Exploiting whole genome sequence variants in cattle breeding

Qianqian Zhang

Supervisors committee in Aarhus University

Main supervisors

Dr. Goutam Sahana
Senior Researcher, Center for Quantitative Genetics and Genomics
Aarhus University, Denmark

Dr. Mario Calus
Associate professor, Animal Breeding and Genomics
Wageningen University & Research, the Netherlands

Co-supervisors

Prof. Dr. Mogens Sandø Lund
Professor of Center for Quantitative Genetics and Genomics
Aarhus University, Denmark

Dr. Bernt Guldbrandtsen
Associate Professor of Center for Quantitative Genetics and Genomics
Aarhus University, Denmark

This research was conducted under the joint auspices of the Graduate School of Science and Technology (GSST), Aarhus University, Denmark and the Graduate School Wageningen Institute of Animal Sciences (WIAS), Wageningen University and is part of the Erasmus Joint Doctorate Program “EGS-ABG”.

Thesis committee Wageningen University

Promotor
Prof. Dr Henk Bovenhuis
Professor of Animal Breeding and Genomics
Wageningen University & Research, the Netherlands

Prof. Dr. Mogens Sandø Lund
Professor of Center for Quantitative Genetics and Genomics
Aarhus University, Denmark

Co-promotors
Dr. Goutam Sahana
Senior Researcher, Center for Quantitative Genetics and Genomics
Aarhus University, Denmark

Dr. Mario Calus
Associate professor, Animal Breeding and Genomics
Wageningen University & Research, the Netherlands

Other members
Prof. Dr Fred van Eeuwijk, Wageningen University & Research, the Netherlands
Prof. Just Jensen, Aarhus University, Denmark
Prof. Dr Johann Sölkner, University of Natural Resources and Life Sciences, Vienna, Austria
Prof. Dr Hubert Pausch, Swiss Federal Institute of Technology, Zurich, Switzerland

This research was conducted under the joint auspices of the Graduate School of Science and Techonology (GSST), Aarhus University, Denmark and the Graduate School Wageningen Institute of Animal Sciences (WIAS), Wageningen University and is part of the Erasmus Joint Doctorate Program “EGS-ABG”.

Exploiting whole genome sequence variants in cattle breeding

Unraveling the distribution of genetic variants and role of rare variants in genomic evaluation

Qianqian Zhang

Thesis

submitted in fulfillment of the requirements for the joint degree of doctor between
Aarhus University
by the authority of the Head of Graduate School of Science and Technology and
Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Head of Graduate School of Science and Technology at Aarhus University and the Academic Board of Wageningen University
to be defended in public
on Tuesday 19 December 2017
at 9.00 a.m. in Foulum, Aarhus University

Qianqian Zhang
Exploiting whole genome sequence variants in cattle breeding

PhD thesis, Aarhus University, Foulum, Denmark and Wageningen University, Wageningen, the Netherlands (2017)
With summaries in English and Danish
ISBN: 978-87-93643-14-7
DOI: https://doi.org/10.18174/428523

Abstract

Zhang, Q. (2017). Exploiting whole genome sequence variants in cattle breeding: Unraveling the distribution of genetic variants and role of rare variants in genomic evaluation. Joint PhD thesis, Aarhus University, Denmark and Wageningen University & Research, the Netherlands.

The availability of whole genome sequence data enables to better explore the genetic mechanisms underlying different quantitative traits that are targeted in animal breeding. This thesis presents different strategies and perspectives on utilization of whole genome sequence variants in cattle breeding. Using whole genome sequence variants, I show the genetic variation, recent and ancient inbreeding, and genome-wide pattern of introgression across the demographic and breeding history in different cattle populations. Using the latest genomic tools, I demonstrate that recent inbreeding can accurately be estimated by runs of homozygosity (ROH). This can further be utilized in breeding programs to control inbreeding in breeding programs. In chapter 2 and 4, by in-depth genomic analysis on whole genome sequence data, I demonstrate that the distribution of functional genetic variants in ROH regions and introgressed haplotypes was shaped by recent selective breeding in cattle populations. The contribution of whole genome sequence variants to the phenotypic variation partly depends on their allele frequencies. Common variants associated with different traits have been identified and explain a considerable proportion of the genetic variance. For example, common variants from whole genome sequence associated with longevity have been identified in chapter 5. However, the identified common variants cannot explain the full genetic variance, and rare variants might play an important role here. Rare variants may account for a large proportion of the whole genome sequence variants, but are often ignored in genomic evaluation, partly because of difficulty to identify associations between rare variants and phenotypes. I compared the powers of different gene-based association mapping methods that combine the rare variants within a gene using a simulation study. Those gene-based methods had a higher power for mapping rare variants compared with mixed linear models applying single marker tests that are commonly used for common variants. Moreover, I explored the role of rare and low-frequency variants in the variation of different complex traits and their impact on genomic prediction reliability. Rare and low-frequency variants contributed relatively more to variation for health-related traits than production traits, reflecting the potential of improving prediction reliability using rare and low-frequency variants for health-related traits. However, in practice, only marginal improvement was observed using selected rare

and low-frequency variants when combined with 50k SNP genotype data on the reliability of genomic prediction for fertility, longevity and health traits. A simulation study did show that reliability of genomic prediction could be improved provided that causal rare and low-frequency variants affecting a trait are known.

To my grandfather, always deeply missed.

Contents

1

General introduction

1.1 Utilizing genomic information in cattle breeding

Animal breeding has been very successful in achieving genetic gain to improve economically relevant traits in breeding goal even without knowing underlying genetic mechanisms controlling the traits. However, to reach a higher genetic gain and more accurate selection, it is of great importance to understand genetic mechanisms underlying including genetic variants affecting the traits, but also how selective breeding shapes the distribution of these genetic variants in the current population. If all the causal variants affecting a trait can be identified, the breeding goal for the trait will be achieved simply using the known genetic mechanisms with 100% accuracy in selection (Goddard, 2017). Recently, genomic information generated is helping to explore these genetic mechanisms (Dekkers, 2012, Kiplagat et al., 2012, Brondum et al., 2015). For example, selection for functional traits such as health-related traits will probably be improved with mining the deleterious genetic variants in the population. Therefore, I believe that the next generation animal breeding will be directly selecting on the variants controlling the trait with more accurate biological knowledge underlying different traits, once identified.

The availability of genomic data has provided the unique opportunity to examine the changes in genomic composition in different populations over the period under selective breeding, and to start identifying the causal genetic variants underlying different phenotypic traits using statistical methods. The causal genetic variants and their combinations are not randomly distributed across individuals and populations. Artificial selection has shaped the variation landscape in cattle genomes, by eliminating the variants with negative effects and increasing the frequency of alleles with favorable effects on phenotypic traits (Xu et al., 2015). Using genomic data, it is possible to study the effect of intense artificial selection on the distribution of genetic variants in cattle populations.

The genomic data are routinely generated for genomic selection in livestock breeding (Bouquet and Juga, 2013). Genomic selection in cattle uses genotype information such as 50k and high density (HD) chip data to predict breeding values for selection candidates (Meuwissen, 2007). By using genomic information, the time and costs needed for measuring and obtaining the phenotypic data from large number of offspring has been by-passed (Pryce and Daetwyler, 2012, Bouquet and Juga, 2013). In fact, large improvements in genetic gain can be achieved in dairy cattle compared with conventional progeny testing (12%-100%) due to significant

reduction of the generation interval (Meuwissen, 2007, Pryce and Daetwyler, 2012, Bouquet and Juga, 2013). With the availability of whole genome sequencing data, it is possible to map variants associated with traits at high resolution. It is also possible to explore the role of rare variants, which are not tagged by 50k and HD chip SNPs, in the variation of different traits. Thereby, the identified variants including rare variants, potentially affecting traits can be exploited in genomic selection in cattle.

One aim of this thesis is to contribute to the understanding of the landscape of distribution of genetic variants in cattle genomes, and how selective breeding including crossbreeding has shaped the distribution of these genetic variants, thereby, this kind of information can later be utilized in genomic selection in cattle breeding. Another aim of this thesis is to provide insight into the identification of, especially rare, genetic variants associated with traits and the impact of including them in genomic selection model along with common variants included in SNP arrays.

