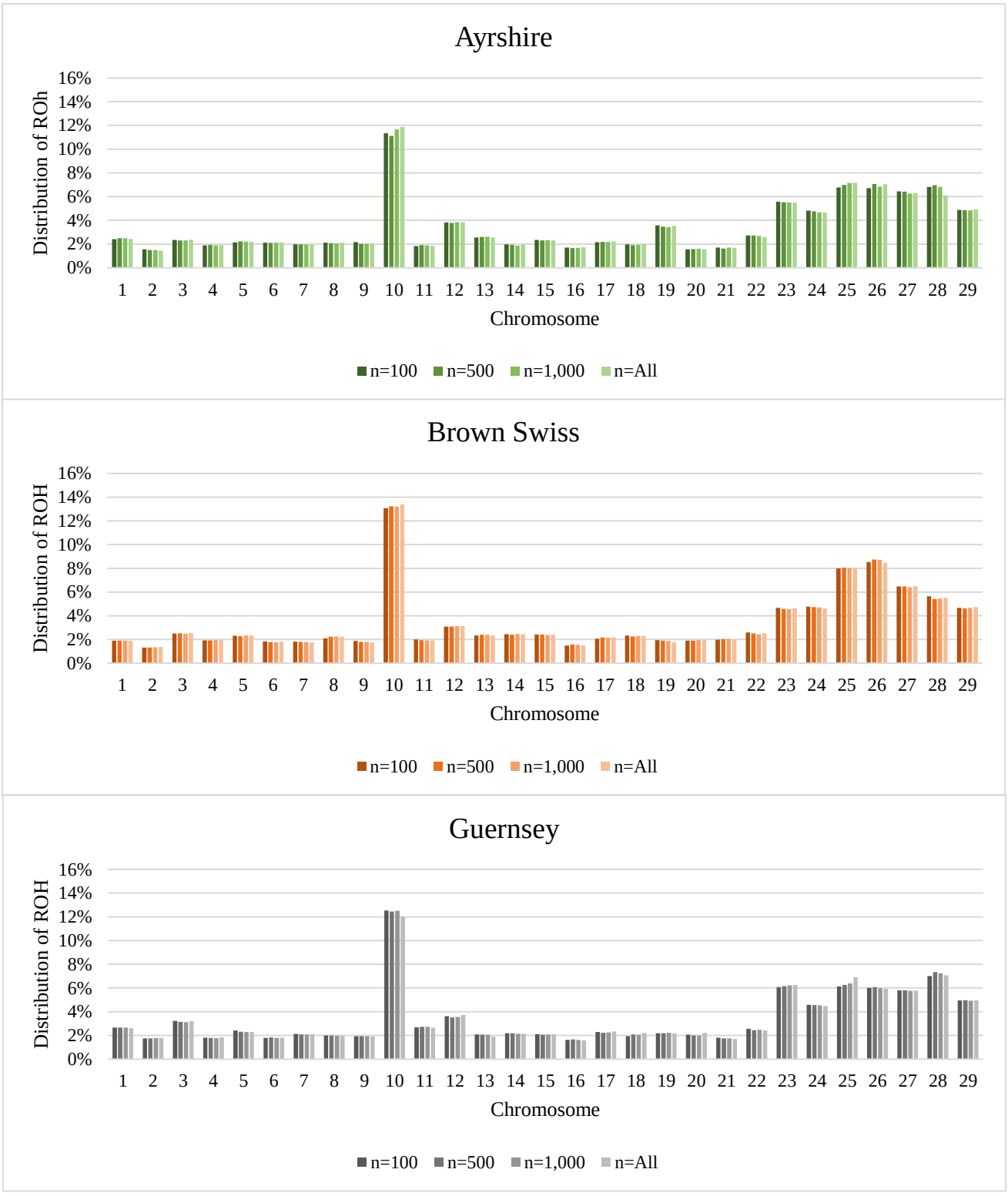


Figure 7: Distribution of ROH across the genome for samples sizes (n=100, n=500, n=1,000) and for the population of each breed in Scenario 2. Ayrshire (SE ±9,46x10^{-5} to ±1.68x10^{-3}); Brown Swiss (SE ±4.59x10^{-5} to ±1.96x10^{-3}); Guernsey (SE ±2.60x10^{-5} to ±2.01x10^{-3}).