1.2 Whole genome sequence variants

Nowadays, the genomic information used in cattle breeding comprises primarily single nucleotide polymorphisms (SNPs) with different densities SNP arrays (50k and HD). However, it is well known that due to the procedure used to select SNPs, 50k or HD SNP chips are suffering from ascertainment bias which affects the allele frequency spectrum (Nielsen, 2000, Eller, 2001, Nielsen and Signorovitch, 2003, Clark et al., 2005, Foll et al., 2008, Guillot and Foll, 2009). Consequently, the population genetic parameters estimated using SNP chip genotype data tend to be biased; even the accuracy in genomic evaluation can be affected by ascertainment bias from SNP chips (Heslot et al., 2013, Wientjes et al., 2015).

The availability of next generation sequencing data enables the study of genetic variants in single base pairs without suffering from ascertainment bias. However, the number of individuals with sequence data such as the available 1000 bull genome project (Daetwyler et al., 2014) is still much lower than the number needed for a training population for genomic selection and powerful investigation of association between the genetic variants and phenotypes. The number of individuals with higher density genotypes can be substantially increased by imputation, which infers the missing genotypes in lower density markers to higher

density markers (Ma et al., 2013, Brondum et al., 2014). Imputation is relying on the linkage disequilibrium (LD) between markers, which is defined as the non-random association of alleles between different loci in a given population (Marchini and Howie. 2010). In cattle, whole genome sequence variants are usually imputed in 2 steps i.e. from 50k markers to HD markers and from imputed HD markers to whole genome sequence variants (Brondum et al., 2014). The imputation for common variants is usually very accurate. However, the accuracy of imputation is relatively low for rare variants due to the low LD between mostly common markers on the SNP chips and rare sequence variants. It has been shown in cattle that imputation accuracy is dropping very fast when allele frequency is less than 0.1 especially when the size of the reference population is small (Bouwman and Veerkamp, 2014, Brondum et al., 2014, van Binsbergen et al., 2014). The accuracy for imputing rare variants might be improved for individuals with sequenced relatives if pedigree information is used to inform the imputation. Combining the sequence from individuals in multiple populations (e.g. 1000 bull genome project) can also improve the imputation accuracy (Brondum et al., 2014). Certain software, such as IMPUTE2, tends to perform better in imputation for rare variants, but the accuracy is still relatively low compared with common variants (Brondum et al., 2014). Therefore, accurate imputation of rare variants is still a challenge.

1.3 Genetic variation and inbreeding

With the advent of next generation sequencing technology, the genetic variation in an individual genome can be fully characterized at base-pair resolution and compared with other individuals' genomes from the same or different populations. The average nucleotide diversity across individual genomes within a population reflects the levels of genetic variation, genetic diversity and effective population size in this population. Furthermore, with whole genome sequence variants, it is possible to examine the genomic regions with different levels of nucleotide diversity across loci on a chromosome, and the persistence of loci in a population under different evolutionary and artificial forces.

The genetic variation is not randomly distributed across the genome and some genomic regions tend to have higher or lower nucleotide diversity (Bosse et al., 2012). A set of genetic variants in an organism that is inherited together from a single parent is defined as a haplotype (Cox et al., 2016). If two haplotypes transmitted from both parents are identical contiguous homozygous stretches in an

individual genome, then the resulting homozygous stretches in an individuals' genome is defined as Runs of Homozygosity (ROH) (McQuillan et al., 2008). The length and occurrence of ROH on the genome is affected by the local recombination rate, the number of generations since the common ancestor and selection on favorable alleles across generations (Bosse et al., 2012, Szpiech et al., 2013). Selection on favorable alleles can result in excess of homozygosity on loci in LD with the favorable alleles i.e. genetic hitchhiking (Smith and Haigh, 2007). The haplotypes containing favorable alleles with hitchhiked alleles will exhibit high extent of homozygosity, and soon be fixed in the population during the process of selection. As a result, these selected haplotypes will accumulate on the same location of genomes across individuals. The genetic variants located on the genomic regions showing excess of homozygosity across individuals show a signal of selection in the population (Sabeti et al., 2002). In addition, mating two close relatives will result in long stretches of ROH on the genome and high levels of inbreeding for the offspring. This will result in inbreeding depression, which consequently leads to reduced selection response in breeding programs and reduced fitness due to the increased risk of homozygosity for deleterious alleles (Miglior et al., 1992, Gonzalez-Recio et al., 2007). Accurate estimate of inbreeding is crucial for monitoring inbreeding in breeding programs to avoid inbreeding depression.

The inbreeding coefficient is defined as the probability that two alleles in an individual are identical by descent (IBD) relative to a base population where all alleles are assumed unrelated (Wright, 1922). ROH can identify genomic regions which are IBD (Broman and Weber, 1999, McQuillan et al., 2008). The inbreeding coefficient can be calculated as the average proportion of ROH on a genome (Purfield et al., 2012). Theoretically, ROH computed from next generation sequencing data is supposed to be the most accurate because the genetic variants are called on a single base-pair resolution and estimation does not suffer from sampling of genetic variants (Ferencakovic et al. , 2013a, Ferencakovic et al., 2013b). Therefore, an important question is whether inbreeding coefficient estimated based on ROH is comparable with other estimators using other source of data such as pedigree and SNP chip genotype data. Several estimates of inbreeding coefficients based on genomic relationship matrix (GRM), excess of homozygosity and correlation of uniting gametes using different density of markers (sequence data, 50k and HD SNP chip) are highly dependent on allele frequencies (Wright, 1948, Broman and Weber, 1999, VanRaden, 2008, Yang et al., 2011, Zhang et al., 2015a). Different allele frequencies across loci in different populations may lead to

various estimation of inbreeding coefficients, therefore, compared with ROH based estimate, unstable estimation might be observed (Zhang et al., 2015a). In this thesis, different estimators of inbreeding coefficient estimated from different data sources were compared to investigate which estimator is most appropriate.

1.4 Introgression

A population is characterized by its genetic variation shaped by different evolutionary forces such as introgression, selection and drift. When gene flow happens from one population to another population, the genetic variation tends to be higher in the hybrid population. The gene flow between populations or even between species is called "introgression". The introgressed haplotypes in the hybrid individual's genome can reflect the history of source populations and the process of introgression (Hedrick, 2013). After years of introgression, the haplotypes are fragmented by recombination and the identification of the introgressed haplotypes is complicated by other evolutionary forces such as selection (Bosse et al., 2014). Although how evolutionary forces drive the distribution of introgressed haplotypes is not certain, some of these introgressed haplotypes indeed remain in the admixed genomes containing favorable alleles with selective advantage from a source population (i.e. adaptive introgression) (Crispo et al., 2011, Hedrick, 2013). Years of selection in livestock species after introgression has in general resulted in increasing frequency of favorable alleles and decreasing the frequency of or even eliminated unfavorable background alleles in the hybrid genomes (Hedrick, 2013).

The implementation of selective breeding in dairy cattle in Europe mostly happened after Second World War (Hartwig et al. 2015). In that period, the main breeding goal of cattle breeding was to improve milk related traits. Therefore, high milk yielding breed have been introduced to the local breed to improve milk production. It has been shown that years of introgression results in significant improvement of milk related traits towards high milk yield performance in the local breed (Hartwig et al., 2015). It will be very interesting to explore how the distribution of introgressed haplotypes is shaped by adaptive introgression and selection, and what kind of genetic variants from different source populations are maintained across generations in the admixed breed. This kind of genomic information can be further utilized in genomic selection to improve different economic traits. In this thesis, due to the known history of admixture from the

high-yielding breeds Holstein and Brown Swiss, the Nordic Red Dairy cattle (RDC) was used as an example to illustrate the global admixture and genome-wide signals of introgression. This study can serve as a model to understand the change of genetic architecture for complex traits during adaptive introgression.

1.5 Distribution of genetic variants

Cattle populations have been subject to intense artificial selection for more than 60 years and the distribution of genetic variants is heavily shaped by artificial selection (Kim et al., 2013). Genomic studies on the distribution of whole genome sequence variants can reveal the genomic architecture shaping by different evolutionary processes for different populations and may contribute to the identification of the causal variants affecting different economic traits. With availability of whole genome sequence variants, it is possible to study the distribution of functional variants shaped by artificial selection in cattle populations. One type of important variants with large impact are those annotated as “deleterious”, due to their large impact on amino acid change and protein synthesis (Velankar et al., 2013). The accumulation of deleterious variants in ROH or non-ROH regions leads to inbreeding depression (Szpiech et al., 2013). The investigations of patterns of genetic variants which are homozygous across the genome (i.e. in ROH regions) can contribute insights into mapping genetic variants associated with inbreeding depression (Gonzalez-Recio et al. 2007). The functional variants are enriched in different length of ROH. The longer an ROH is, the more deleterious variants are enriched in human population (Szpiech et al., 2013). In human populations, deleterious variants are more enriched in long ROH regions reflecting more recent inbreeding (Szpiech et al., 2013). Some of the annotated “deleterious” variants, however, could also have beneficial effect on economic traits. The observed standing genetic variants might be accumulating in stretches of homozygosity in the genome when the standing variants are functional and under positive selection (Zhang et al. 2015b). Therefore, in cattle populations, the enrichment of these variants with large impact in ROH regions can be due to not only inbreeding but also artificial selection (Zhang et al. 2015b). Years of selection will result in a common sharing of relatively short ROH enriched with genetic variants with large impact among individuals in the population (Ferencakovic et al., 2013a). There is a positive correlation between enrichment of functional variants and different length of ROH regions affecting by the different history of populations (Kim et al., 2013, Purfield et al., 2017). In this thesis, I studied the distribution of functional variants

enriched in different length of ROH regions in cattle populations and compared with the distribution observed in human genome.

Similarly, the distribution of genetic variants in introgressed regions is also shaped by different evolutionary forces as well as artificial breeding practice (Bosse et al. 2014). To improve performance of a local breed, high performance breeds are hybridized with the local breed (Hartwig et al. 2015). With following selection, the frequencies of beneficial alleles increase and the performance of local breed improves. The introgressed haplotypes with beneficial effects will be preserved under the force of long time artificial selection. Detection of genomic regions showing signature of selection is necessary to identify regions that have both been introgressed and under selection. If the introgressed haplotypes are indeed selection candidates, only the ones with selective advantage will be preserved. These introgressed haplotypes have a high genetic diversity shortly after introgression, which decreases over generations due to selection. Different methods and strategies are necessary to be implemented to get a full overview of selection signatures. These introgressed and selected regions have the potential to be utilized further in genomic selection in livestock breeding. In this thesis, I used different methods to detect selection signature using whole genome sequence data and compare the signals under selection with introgression signals in Danish cattle breeds.

1.6 Common vs. Rare variants

There is an ongoing debate in human genetics on the contribution of common and rare variants to the genetic basis of complex traits (Schork et al., 2009, Gibson, 2012). Common variants are the variants with minor allele frequencies (MAF) higher than 5%. Under the infinitesimal model, it is assumed that an infinite number of unlinked loci are affecting a quantitative trait and each with an infinitesimal effect (Norton and Pearson, 1976, Vilela et al., 2008). The infinitesimal model has been applied in animal breeding for breeding value estimation and led to great genetic progress (Bouquet and Juga, 2013). In fact, it has been suggested that common variants with small effect sizes collectively are the main source of genetic variation for quantitative traits (Yang et al., 2010, Yang et al., 2015). Numerous quantitative trait loci have been identified to be associated with complex traits and they are mostly common variants. For example, the top QTLs affecting fat yield in dairy cattle are common variants (Iso-Touru et al., 2016).

However, common variants that are identified to be associated with complex traits only account for a small fraction of traits' estimated heritabilities (Maher, 2008, Manolio et al., 2009, Gibson, 2012). The unidentified part of estimated heritability is the so-called "missing heritability", which has received a lot of attention in human genetics (Manolio et al., 2009). Theoretical and empirical studies suggest that rare variants (defined by convention as those that have MAF <1%), which are poorly captured by single-nucleotide polymorphism chips, may play a significant role in quantitative traits' variation (Gibson, 2012). The advent of whole genome sequence variants enables to study rare variants. To identify rare variants with small effects on complex traits, very large sample size are needed in association studies (Swartz et al., 2014).

Rare and low-frequency variants have received a lot attention in human genetics (Gibson, 2012). For human diseases, the important role of rare and low-frequency variants has been confirmed for monogenetic diseases (Duncan et al. 2014). More and more efforts have been made to understand the effects of rare and low-frequency variants and close the gap of missing heritability in human complex traits (Yang et al., 2015). Rare and low-frequency variants play an important role in the genetic basis of human diseases. Recently, studies with very large sample size shown that rare and low-frequency variants contribute substantially to the phenotypic variance of complex traits (e.g. prostate cancer, blood metabolites and height) in humans (Yang et al., 2015, Mancuso et al., 2016, Long et al., 2017), while a study on the genetic architecture of type 2 diabetes shows that lower-frequency variants do not have a major role in predisposition to type 2 diabetes (Cauchi et al., 2008). To identify these rare and low-frequency variants with small or intermedium effects, very large sample size for association studies is required (Swartz et al., 2014). Until now, the contribution of rare and low-frequency variants has not been explored in cattle genetics. With large number of individuals with imputed sequence data and phenotypes open opportunity to study rare variants contribution in cattle phenotypes.

1.7 Genome-wide association study

1.7.1 Association mapping for common variants

The aim of a genome-wide association study is to identify genetic variants associated with a trait of interest. The most common method is to fit genetic

markers across the whole genome one by one in the single marker test (Cauchi et al., 2008). It is expected that whole genome sequence will include the causal variants. However, significant SNPs from mapping based on whole genome sequence data are not only the causal variants but majority are variants in LD with the causal variants affecting the phenotypic traits (Pearson and Manolio, 2008, Bush and Moore, 2012), and it is not possible to distinguish between them only based on association statistics (Andersson and Georges, 2004). Verification of significant SNPs in another population might filter out part of the variants in LD with causal variants. Follow-up functional studies of highly significant SNPs are needed to distinguish the causal variants and variants in LD with causal variants.

1.7.2 Association mapping for rare variants

Methods for genome-wide association studies with common SNP variants are well established, and have been successful in identifying common variants associated with complex traits (Iso-Touru et al., 2016, Bush and Moore, 2012). Mapping rare variants, however, remains challenging. Several classes of statistical methods have been developed for the analysis of rare variants especially for 'case-control' designs and complex traits in human genetics, both for samples of unrelated and related individuals (Madsen and Browning, 2009, Price et al., 2010, Neale et al., 2011, Wu et al., 2011). One broad class of such methods is "burden tests", which collapses multiple rare variants in a region into a single variable to represent a genetic burden score (Morgenthaler and Thilly, 2007, Li and Leal, 2008, Madsen and Browning, 2009, Price et al., 2010). Another broad class of methods is variance component tests (Neale et al., 2011, Wu et al., 2011), which aggregate individual variant statistics measuring strength of association with each site, and incorporate flexible weight functions to boost analysis power. The third category is a combination of the two approaches (burden tests and variance component tests) to exploit the strengths of both approaches e.g. (Lee et al., 2012, Jiang and McPeek, 2014). This combination of both methods will be optimally balanced by the data itself and could simultaneously detect the common effect across rare variants (as in burden tests) and the individual deviations from the average effect (as in variance component tests).

However, best of my knowledge, association studies for rare variant have not been carried out so far in cattle and other livestock species. The initiative to collate a large number of whole genome sequences (Daetwyler et al., 2014) and availability of exome sequence data in the near future could be used for mapping rare variants in cattle. Suitability of the above mentioned statistical methods developed for

mapping rare variants for quantitative traits in human still remains unexamined in data structure like cattle. One of the objectives of this thesis, therefore, was to investigate power and type I error rate for several approaches for mapping rare variants applied to cattle population.

1.8 Genomic prediction

With availability of denser markers such as 50k and HD SNP chips, genomic breeding values are estimated based on markers by estimating the effect of each marker in a process known as genomic prediction (Meuwissen, 2007, Bouquet and Juga, 2013). Genomic prediction relies on the LD between markers and causal variants controlling the trait of interest. Large numbers of markers (50k or HD markers) is used to estimate genomic breeding values, because genetic variation can be captured as markers in LD with the causal variants covering the whole genome.

The advantage of genomic selection, i.e. selection based on genomic breeding values, in dairy cattle is that it significantly reduces the generation interval and thereby improves the rate of genetic gain (Schaeffer, 2006, Bouquet and Juga, 2013). The breeding values are predicted in two steps. Firstly, the marker effects are estimated in the training population with both phenotypes and genotypes. Secondly, the estimated breeding values for the selection candidate are calculated using their genotypes by summing up the estimated marker effects calculated from the training population. The most commonly used model in genomic prediction in dairy cattle is genomic best linear unbiased prediction (GBLUP) (VanRaden, 2008).