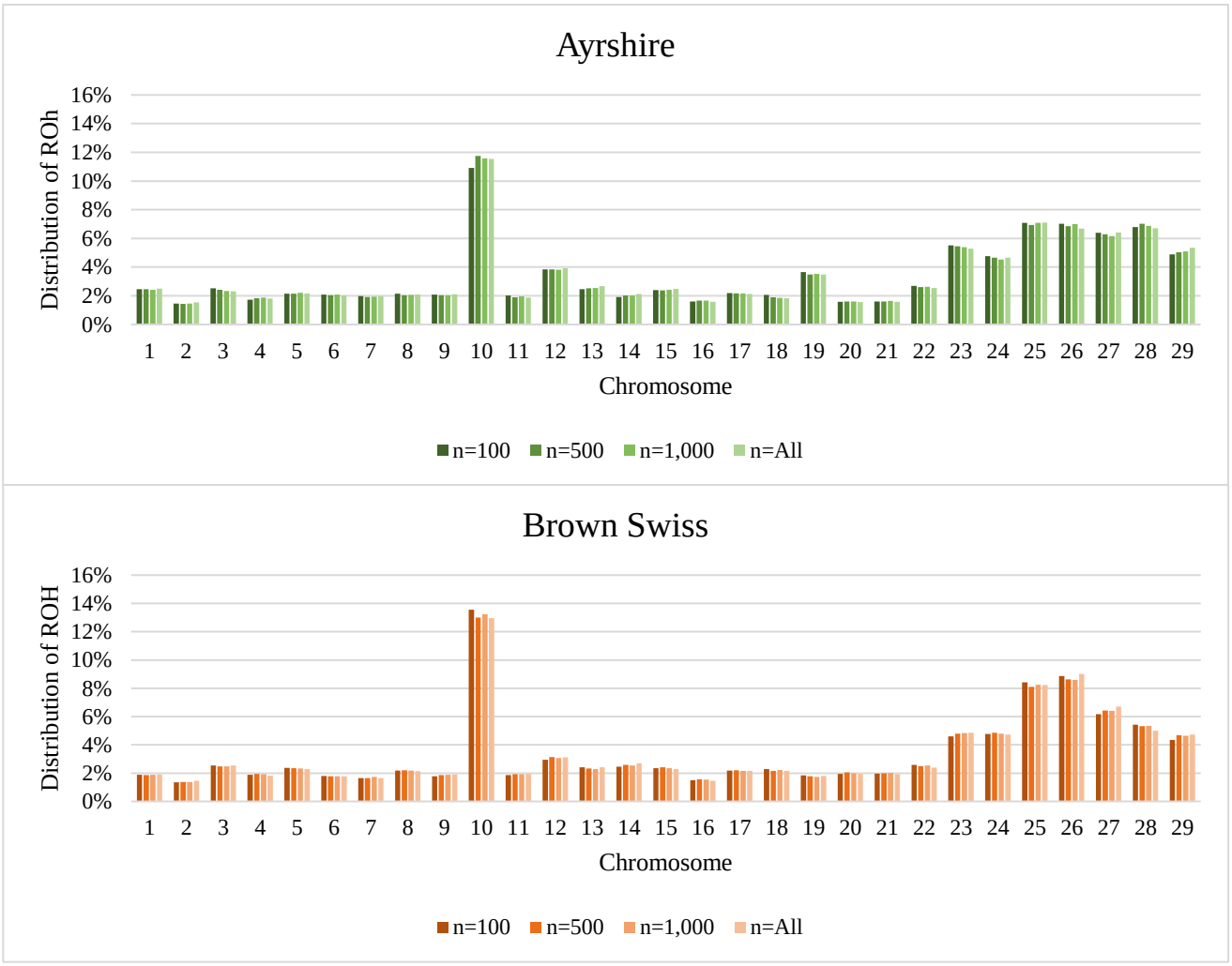


Figure 8: Distribution of ROH across the genome for samples sizes (n=100, n=500, n=1,000) and for the population of each breed in Scenario 3. Ayrshire ($\pm 1.38 \times 10^{-4}$ to $\pm 1.92 \times 10^{-3}$); Brown Swiss (SE $\pm 3.69 \times 10^{-5}$ to $\pm 1.83 \times 10^{-3}$); Jersey (SE $\pm 6.47 \times 10^{-5}$ to $\pm 1.13 \times 10^{-3}$); Holstein (SE $\pm 1.00 \times 10^{-4}$ to $\pm 2.15 \times 10^{-3}$).