Increased marker density is expected to increase the reliability of prediction (Meuwissen, 2009, Meuwissen and Goddard, 2010, Clark et al., 2011). However, the improvement of reliability is small for the prediction based on HD instead of 50k markers (Su et al., 2012, VanRaden et al., 2013, Erbe et al., 2014). The reliability of genomic prediction using the whole genome sequence variants had no additional improvement (Calus et al., 2016, Heidaritabar et al., 2016, Veerkamp et al., 2016). It is expected the whole genome sequence variants include majority of causal variants as well as markers in high LD with causal variants. However, at the same time adding a large number of markers with no association with the trait adds noise to the prediction. In summary, using all sequence variants in genomic prediction results in no added improvement in terms of the reliability of genomic

prediction but it may be advantageous to only include a subset of informative variants in the genomic prediction model.

Therefore, including rare variants might be the key factor to improve the accuracy of genomic prediction. In this thesis, I explored the role of rare variants in traits' variation and the impact of including rare variants on accuracy of genomic prediction in dairy cattle.

1.9 This thesis

The overall objective of this thesis is to utilize whole genome sequence variants in cattle breeding by unraveling the distribution of genetic variants and evaluating the possible role of rare variants in trait variation. The whole genome sequence variants were analyzed to reveal the genetic variation, inbreeding and introgression among different cattle populations. The genomic regions with high or low nucleotide diversity were identified and how different evolutionary and artificial processes shaped the distribution of genetic variants in these genomic regions was examined. The distribution of the genetic variants reflected the genomic architecture of modern cattle breeds is affected by intense artificial selection, gene-flow and genetic drift. The knowledge gained contributes to the understanding of the relationship between selective breeding and dynamics of genomic architecture in breeding populations. Furthermore, the availability of whole genome sequence data has enabled to study rare variants, which are poorly tagged by SNP chip. In **chapter 2**, I investigated the occurrence of ROH in different cattle breeds and the distribution of functional variants in ROH regions in cattle genomes. In **chapter 3**, the levels of genomic inbreeding were examined in different cattle breeds, and different estimators of inbreeding coefficient using different data source and methods were compared. In **chapter 4**, I used the Nordic Red Dairy cattle with past history of admixture, to explore the introgressed haplotypes and studied how selection shapes these introgressed haplotypes. In **chapter 5**, the whole genome sequence variants associated with longevity were identified in three cattle breeds. **Chapter 6** compared different genome-wide association methods to map rare variants in cattle populations. In **chapter 7**, I explored the contribution of rare and low-frequency variants in different complex traits in cattle. In **chapter 8**, the impact of including rare and low-frequency variants on the reliability of genomic prediction in cattle was evaluated. Finally, I

discussed the different future perspectives on exploiting whole genome sequence variants in the context of cattle breeding in **chapter 9**.

References

Andersson, L. and M. Georges. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. Nature Reviews Genetics 5(3):202-212.

Cox, C. B., Moore, P. D., & Ladle, R. 2016. Biogeography: an ecological and evolutionary approach. John Wiley & Sons.

Bosse, M., H. J. Megens, O. Madsen, Y. Paudel, L. A. Frantz, L. B. Schook, R. P. Crooijmans, and M. A. Groenen. 2012. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genetics 8(11):e1003100.

Bosse, M., H. J. Megens, L. A. F. Frantz, O. Madsen, G. Larson, Y. Paudel, N. Duijvesteijn, B. Harlizius, Y. Hagemeijer, R. P. M. A. Crooijmans, and M. A. M. Groenen. 2014. Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. Nature Communications 5.

Bouquet, A. and J. Juga. 2013. Integrating genomic selection into dairy cattle breeding programmes: a review. Animal 7(5):705-713.

Bouwman, A. C. and R. F. Veerkamp. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genetics 15(1):105.

Broman, K. W. and J. L. Weber. 1999. Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. American Journal of Human Genetics 65(6):1493-1500.

Brondum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. S. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15(1):728.

Brondum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. Journal of Dairy Science 98(6):4107-4116.

Bush, W. S. and J. H. Moore. 2012. Chapter 11: Genome-Wide Association Studies. PloS Computional Biology 8(12):e1002822.

Calus, M. P. L., A. C. Bouwman, C. Schrooten, and R. F. Veerkamp. 2016. Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. Genetics Selection Evolution 48(1):49.

Cauchi, S., K. Nead, H. Choquet, F. Horber, N. Potoczna, B. Balkau, M. Marre, G. Charpentier, P. Froguel, and D. Meyre. 2008. Genetic architecture of type 2 diabetes is modulated by the status of obesity. Diabetes Metab 34:A38-A38.

Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen. 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Research 15(11):1496-1502.

Clark, S. A., J. M. Hickey, and J. H. J. van der Werf. 2011. Different models of genetic variation and their effect on genomic evaluation. Genetics Selection Evolution 43(1):18.

Crispo, E., J. S. Moore, J. A. Lee-Yaw, S. M. Gray, and B. C. Haller. 2011. Broken barriers: Human-induced changes to gene flow and introgression in animals. Bioessays 33(7):508-518.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brondum, X. P. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46(8):858-865.

Dekkers, J. C. M. 2012. Application of Genomics Tools to Animal Breeding. Current Genomics 13(3):207-212.

Duncan, E., M. Brown, & E. M. Shore. 2014. The revolution in human monogenic disease mapping. Genes, 5(3), 792-803.

Eller, E. 2001. Effects of ascertainment bias on recovering human demographic history. Human Biology 73(3):411-427.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2014. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 97(10):6622-6622.

Ferencakovic, M., E. Hamzic, B. Gredler, T. R. Solberg, G. Klemetsdal, I. Curik, and J. Solkner. 2013a. Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. Journal of Animal Breeding and Genetics 130(4):286-293.

Ferencakovic, M., J. Solkner, & I. Curik. 2013b. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genetics Selection Evolution 45(1):42.

Foll, M., M. A. Beaumont, and O. Gaggiotti. 2008. An approximate Bayesian computation approach to overcome biases that arise when using amplified

fragment length polymorphism markers to study population structure. Genetics 179(2):927-939.

Gibson, G. 2012. Rare and common variants: twenty arguments. Nature Reviews Genetics 13(2):135-145.

Goddard, M. E. 2017. Can we make genomic selection 100% accurate? Journal of Animal Breeding and Genetics 134(4):287-288.

Gonzalez-Recio, O., E. L. de Maturana, and J. P. Gutierrez. 2007. Inbreeding depression on female fertility and calving ease in Spanish dairy cattle. Journal of Dairy Science 90(12):5744-5752.

Guillot, G. and M. Foll. 2009. Correcting for ascertainment bias in the inference of population structure. Bioinformatics 25(4):552-554.

Hartwig, S., R. Wellmann, R. Emmerling, H. Hamann, and J. Bennewitz. 2015. Short communication: Importance of introgression for milk traits in the German Vorderwald and Hinterwald cattle. Journal of Dairy Science 98(3):2033-2038.

Hedrick, P. W. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular Ecology 22(18):4606-4618.

Heidaritabar, M., M. P. L. Calus, H. J. Megens, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics 133(3):167-179.

Heslot, N., J. Rutkoski, J. Poland, J. L. Jannink, and M. E. Sorrells. 2013. Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. Plos One 8(9): e74612.

Iso-Touru, T., G. Sahana, B. Guldbrandtsen, M. S. Lund, and J. Vilkki. 2016. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genetics 17(1):1.

Jiang, D. and M. S. McPeek. 2014. Robust Rare Variant Association Testing for Quantitative Traits in Samples With Related Individuals. Genetic epidemiology 38(1):10-20.

Kiplagat, S. K., Limo, M. K., & Kosgey, I. S. 2012. Genetic improvement of livestock for milk production. InTech Cellular Mechanism(Animal Management and Health).

Kim, E. S., J. B. Cole, H. Huson, 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 US Holstein Cattle. PLoS One 8(11): e80813.

Lee, S., M. C. Wu, and X. H. Lin. 2012. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762-775.

Li, B. S. and S. M. Leal. 2008. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. American Journal of Human Genetics 83(3):311-321.

Long, T., M. Hicks, H. C. Yu, W. H. Biggs, E. F. Kirkness, C. Menni, J. Zierer, K. S. Small, M. Mangino, H. Messier, S. Brewerton, Y. Turpaz, B. A. Perkins, A. M. Evans, L. A. D. Miller, L. N. Guo, C. T. Caskey, N. J. Schork, C. Garner, T. D. Spector, J. C. Venter, and A. Telenti. 2017. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nature Genetics 49(4):568-578.

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

Ma, P., R. F. Brondum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. Journal of Dairy Science 96(7):4666-4677.

Madsen, B. E. and S. R. Browning. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genetics 5(2):e1000384.

Maher, B. 2008. Personal genomes: The case of the missing heritability. Nature 456(7218):18-21.

Mancuso, N., N. Rohland, K. A. Rand, A. Tandon, A. Allen, D. Quinque, S. Mallick, H. Li, A. Stram, X. Sheng, Z. Kote-Jarai, D. F. Easton, R. A. Eeles, L. Le Marchand, A. Lubwama, D. Stram, S. Watya, D. V. Conti, B. Henderson, C. A. Haiman, B. Pasaniuc, D. Reich, and P. Consortium. 2016. The contribution of rare variation to prostate cancer heritability. Nature Genetics 48(1):30.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-753.

Marchini, J., & B. Howie. 2010. Genotype imputation for genome-wide association studies. Nature reviews. Genetics, 11(7), 499.

McQuillan, R., A. L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, L. Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, O. Polasek, 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. American Journal of Human Genetics 83(3):359-372.

Meuwissen, T. 2007. Genomic selection : marker assisted selection on a genome wide scale. Journal of Animal Breeding and Genetics 124(6):321-322.

Meuwissen, T. and M. Goddard. 2010. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics 185(2):623-U338.

Meuwissen, T. H. E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genetics Selection Evolution 41(1):35.

Miglior, F., B. Szkotnicki, and E. B. Burnside. 1992. Analysis of Levels of Inbreeding and Inbreeding Depression in Jersey Cattle. Journal of Dairy Science 75(4):1112-1118.

Morgenthaler, S. and W. G. Thilly. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 615(1), 28-56.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly. 2011. Testing for an Unusual Distribution of Rare Variants. PLoS Genetics 7(3): e1001322.

Nielsen, R. 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154(2):931-942.

Nielsen, R. and J. Signorovitch. 2003. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theoretical Population Biology 63(3):245-255.

Norton, B. and E. S. Pearson. 1976. Note on the Background to, and Refereeing of, Fisher,R.A. 1918 Paper on the Correlation between Relatives on the Supposition of Mendelian Inheritance. Notes and Records. The Royal Society Journal of the History of Science 31(1):151-162.

Pearson, T. A. and T. A. Manolio. 2008. How to interpret a genome-wide association study. Jama 299(11):1335-1344.

Price, A. L., G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples, L. J. Wei, and S. R. Sunyaev. 2010. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. American Journal of Human Genetics 86(6):832-838.

Pryce, J. E. and H. D. Daetwyler. 2012. Designing dairy cattle breeding schemes under genomic selection: a review of international research. Animal Production Science 52(2-3):107-114.

Purfield D C, B. D. P., McParland S, et al. 2012. Runs of homozygosity and population history in cattle. BMC Genetics 13(1):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(5): e0176780.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832-837.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123(4):218-223.

Schork, N. J., S. S. Murray, K. A. Frazer, and E. J. Topol. 2009. Common vs. rare allele hypotheses for complex diseases. Current Opinion in Genetics & Development 19(3):212-219.

Smith, J. M. and J. Haigh. 2007. The hitch-hiking effect of a favourable gene. Genetics Research 89(5-6):391-403.

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

Swartz, M. D., T. Kim, J. Niu, R. K. Yu, S. Shete, and I. Ionita-Laza. 2014. Small sample properties of rare variant analysis methods. BMC proceedings 8(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo):S13.

Szpiech, Z. A., J. Xu, T. J. Pemberton, W. Peng, S. Zollner, N. A. Rosenberg, J. Z. Li. 2013. Long runs of homozygosity are enriched for deleterious variation. American Journal of Human Genetics 93(1):90-102.

van Binsbergen, R., M. C. A. M. Bink, M. P. L. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 46(1):41.

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. C. H. M. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, and G. A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of Dairy Science 96(1):668-678.

Veerkamp, R. F., A. C. Bouwman, C. Schrooten, and M. P. Calus. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. Genetics Selection Evolution 48(1):95.

Velankar, S., J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M. J. Martin, and G. J. Kleywegt. 2013. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research 41(D1):D483-D489.

Vilela, F. O., A. T. do Amaral, M. G. Pereira, C. A. Scapim, A. P. Viana, and S. D. P. Freitas. 2008. Effect of recurrent selection on the genetic variability of the UNB-2U popcorn population using RAPD markers. Acta Scientiarum. Agronomy 30(1):25-30.

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

Wright, S. 1922. Coefficients of inbreeding and relationship. The American Naturalist 56(645):330-338.

Wright, S. 1948. Genetics of Populations. Encyclopaedia Britannica 10:111-A-D-112.

Wu, M. C., S. Lee, T. X. Cai, Y. Li, M. Boehnke, and X. H. Lin. 2011. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. American Journal of Human Genetics 89(1):82-93.

Xu, L. Y., D. M. Bickhart, J. B. Cole, S. G. Schroeder, J. Z. Song, C. P. Van Tassell, T. S. Sonstegard, and G. E. Liu. 2015. Genomic Signatures Reveal New Evidences for Selection of Important Traits in Domestic Cattle. Molecular Biology Evolution 32(3):711-725.

Yang, J., A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. B. Perry, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Magi, A. Metspalu, A. Hamsten, P. K. E. Magnusson, N. L. Pedersen, E. Ingelsson, N. Soranzo, M. C. Keller, N. R. Wray, M. E. Goddard, P. M. Visscher, and L. C. Study. 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature Genetics 47(10):1114.

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42(7):565-U131.

Yang, J. A., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. American Journal of Human Genetics 88(1):76-82.

Zhang, Q. Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015a. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics 16(1):88.

Zhang, Q., B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. 2015b. Runs of homozygosity and distribution of functional variants in the cattle genome. BMC Genomics 16:542.

2

Runs of homozygosity and distribution of functional variants in the cattle genome

Qianqian Zhang[1,2], Bernt Guldbrandtsen[1], Mirte Bosse[2], Mogens S Lund[1] and Goutam Sahana[1]

[1] Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, Tjele DK-8830, Denmark; [2] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen 6700 AH, The Netherlands.

BMC Genomics (2015) 16:542

Abstract

Background: Recent developments in sequencing technology have facilitated widespread investigations of genomic variants, including continuous stretches of homozygous genomic regions. For cattle, a large proportion of these runs of homozygosity (ROH) are likely the result of inbreeding due to the accumulation of elite alleles from long-term selective breeding programs. In the present study, ROH were characterized in four cattle breeds with whole genome sequence data and the distribution of predicted functional variants was detected in ROH regions and across different ROH length classes.

Results: On average, 19.5 % of the genome was located in ROH across four cattle breeds. There were an average of 715.5 ROH per genome with an average size of ~750 kbp, ranging from 10 (minimum size considered) to 49,290 kbp. There was a significant correlation between shared short ROH regions and regions putatively under selection ($p < 0.001$). By investigating the relationship between ROH and the predicted deleterious and non-deleterious variants, we gained insight into the distribution of functional variation in inbred (ROH) regions. Predicted deleterious variants were more enriched in ROH regions than predicted non-deleterious variants, which is consistent with observations in the human genome. We also found that increased enrichment of deleterious variants was significantly higher in short (<100 kbp) and medium (0.1 to 3 Mbp) ROH regions compared with long (>3 Mbp) ROH regions ($P < 0.001$), which is different than what has been observed in the human genome.

Conclusions: This study illustrates the distribution of ROH and functional variants within ROH in cattle populations. These patterns are different from those in the human genome but consistent with the natural history of cattle populations, which is confirmed by the significant correlation between shared short ROH regions and regions putatively under selection. These findings contribute to understanding the effects of inbreeding and probably selection in shaping the distribution of functional variants in the cattle genome.

Key words: Runs of homozygosity, Polymorphisms, Inbreeding, Cattle, Genome sequencing

2.1 Background

Dairy cattle have been subjected to more than 60 years of intense selection for traits to enhance milk production (Christensen, 1989, Ramachandran et al., 2005, Hayes et al., 2009, Brade and Brade, 2013). Relatively few bulls were chosen to produce thousands of daughters, resulting in large half-sib families. Traditionally, cattle breeders estimate the inbreeding coefficient by the degree of parental relatedness using pedigree and genotype data (Wright, 1921, Nomura, 2001, Sorensen et al., 2005, Bjelland et al., 2013). The rate of inbreeding in cattle populations has increased in recent years (Miglior et al., 2005, Sorensen et al., 2005, Gonzalez-Recio et al., 2007), and there is a strong correlation between inbreeding levels and reduced fitness (Freyer et al., 2005). This can be explained by the increased risk of homozygosity for deleterious alleles as inbreeding increases (Ku et al., 2011). Thus, high levels of inbreeding in populations will result in inbreeding depression (Miglior et al., 1992, Gonzalez-Recio et al., 2007). Reduced variability also leads to a reduced selection response in breeding programs (Weigel, 2006). Thus, maintaining genetic diversity is crucial in cattle breeding populations.