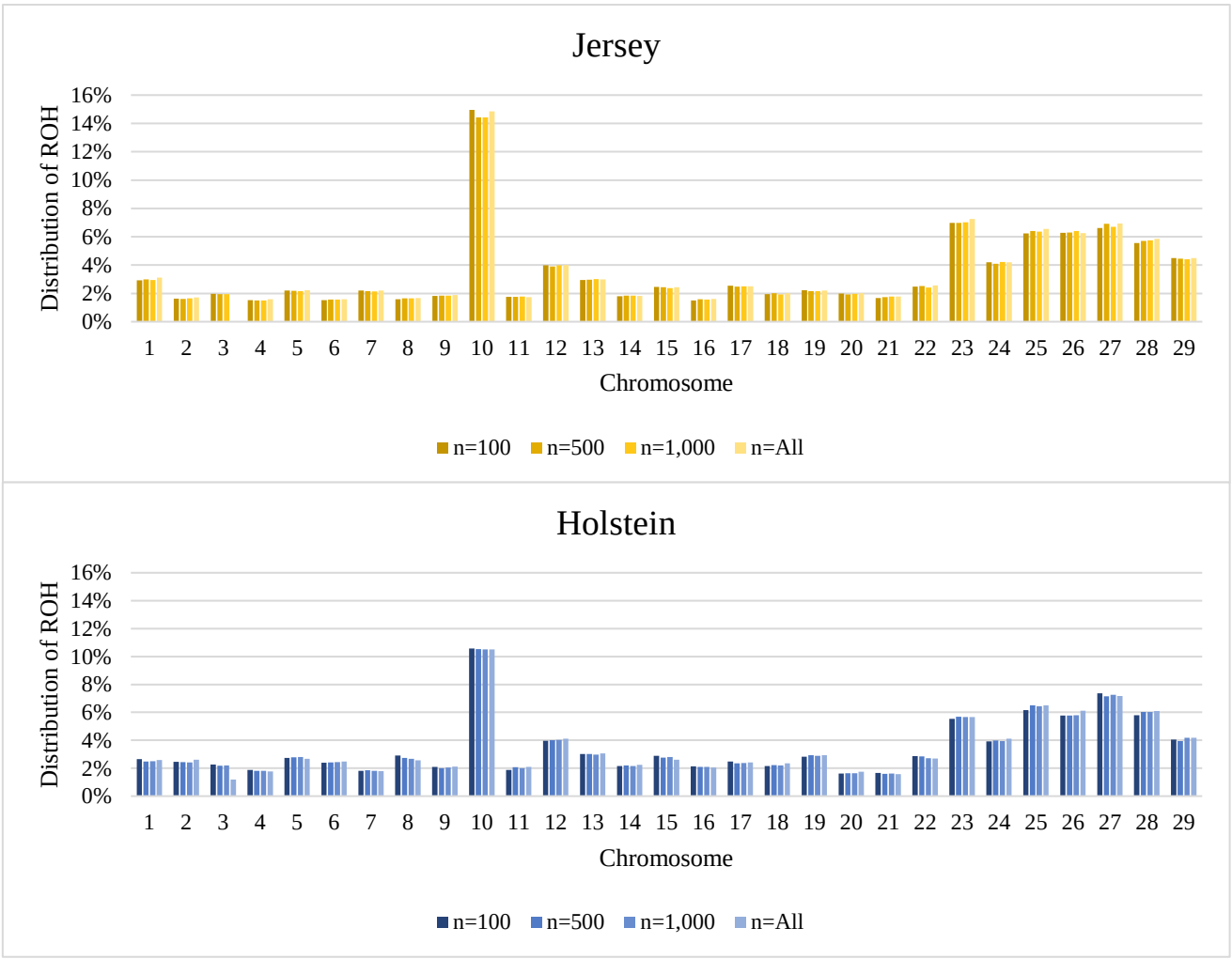

Jersey
Distribution of ROH
16%
14%
12%
10%
8%
6%
4%
2%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Chromosome
n=100 n=500 n=1,000 n=All
Holstein
Distribution of ROH
16%
14%
12%
10%
8%
6%
4%
2%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
Chromosome
n=100 n=500 n=1,000 n=All

Table 7: FROH, FPED, mean ROH length (Mb), and mean number of ROH per breed, scenario, and sample size. Standard errors presented in brackets ([a]SEx10^{-3}, [b]SEx10^{-2}, [c]SEx10^{-6}, [d]SEx10^{-5}).

Breed	Scenario	Sample Size	F_{ROH}[a]	F_{PED}[a]	Average ROH Length (Mb)[b]	Average Number ROH
Ayrshire	Scenario 1	100	0.09 (1.10)	0.052 (2.70)	6.54 (8.58)	37 (0.18)
		500	0.095 (0.69)	0.054 (0.55)	6.59 (2.84)	38 (0.13)
		1000	0.094 (0.41)	0.054 (0.32)	6.58 (2.82)	38 (4.67)[b]
		All	0.095 (0.41)	0.054 (0.32)	6.58 (1.94)	38 (0.10)
	Scenario 2	300	0.10 (1.20)	0.054 (1.30)	5.77 (4.88)	46 (3.59)[b]
		1500	0.10 (0.77)	0.055 (0.43)	5.90 (2.62)	45 (0.13)
		3000	0.10 (0.35)	0.054 (0.29)	5.85 (2.14)	46 (8.10)[b]
		All	0.10 (0.41)	0.054 (0.32)	6.03 (1.74)	43 (0.11)
	Scenario 3	100	0.10 (1.90)	0.053 (1.70)	5.50 (9.28)	50 (0.36)
		500	0.10 (0.92)	0.054 (0.40)	5.87 (3.29)	46 (0.25)
		1000	0.10 (0.32)	0.054 (0.40)	5.83 (1.58)	46 (7.73)[b]
		All	0.11 (0.42)	0.054 (0.32)	5.50 (1.57)	50 (0.12)
Brown Swiss	Scenario 1	100	0.14 (1.40)	0.065 (1.10)	8.12 (7.48)	46 (0.34)
		500	0.14 (0.68)	0.063 (0.56)	8.05 (2.63)	47 (9.41)[b]
		1000	0.14 (0.56)	0.063 (0.41)	8.04 (2.40)	47 (6.93)[b]
		All	0.14 (5.84)[c]	0.063 (0.37)	8.03 (1.79)	47 (0.10)
	Scenario 2	100	0.15 (0.87)	0.064 (0.57)	7.17 (2.53)	56 (0.24)
		500	0.15 (0.81)	0.063 (0.43)	7.15 (2.70)	57 (8.78)[b]
		1000	0.15 (0.22)	0.063 (0.14)	7.14 (0.96)	57 (8.35)[b]
		All	0.15 (0.44)	0.063 (0.40)	7.25 (1.59)	55 (0.11)
	Scenario 3	100	0.16 (1.60)	0.064 (0.85)	6.57 (5.89)	64 (0.17)
		500	0.15 (0.53)	0.063 (0.33)	7.08 (1.38)	58 (0.17)
		1000	0.15 (0.37)	0.063 (0.25)	7.03 (1.49)	58 (8.72)[b]
		All	0.16 (0.45)	0.063 (0.37)	6.57 (1.46)	64 (0.12)

Breed	Scenario	Sample Size	F_{ROH}[a]	F_{PED}[a]	Average ROH Length (Mb)[b]	Average Number ROH
Guernsey	Scenario 1	100	0.12 (0.95)	0.060 (0.73)	6.73 (4.24)	47 (0.24)
		500	0.12 (0.48)	0.057 (0.45)	6.57 (2.52)	47 (6.65)[b]
		1000	0.12 (0.084)	0.057 (0.12)	6.59 (0.25)	47 (3.97)[b]
		All	0.12 (0.94)	0.057 (0.94)	6.59 (0.11)	47 (4.24)[b]
	Scenario 2	100	0.13 (1.0)	0.055 (1.10)	5.70 (5.81)	62 (0.39)
		500	0.13 (0.29)	0.057 (0.30)	5.69 (0.76)	62 (0.11)
		1000	0.13 (0.13)	0.057 (1.60)	5.69 (0.51)	62 (4.80)[b]
		All	0.14 (0.98)	0.057 (0.94)	5.24 (3.08)	71 (0.32)
Jersey	Scenario 1	100	0.14 (1.80)	0.052 (1.5)	6.85 (6.99)	54 (0.37)
		500	0.14 (1.10)	0.053 (0.71)	6.80 (3.95)	55 (0.17)
		1000	0.14 (0.50)	0.053 (0.35)	6.80 (1.67)	55 (8.77)[b]
		All	0.14 (0.13)	0.054 (0.11)	6.81 (0.46)	55 (3.29)[b]
	Scenario 3	100	0.16 (1.50)	0.054 (1.20)	5.87 (3.87)	73 (0.55)
		500	0.16 (0.81)	0.053 (0.63)	5.82 (2.19)	75 (0.21)
		1000	0.16 (0.53)	0.054 (0.37)	5.81 (1.50)	75 (0.21)
		All	0.16 (0.13)	0.054 (0.11)	5.87 (0.38)	73 (3.86)[b]
Holstein	Scenario 1	100	0.11 (1.60)	0.064 (1.30)	7.21 (3.84)	38 (0.30)
		500	0.10 (0.41)	0.063 (0.47)	7.03 (2.84)	38 (8.15)[b]
		1000	0.10 (0.43)	0.064 (0.38)	7.02 (1.56)	38 (0.11)
		All	0.10 (8.39)[d]	0.063 (6.38)[d]	7.02 (0.40)	38 (2.02)[b]
	Scenario 3	100	0.11 (1.60)	0.064 (1.5)	6.62 (3.95)	42 (0.48)
		500	0.11 (0.59)	0.064 (0.34)	6.21 (2.82)	47 (0.12)
		1000	0.10 (0.33)	0.063 (0.19)	6.19 (1.32)	46 (4.61)[b]
		All	0.10 (8.40)[d]	0.063 (6.38)[d]	6.62 (0.37)	42 (2.15)[b]