The availability of high-throughput whole genome sequencing for substantial numbers of animals has opened new avenues in examining genetic diversity and led to reliable and detailed investigation of large chromosome segments, including stretches of homozygous genomic regions (Bosse et al., 2012). Runs of homozygosity (ROH) are contiguous homozygous stretches in an individual genome due to transmission of identical haplotypes from parents to offspring. ROH detection can be used to improve mating systems and minimize inbreeding. However, the effects of inbreeding vary among individuals and populations, and it has long been of interest to explore the mechanisms of inbreeding depression and deleterious variants at the genomic level (Margolin and Bartlett, 1945, Pusey and Wolf, 1996, Koenig and Simianer, 2006). Recently, ROH extracted from SNP chip data have been used to study the population history of different cattle breeds (Margolin and Bartlett, 1945). Purfield et al. (2012) claimed that both natural and artificial selection of cattle, as well as demographic processes, have resulted in breeds with extensive phenotype variation. ROH may provide useful information on how these processes work in disparate populations, especially for cattle due to recent intense selection of sires, artificial insemination, and embryo transfer in some cattle breeds (Purfield et al., 2012). However, ROH estimated from SNP chip data may miss short and medium ROH due to limited resolution. Additionally, SNP

arrays suffer from ascertainment biases due to inclusion of SNPs with high minor allele frequencies (Lencz et al., 2007).

Consanguineous mating, population size reduction, and selection result in long homozygous regions along the genome (Pusey and Wolf, 1996, Bosse et al., 2012). Two copies of a genome segment in an individual inherited from a common ancestor without recombination are identical-by-descent (IBD). ROH that arise due to inbreeding tend to be fairly long and are dispersed more or less randomly throughout the genome. Accumulation of deleterious variants by definition creates fitness consequences, particularly when homozygous (Lencz et al., 2007, Nalls, 2009). Distribution of functional homozygote variants provides information on enrichment of deleterious homozygotes within homozygous regions compared with neutral homozygotes. Charlesworth et al. (1993) suggested that a large number of weakly deleterious variants are purged by negative selection. A human genome study (Lohmueller et al., 2011) indicated that purifying selection interacted with founder effects during demographic processes, affecting the proportion of recessive deleterious variants. Recently, based on variant annotations, Szpiech et al. (2013) reported long ROH enriched deleterious variation in the human genome. By counting deleterious and neutral variants inside and outside ROH in humans, they determined that out of the two competing hypotheses regarding patterns of deleterious variation in the human genome — namely, whether a smaller or larger proportion of all deleterious homozygotes resided in ROH regions compared with neutral homozygotes (Szpiech et al., 2013), a larger proportion of deleterious homozygote variants were found, and deleterious variants in long ROH accumulated more rapidly compared to short and medium ROH in the human genome.

Cattle domestication was initiated approximately 10,000 years ago, with evidence of at least two separate domestication events (Loftus, 1994). During the last few decades, breeders in modern dairy cattle breeding programs have implemented strong artificial selection. Kim et al. showed that selection increases overall autozygosity across the genome, whereas the autozygosity in an unselected line does not change significantly across most chromosomes in cattle populations (Kim et al., 2013). Genomic regions under positive selection show increased ROH levels due to local reductions in haplotype diversity, i.e. selective sweeps (Leocard, 2009). Certain haplotypes under these conditions constitute a large proportion of the total haplotype pool for the portions of the genome targeted for selective sweeps. In these conditions, a portion of the ROH is the actual selection target and is retained

in the ROH region (Pemberton et al., 2012). These ROH will increase in frequency and be subject to purging, as the spread of haplotypes carrying deleterious recessive alleles are restrained by forces counteracting the selection pressure driving the selective sweep (Lohmueller et al., 2008). Therefore, we expect that this ROH type would be on average shorter, composed of more common haplotypes, be concentrated in fewer genomic locations, and contain fewer deleterious alleles than the genomic average.

Short ROH are generally due to older haplotype relatedness, while longer ROH result from more recent parental relatedness (Kirin et al., 2010). Thus, short and medium sized ROH have been subject to selection for a longer period of time than longer ROH. Furthermore, recombination will have had more time to trim down ROH that have been the target of selective sweeps (Carbone et al., 2007, Kim et al., 2013). ROH regions are expected to exhibit an increased frequency of homozygotes compared to non-ROH regions, as the homozygote allelic frequency in non-ROH regions is p^2 and the allelic frequency in ROH regions is p, given a population allelic frequency of p. Therefore, IBD results in enrichment of homozygotes in ROH. Given the nature of cattle breeding, artificial selection is expected to play a more crucial role in shaping the frequency and distribution of functional variants in ROH in modern cattle relative to human populations.

Therefore, we expect that selection pressures could reduce the number of regions with an increased frequency of deleterious variations, and, at the same time, enrich for short ROH regions with substantial beneficial effects in cattle breeding populations. This remains an ongoing process, as the properties of genomic variants within ROH regions continue to be discovered in cattle (Stothard et al., 2011, Zhan et al., 2011, Jansen et al., 2013). An increase in functional variant frequency in different ROH length classes should also be examined under different population-genetic processes, such as the number of generations of inbreeding (Kirin et al., 2010).

Genome scale bioinformatics annotations are available from a number of sources, including the Variants Effect Predictor from ENSEMBL (McLaren et al., 2010), with several available levels of annotation. Generally, synonymous polymorphisms are the least subject to selective forces as they have the least effect on the resultant protein. Tools such as SIFT can be used to predict non-synonymous change effects on proteins, and, thus, give an idea of their likely selective pressure (Velankar et al., 2013). SIFT classifies non-synonymous changes as either non-deleterious or

deleterious, based on the predicted effect on the protein. By comparing these classes (synonymous, non-deleterious non-synonymous, and deleterious non-synonymous), we can study how selection has shaped deleterious variants within the cattle genomic ROH.

We hypothesize that due to strong artificial selection pressures and demographic processes in cattle, deleterious variants increased in frequency within ROH compared with non-deleterious variants and that the deleterious allelic frequency and distribution in ROH classes differs markedly from the human genome. Testing this hypothesis will contribute to our understanding of inbreeding in cattle and help elucidate how artificial selection and other population level processes affect the distribution of functional variants. To accomplish this, we examined patterns of ROH detected using full genome sequencing in four Danish cattle breeds and studied the distribution and frequency of deleterious and non-deleterious variations in different length ROH regions.

2.2 Methods

As previously obtained cattle genomic sequences were used exclusively in this project and no live animal experiments were performed, no animal use and care committee approval was required.

2.2.1 SNP genotyping, sequencing, variant calling, and quality control

A total of 104 bulls, (i.e. 32 HOL, 27 JER, 15 old-RED, and 30 new-RED) with high genetic contributions to the current Danish dairy cattle populations, including Holstein (HOL), Jersey (JER), old Red Danish Dairy cattle (old-RED), and New Danish Red Dairy cattle (new-RED) were selected for sequencing. In addition, 81 and 85 individuals among those sequenced were respectively genotyped using High Density SNP assays (Infinium BovineHD BeadChip), and the 50 k assay Infinium BovineSNP50 v.1 BeadChip (Illumina, San Diego CA). SNP genotyping was performed as described by Höglund et al. (2014).

All selected individuals' genomes were sequenced to ~10× depth using Illumina paired-end sequencing. Sequencing was undertaken at the Beijing Genomics (Shenzhen, China), Aros Applied Biotechnology A/S (Aarhus, Denmark), and at Aarhus University (Foulum, Denmark). Reads were aligned to the cattle genome assembly UMD3.1 (Zimin et al., 2009) using bwa (Li and Durbin, 2009). Aligned sequences were converted to raw BAM files using samtools (Li et al., 2009).

Duplicate reads were removed using the samtools rmdup option (Li et al., 2009). The Genome Analysis Toolkit (McKenna et al., 2010) was used for local realignment around insertion/deletion (indels) regions, and recalibration following the Human 1000 Genome guidelines incorporating information from dbSNP (Sherry et al., 2001). Finally, variants were called using the Genome Analysis Toolkit (McKenna et al., 2010), which simultaneously calls short indels and SNPs by incorporating information from dbSNP (McKenna et al., 2010). Indels were excluded in further analyses, and variants with phred scores exceeding 100 were included in nucleotide diversity calculations and ROH computations. Nucleotide diversity was calculated for bins of 10 kbp over the entire genome in all 104 sequenced individuals following the procedures of Bosse et al. (2012) and Nei and Li (1979). SNP counts per 10 kbp bin were corrected for the number of bases within a 10 kbp bin, which is proportional to 10,000 covered bases. A correction factor must be applied, significant portions (0.5 – 2x) were not covered. The correction factor equaled DP/bin size, where DP is the coverage in bp/bin.