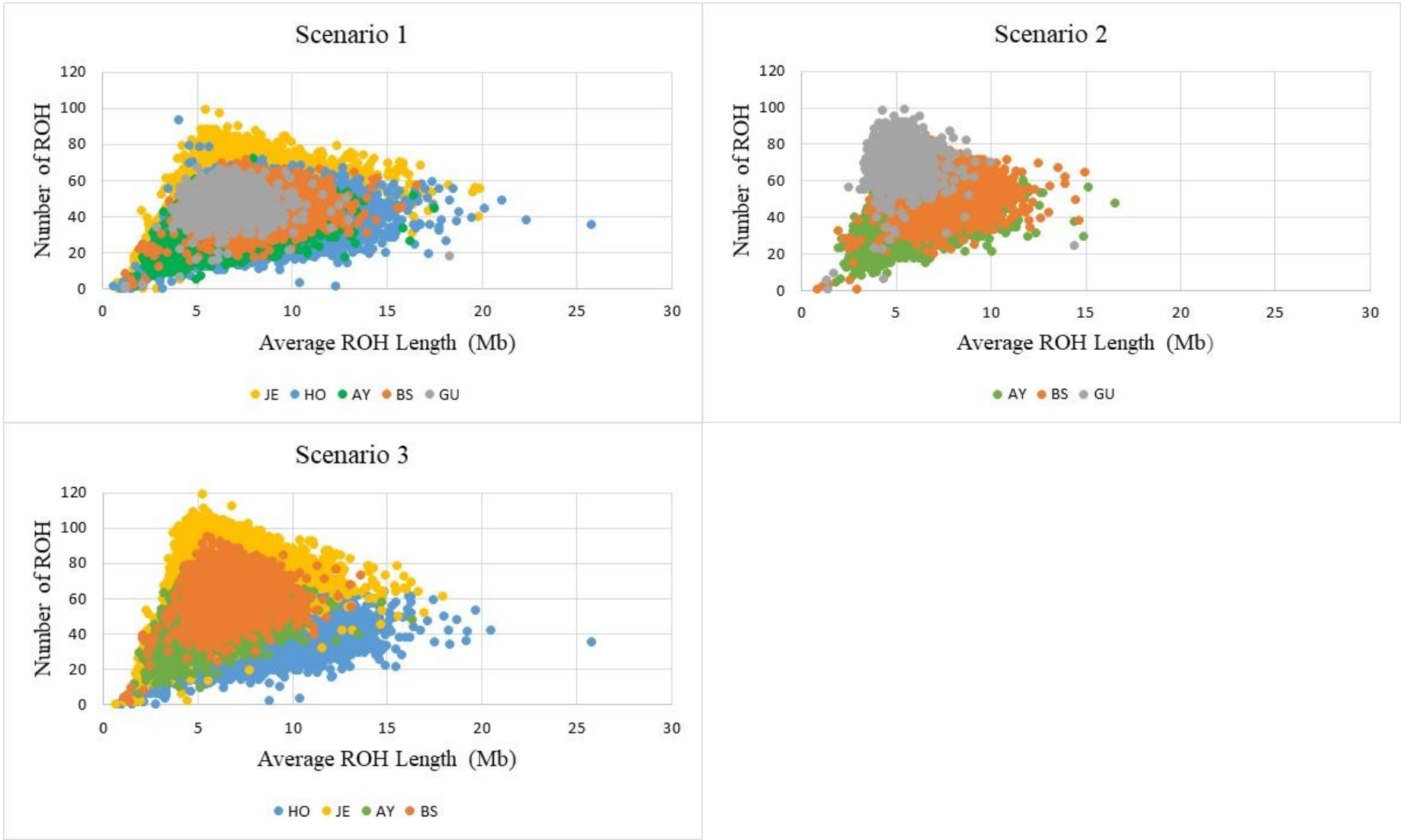


Figure 9: Average ROH length vs. number of ROH per individual for each scenario and breed (AY – Ayrshire; BS – Brown Swiss; GU – Guernsey; JE – Jersey; HO – Holstein).

APPENDIX I

BCFtools Analysis – Chapter 2 – M.Sc. Thesis

Calista Vogelzang

M.Sc. Thesis 2018

##

To run this script, you will need:

Binary PLINK files (BED/BIM/FAM) of filtered data

Genetic map file

Optional:

Animal subset text files

This script includes commands to create and compress VCF files, run BCFtools ROH detection and statistics, and run a python script to convert the format of BCFtools output into format of PLINK ROH output

This script has been created to run multiple scenarios (single- and multiple-breed analyses), breeds, sample sizes, and sampled rounds in a loop

Run this script where your filtered data is located

##

```
breed=c("AY", "BS", "GU","JE","HO","AYBSGU","AYBSJEHO")
subset_sizes=c("100","500","1000")
subset_sizes=c("All")

for (a in breed){
  for (r in subset_sizes){
    for (b in 1:5){
      for (i in 1:29){

        #Creating VCF files (SNP already filtered, genetic maps already prepared)
        com1=paste("plink --cow --chr ",i," --bfile Data/All/ADSA2018_SNPs_",a,"-72 --keep
Data/",r,"/rand",a,r,"_Round",b,".txt --recode vcf --out
VCF_Files/",r,"/",a,"/rand",a,r,"_Round",b,"-chr",i, sep="")

        #for 'All animals'
        #com1=paste("plink --cow --chr ",i," --bfile Data/All/ADSA2018_SNPs_",a,"-72 --keep
Data/All/subset_only_",a,".txt --recode vcf --out VCF_Files/All/",a,"/",a,"_All-chr",i, sep="")
        system(com1)

        #Compressing VCF files
        com2=paste("bgzip -c VCF_Files/",r,"/",a,"/rand",a,r,"_Round",b,"-chr",i,".vcf >
VCF_Files/",r,"/",a,"/rand",a,r,"_Round",b,"-chr",i,".vcf.gz", sep="")
        com3=paste("zcat VCF_Files/",r,"/",a,"/rand",a,r,"_Round",b,"-chr",i,".vcf.gz  | bgzip -c >
VCF_Files/",r,"/",a,"/rand",a,r,"_Round",b,"-chr",i,".new.vcf.gz && tabix
VCF_Files/",r,"/",a,"/rand",a,r,"_Round",b,"-chr",i,".new.vcf.gz",sep="")

        #for 'All animals'
        #com2=paste("bgzip -c VCF_Files/All/",a,"/",a,"_All-chr",i,".vcf >
VCF_Files/All/",a,"/",a,"_All-chr",i,".vcf.gz", sep="")
```

```
#com3=paste("zcat VCF_Files/All/",a,"/",a,"_All-chr",i,".vcf.gz  | bgzip -c > VCF_Files/All/",a,"/",a,"_All-chr",i,".new.vcf.gz && tabix VCF_Files/All/",a,"/",a,"_All-chr",i,".new.vcf.gz",sep="")


system(com2)

system(com3)


#Running BCFtools analysis

com4=paste("bcftools stats VCF_Files/",r,"/",a,"/rand",a,r,"_Round",b,"-chr",i,".new.vcf.gz > Results/",r,"/",a,"/stats_ROH-",a,"-",r,"-Round",b,"-",i, sep="")

com5=paste("bcftools roh -m Final_Maps/map",i,".txt -e - -G30 -V 1e-10 VCF_Files/",r,"/",a,"/rand",a,r,"_Round",b,"-chr",i,".new.vcf.gz  > Results/",r,"/",a,"/ROH-",a,"-",r,"-Round",b,"-",i,sep="")



#for 'All animals'

#com4=paste("bcftools stats VCF_Files/All/",a,"/",a,"_All-chr",i,".new.vcf.gz  > Results/All/",a,"/stats_ROH-",a,"-All-",i, sep="")

#com5=paste("bcftools roh -m Final_Maps/map",i,".txt -e - -G30 -V 1e-10 VCF_Files/All/",a,"/",a,"_All-chr",i,".new.vcf.gz  > Results/All/",a,"/ROH-",a,"-All-",i,sep="")


system(com4)

system(com5)


#Converting BCFtools output using a python script into a format detectRUNS will be able to read (in this case PLINK ROH output format)

com6=paste("python Conversion_BCFtools.py Results/",r,"/",a,"/ROH-",a,"-",r,"-Round",b,"-",i," Results/",r,"/",a,"/Converted_Files/ROH-",a,"-",r,"-Round",b,"-",i," Data/All/mapFile_ALLSNP.map", sep="")


#for 'All animals'
```

```
      #com6=paste("python Conversion_BCFtools.py Results/All/",a,"/ROH-",a,"-All-",i," Results/All/",a,"/Converted_Files/ROH-",a,"-All-",i," Data/All/mapFile_ALLSNP.map", sep="")
      system(com6)


    }
   }
  }
}
```