2.2.2 Principal component analysis

Genotypes were extracted from the sequence data sets (32 HOL, 27 JER, 15 old-RED, and 30 new-RED bulls) following variant calling using a perl script. Bi-allelic variant calls with phred scores exceeding 100 and higher than average read depths were used in genotype construction. Genome-wide Complex Trait Analysis (GCTA) (Yang et al., 2011) was employed to construct a genetic relationship matrix for chromosome 1 using all sequenced individuals. In addition, the population structure was examined for four breeds using a principal component method (Price et al., 2006) using GCTA (Yang et al., 2011).

2.2.3 Runs of homozygosity

The method developed by Bosse et al. (2012) was applied to identify ROH on all autosomes of the 104 sequenced individuals. The threshold to declare a ROH was set to a SNP count maximum of 0.25x the genome coverage following Bosse et al. (2012). ROH were also detected in 50 k and HD chip genotyped animals using the Runs of Homozygosity tool in PLINK (v. 1.07) (Purcell et al., 2007), with parameters similarly adapted to sequence data. Extracted ROH based on the technique of Bosse et al. (2012) were compared with ROH calculated from PLINK (v. 1.07) (Purcell et al., 2007) for the same individual and chromosome. ROH extracted from sequence data were further classified into three size categories: short ROH are smaller than 100 kbp, and reflect ancient homozygosity haplotypes; medium ROH exhibit sizes from 0.1 to 3 Mbp and arise from relatedness within populations; and

long ROH result from recently related individuals, with ROH sized larger than 3 Mbp (Kirin et al., 2010, Bosse et al., 2012).

The distribution of functional variants in the cattle genome was detected by computing the proportion of an individual's genome covered by any sized ROH region (j = ROH genome coverage; subscripts denote the ROH size). All comparison tests among breeds were two-tailed t-tests performed with the t.test function in R (v.3.1.0). Correlation coefficient significance tests were evaluated using the cor and cor.test function in R (v.3.1.0). The sharing of ROH between individuals among HOL, JER, Old-RED and New-RED was computed by counting the overlap ROH regions between individuals in the 10 kb bin over the full length of genome. To examine if the ROH distribution is the result of pure demography, we randomized the number and length of ROH regions for each individual over the genome and re-distributed them randomly throughout their genomes. These were then compared with the actual sharing of ROH regions between individuals as previously described.

2.2.4 Detection of selection signatures

2.2.4.1 Fst analysis

The genetic differentiation between individuals among HOL, JER, Old-RED and New-RED was measured by pairwise Fst analysis following Weir and Cockerham (W1984). The pairwise Fst between the defined breeds was computed with Genepop 4.2 in bins of 10 kb over the full length of the genome (Weir and Cockerham, 1984). Correlation between Fst and sharing of ROH averaged for the same bins of 500 kb was calculated with Pearson correlation in R (v.3.1.1).

2.2.4.2 Extended haplotype homozygosity tests

The extended haplotype homozygosity tests were implemented between the breeds for the sequenced individuals. The genome-wide scan for integrated haplotype score (iHS) within each breed was performed using the R package rehh (Sabeti et al., 2002, Gautier and Vitalis, 2012), and the four breeds were compared using the ies2rsb function in rehh (Tang et al., 2007, Gautier and Vitalis, 2012). Finally, the significance levels (the corresponding p values, assuming iHS or rSB are normally distributed under the neutral hypothesis) between breeds were averaged for a bin of 500 kb and were correlated with ROH sharing for the same bin of 500 kb by Person's correlation.

2.2.5 Variant annotations and classifications by predicted functional impacts

The called variants of each genomic site were annotated using ENSEMBL (v.67) databases with Variant Effect Predictor (VEP) (McLaren et al., 2010). Any sites with multiple transcripts resulting in multiple annotations were annotated only once using the by-gene option in VEP (McLaren et al., 2010). VEP determines variant effects (i.e. SNPs, insertions, deletions, CNVs, or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. It predicts genes and transcripts affected by variants, variant locations (e.g. upstream of a transcript, in a coding sequence, in non-coding RNA, in regulatory regions), and any variant consequence on protein sequence (e.g. gain or loss of a stop codon, missense, frameshift). SIFT scores were used to classify annotations for non-reference alleles. Given a set of mutations, SIFT predicts the potential effect a non-reference allele has on encoded proteins, and integrates effects of amino acid change, folded structure (predicted or known), and conservation score. SIFT categorizes the non-reference mutations as “deleterious” or “tolerated”. In this analysis, we classified non-reference alleles with a “deleterious” predicted effect as “damaging”, while “non-deleterious” or “tolerated” non-reference alleles were classified as “non-damaging”. It should be noted that non-reference alleles predicted as “deleterious” could just be different from reference alleles (in cattle, the reference genome was constructed from a beef breed rather than a dairy breed), which could exhibit substantial effects on amino acid change. Non-reference homozygotes were compared between non-deleterious (non-damaging), deleterious (damaging), and synonymous groups. Although truly damaging alleles could falsely be classified as non-damaging, the objective of this analysis was to detect the distribution of functional variants in regions of homozygosity at the whole genome level.

2.2.6 Distribution of functional variants in ROH regions

2.2.6.1 Number of deleterious homozygous genotypes in ROH

We followed the method proposed by Szpiech et al. (2013) to detect predicted functional variant distribution in ROH regions in these three cattle breeds (Zhang et al., 2015). We partitioned genotypes in our data into those occurring at deleterious versus non-deleterious sites and those occurring outside or inside ROH regions for the given ROH size. Homozygous non-reference genotypes (1/1) in all sequenced individuals were chosen. Alternate non-reference alleles were classified as deleterious or non-deleterious based on predicted effects as previously described (McLaren et al., 2010). Congruence with Szpiech et al. (2013) was maintained for individual i, across all sites, by denoting $g_i^{n,k}$ and $g_i^{d,k}$ the total number of sites with

k ∈ {0, 1, 2} alternate alleles at non-deleterious and deleterious sites, respectively. For an individual i, $g_{n,k}^{i,j}$ and $g_{d,k}^{i,j}$ represent the total number of sites with k ∈ {0, 1, 2} alternate alleles falling in ROH class j ∈ {S, M, L, R, N} at non-deleterious and deleterious sites, respectively (Szpiech et al., 2013). Here, S, M, and L indicate the different ROH length classes (S: small; M: medium; L: long), R is the union of all three ROH classes, and N represents sites located outside any ROH region (Szpiech et al., 2013). Therefore,

$$g_{i,R}^{n,k} = g_{i,S}^{n,k} + g_{i,M}^{n,k} + g_{i,L}^{n,k}$$
$$g_{i,R}^{d,k} = g_{i,S}^{d,k} + g_{i,M}^{d,k} + g_{i,L}^{d,k}$$
$$g_{i,N}^{n,k} = g_{i}^{n,k} - g_{i,R}^{n,k}$$
$$g_{i,N}^{d,k} = g_{i}^{d,k} - g_{i,R}^{d,k}$$

2.2.6.2 Deleterious and non-deleterious homozygotes in ROH of any size

Following Szpiech et al. (2013), we compared the proportion of deleterious non reference homozygotes inside and outside ROH regions to the corresponding proportion of non-deleterious non-reference homozygotes using the formula:

$$f_{i,R}^{n} = \frac{g_{i,R}^{n,2}}{g_{i}^{n,2}}$$

where $f_{i,R}^{n}$ is the proportion of non-deleterious 1/1 homozygotes in individual i that fall in any size ROH. These proportions of non-deleterious 1/1 homozygotes represent the distribution of non-deleterious homozygotes in ROH regions. Similarly, we computed

$$f_{i,R}^{d} = \frac{g_{i,R}^{d,2}}{g_{i}^{d,2}}$$

where $f_{i,R}^{d}$ is the proportion of deleterious 1/1 homozygotes in individual i that fall in any ROH region (Szpiech et al., 2013).