CHAPTER 3: GENERAL CONCLUSIONS

Inbreeding is unavoidable in populations under selection, especially in dairy populations with small effective population sizes (Zhang *et al*., 2015a). In some instances, inbreeding is desirable, such as fixing alleles in populations that have a large impact on production traits (i.e. DGAT1), however in most cases, inbreeding has been found to compromise the fitness of an individual and by extension, a population. Inbreeding depression is an issue that the dairy industry is currently facing, and efforts to understand and manage inbreeding are being employed not only in Canada, but worldwide. Runs of homozygosity (**ROH**) can be used to calculate and manage inbreeding at the genomic level. The objective of this thesis was to characterize ROH in five dairy breeds by focusing on how certain factors (i.e. single- and. multiple-breed analyses and sample sizes) affect ROH detection as this will influence how the industry will make breeding decisions using ROH-based inbreeding ($\mathbf{F_{ROH}}$) information in the future.

Factors that may influence ROH detection were explored theoretically in Chapter 1 and practically in Chapter 2. Factors discussed included SNP chip density, ROH-detecting programs, single- and multiple-breed analyses differences, breed differences, and the impact of sample size on characterizing ROH (VanRaden *et al.*, 2011; Ferenčaković *et al.*, 2013a Ferenčaković *et al.*, 2013b; Hay and Rekaya, 2015; Forutan *et al*., 2018). In Chapter 1, the sensitivity of medium (50K) and high density (**HD**) genotype data were compared. 50K genotypes were noted to be less sensitive to detecting short ROH, and therefore were not dense enough to use in population history or breed history studies in which characterizing ancient inbreeding is needed (Kirin *et al*., 2010; Ferenčaković *et al.*, 2013b). 50K genotypes were found to be highly correlated with whole-genome sequencing (**WGS**) data, however, indicating that medium density genotypes were sufficient to study recent inbreeding in dairy breeds (Zhang *et al.*, 2015b). In Chapter 2, the number of short

ROH appeared to increase between single- and multiple-breed analyses suggesting multiple-breed analyses may underestimate recent inbreeding. This will need to be further investigated in the future with amended sample sizes.

In Chapter 1, factors affecting the variation of SNP inclusion for ROH analyses were discussed at length. Genotype errors, differences in allele frequencies between breeds, and parameters imposed during filtering have all been shown to change the number and subset of SNP included in analyses, and consequently the length and number of ROH identified (Ferenčaković *et al.*, 2013b). Multiple-breed analyses augmented SNP biases, especially if small populations were being analysed with a larger population such as Holstein. Creating a subset from large populations before SNP filtering and ROH detection may reduce such biases, but further research is required to ascertain if this method is practical.

Chapter 2 outlined the importance of proper filtering methods and how different scenarios and sample sizes led to different SNP being included in 'clean' data. The way BCFtools managed different datasets was investigated in this chapter as well, and it was determined that the quality of data that is input into the program greatly impacts ROH detection, and by extension, ROH characterization. This has implications for incorporating ROH into breeding programs with the intent of managing herd and population rates of inbreeding.

It is known that ROH can be used to detect differences between breeds. In Chapter 2, the distribution of ROH across length classes and across the genome were found to be unique, while some general trends were maintained across breeds. In each breed, a large proportion of ROH were found on BTA 10. This chromosome has several genes related to udder traits, and as all five breeds considered in this study are selected primarily for milk production, this result is unsurprising (Schrooten *et al.*, 2000; Hiendleder *et al.*, 2003; Ashwell *et al.*, 2005).

In Chapter 2, the precision of different sample sizes was investigated to determine if a subset of animals could accurately estimate the level of F_{ROH} of the population. Precisions were found to increase as sample size increased, however repeated animals across rounds may have biased these results. It was also found that smaller sample sizes (n=1,000) detected ROH that was representative of the entire population in single-breed analyses. This result indicates that computational time and demand could be saved by using a subset of animals from large datasets to detect and characterize ROH.

This work has the potential to be the foundation of creating a standardized method to identify ROH. This will be useful to the dairy industry as well as academic institutions as it will provide more accurate calculations of F_{ROH} and a more accurate picture of the structure of the genome of an individual. This in turn will allow for improved mating decisions, maintained genetic variation for continued genetic gain, and further deconstruction and understanding of the mechanisms behind inbreeding and inbreeding depression, Next steps would include running each scenario with the same sample size to be able to make across-scenario comparisons, as well as to employ a bootstrapping technique to address any bias imposed by repeated animals. Additionally, performing further characterization analyses such as calculating F_{ROH} per year or per generation, will aid in analysing the levels of homozygosity in an individual that are due to identical by descent segments, resulting in more accurate F_{ROH} estimates.

FINAL REMARKS

Runs of homozygosity can be used to calculate genomic inbreeding and potentially be adapted into methods that can be used to manage inbreeding levels while maintaining genetic gain. The results presented in this thesis address the challenges of identifying and characterizing ROH, and how these factors can affect the accuracy of calculating F_{ROH}. The number and length of ROH

can be heavily influenced by genotype densities, genotyping errors, SNP filtering parameters, breed differences in allele frequencies, differences between ROH-detecting programs, and differences in populations sizes of breeds in multiple-breed analyses. Further work is required to construct a standard definition and method of identification of ROH that is robust enough to account for these different factors before F_{ROH} can effectively be employed by the dairy industry to make breeding decisions.

REFERENCES

Ashwell, M. S., D. W. Heyen, J. I. Weller, M. Ron, T. S. Sonstegard, C. P. Van Tassell, and H. A. Lewin. 2005. Detection of quantitative trait loci influencing conformation traits and calving ease in Holstein-Friesian cattle. J. Dairy Sci. 88:4111-4119.

Ferenčaković, M., E. Hamzić, B. Gredler, T. R. Solberg, G. Klemetsdal, I. Curik, and J. Sölkner. 2013a. Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. J. Anim. Breed. Genet. 130:286-293. https://dx.doi.org/10.1111/jbg/12012.

Ferenčaković, M., J. Sölkner, and I. Curik. 2013b. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genet. Sel. Evol. 45:42. https://dx.doi.org/10.1186/1297-9686-45-42.

Forutan, M., S. A. Mahyari, C. Baes, N. Melzer, F. S. Schenkel, and M. Sargolzaei. 2018. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. BMC Genomics 19:98. https://dx.doi.org/10.1186/s12864-018-4453-z.

Hay, E. H., and R. Rekaya. 2015. A multi-compartment model for genomic selection in multi-breed populations. Livest. Sci. 177:1-7. https://dx.doi.org/10.1016/j.livsci.2015.03.027.

Hiendleder, S., H. Thomsen, N. Reinsch, J. Bennewitz, B. Leyhe-Horn, C. Looft, N. Xu, I. Medjugorac, I. Russ, C. Kuhn, G. A. Brockmann, J. Blumel, B. Brenig, F. Reinhardt, R. Reents, G. Averdunk, M. Schwerin, M. Forster, E. Kalm, and G. Erhardt. 2003. Mapping of QTL for body conformation and behavior in cattle. J. Hered. 94:496-506. https://dx.doi.org/10.1093/jhered/esg090.

Howard, J. T., M. Haile-Mariam, J. E. Pryce, and C. Maltecca. 2015b. Investigation of regions

impacting inbreeding depression and their association with the additive genetic effect for United States and Australia Jersey dairy cattle. BMC Genomics 16:813. https://dx.doi.org/10.1186/s12864-015-2001-7

Kirin, M., R. McQuillan, C. S. Franklin, H. Campbell, P. M. McKeigue, and J. F. Wilson. 2010. Genomic runs of homozygosity record population history and consanguinity. PLoS ONE. 5:e13996. https://dx.doi.org/10.1371/journal.pone.0013996.

Schrooten, C., H. Bovenhuis, W. Coppieters, and J. A. M. Van Arendonk. 2000. Whole genome scan to detect quantitative trait loci for conformation and functional traits in dairy cattle. J. Dairy Sci. 83:795-806. https://dx.doi.org/10.3168.jds.S0022-0302(00)74942-3.

VanRaden, P. M., K. M. Olson, G. R. Wiggans, J. B. Cole, and M. E. Tooker. 2011. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J. Dairy Sci. 94:5673-5682. https://dx.doi.org./10.3168/jds.2011-4500.

Zhang, Q., B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. 2015a. Runs of homozygosity and distribution of functional variants in the cattle genome. BMC Genomics 16:542. https://dx.doi.org/10.1186/s12864-015-1715-x.

Zhang, Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015b. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics 16:88. https://dx.doi.org/10.1186/s12863-015-0227-7.