We performed two linear regressions on total genomic ROH coverage for deleterious and non-deleterious genotypes, and tested statistical significance of results following Szpiech et al. (2013). In addition, we fit a linear model

$$f_{i,R} = \beta_0 + \beta_1 G_{i,R} + \beta_2 D_i + \beta_3 G_{i,R} D_i + \varepsilon$$

where $f_{i,R}$ is a vector of length 104 containing, for all individuals, the proportion of genome-wide deleterious homozygotes in any ROH region ($f_{i,R}^{d}$) and the proportion of genome-wide non-deleterious homozygotes in any ROH region ($f_{i,R}^{n}$). $G_{i,R}$ is the proportion of the genome covered by ROH of any size for individual i, and Di is an indicator variable with a value of 1 if the observed response is of deleterious

homozygotes or a value of 0 for non-deleterious homozygotes (Szpiech et al., 2013). A statistically significant β_2 (via a two-tailed t-test) indicates a difference in the intercepts of separate regressions for deleterious and non-deleterious homozygotes, and a statistically significant β_3 (two-tailed t-test) indicates a difference in the regression slopes (Szpiech et al., 2013).

2.2.6.3 Deleterious and non-deleterious homozygotes by ROH size class

We subsequently tested how deleterious and nondeleterious homozygotes showed increased frequency in different size classes of ROH regions. It is interesting to explore which ROH lengths (L, M, or S) exhibited increases in deleterious or non-deleterious homozygotes in the cattle genome. Therefore, we separately evaluated each ROH size class following Szpiech et al. (2013). Similarly, for homozygous genotypes falling in ROH of size class j (j∈{S, M, L, R, N}), we calculated

$$f_{i,R}^{d} = \frac{g_{i,R}^{d,2}}{g_{i}^{d,2}}$$

$$f_{i,R}^{n} = \frac{g_{i,R}^{n,2}}{g_{i}^{n,2}}$$

for deleterious and non-deleterious 1/1 homozygotes, respectively. We investigated data points for each size class for each individual, using the $f_{i,R}^{d}$ and $f_{i,R}^{n}$ values.

We tested the statistical difference in these regressions with a linear model analogous to the equations from Szpiech et al. (2013). The regression model applied to distinguish deleterious homozygote distributions in ROH size classes is as follows:

$$f_{i}^{d} = \beta_0 + \beta_1 G_i + \beta_2 C_i + \beta_3 G_{i,R} D_i + \varepsilon$$

where f_{i}^{d} is a vector of length 104 containing, for all individuals, the proportions of genome-wide deleterious homozygotes in large ROH ($f_{i,L}^{d}$), medium ROH ($f_{i,M}^{d}$), and small ROH ($f_{i,S}^{d}$). G_i is the proportion of the genome covered by either large ($G_{i,L}$), medium ($G_{i,M}$), or small ($G_{i,S}$) ROH for individual i, and D_i is an indicator variable with a value of 1 if the observed response is deleterious homozygotes in large ROH or a value of 0 if the observed response is deleterious homozygotes in small ROH (when comparing large and small ROH). These comparisons were performed between all possible pairings of ROH sizes.

2.2.6.4 Nonsense variants and ROH

Our study classified homozygotes into two predicted classes: deleterious and non-deleterious. Although the deleterious class exhibited increased variants with deleterious effects, a more informative approach would be to examine a subset of variants with an even higher likelihood of being deleterious, e.g. nonsense mutations as suggested by Szpiech et al. (2013), as these are more likely to interfere with normal protein functioning. We tested two sets of predicted nonsense mutations in relationship to their distribution in different ROH lengths. The first set were predicted as stop gain and stop-loss mutations; the second was a mutation set predicted as frame shift and in-frame mutations, which were classified as loss of function variants. Following Szpiech et al. (2013), we divided individuals into two groups: "low-ROH" and "high-ROH" individuals to examine nonsense variants in ROH regions. Individuals with less than 20% genomic ROH coverage were classified as low ROH and those with more than 20% as high ROH.

2.3 Results

2.3.1 General Statistics

Runs of homozygosity (ROH) in the autosomes of 104 resequenced individuals were determined from four Danish dairy cattle breeds: Holstein (HOL), Jersey (JER), old Red Danish Dairy cattle (old-RED), and New Danish Red Dairy cattle (new-RED) (Fig. 2.1 and Additional file 1: Figure S1). The average genomic ROH content was 19.5 % across the four cattle breeds, with HOL, JER, New-RED, and Old-RED having 18.67 %, 24.23 %, 11.84 %, and 23.26 %, respectively. The average number of ROH per genome was 715.5 ± 21.0, with an average size of 750,564.2 bp, ranging from 10 kbp (the minimum size considered) to 49,290 kbp (Additional file 10: Table S1). The mean ROH size varied significantly between HOL, JER, Old-RED, and New-RED ($P < 0.001$) with the exception of JER and Old-RED (Fig. 2.1c). The mean number of ROH was significantly different between HOL, JER, Old-RED, and New-RED cattle ($P < 0.001$) with the exception of HOL and Old-RED (Fig. 2.1d).

The average genome-wide nucleotide diversity (π) was 1.59 (± 0.024) heterozygous positions per kbp across all individuals, including ROH, and 1.89 (± 0.018) heterozygous positions per kbp across all individuals in the genome excluding ROH (π-out) (Additional file 10: Table S1). The minimum and maximum nucleotide diversities were 1.50 (± 0.029) SNPs/kbp in JER and 1.71 (±0.024) SNPs/kbp in New-RED, respectively (Fig. 2.1a and Additional file 2: Figure S2, Additional file 10: Table S1) and were significantly different between all cattle breeds except HOL and New-RED ($P < 0.05$). π-out was significantly different between Old-RED and HOL, JER,

and New-RED (P < 0.001, two-tailed t-test) (Fig. 2.1a and b). Nucleotide diversity was higher in the vicinity of the major histocompatibility complex (MHC) on chromosome 23 (Additional file 1: Figure S1).

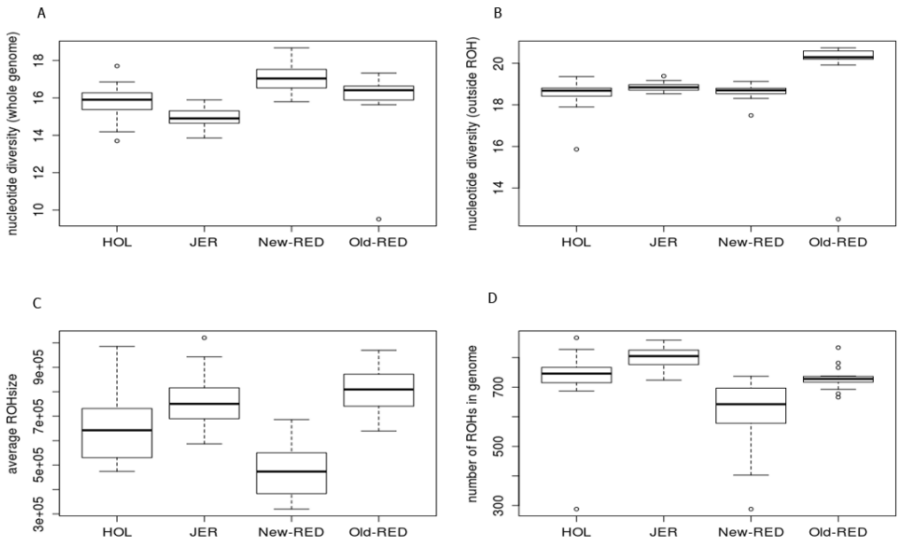


Figure 2.1 ROH general statistics. A. Average genome-wide nucleotide diversity (polymorphic sites per 10,000 bp). B. Average nucleotide diversity outside ROH (polymorphic sites per 10,000 bp). C. Average ROH size (bp). D. Average genome-wide ROH totals.

2.3.2 Genomic patterns of homozygosity

ROH were separated into three size classes: small (10 kbp to 100 kbp), medium (0.1 to 3 Mbp), and large (> 3 Mbp) (as described in the Materials and Methods section). The proportion of ROH in each size class was computed in all 104 sequenced individuals. While small ROH were frequent throughout the genome, they constituted a small proportion of the entire genome (Fig. 2.2). In contrast, medium ROH were much less frequent, they constituted significantly more of the genome than either small or large ROH. Large ROH were at least ten-fold less numerous than medium ROH, but nevertheless covered a sizable proportion of the total genome length. Old-RED cattle on average had the largest proportion of their genome in large ROH. New-RED cattle had fewer genomic ROH and a smaller genomic proportion of ROH than HOL and JER cattle (P < 0.001). Old-RED and JER cattle on average had more ROH and increased proportion of genomic homozygosity.

Principal component analysis successfully differentiated the four cattle breed individuals into different clusters based on SSR sequence data (Additional file 4: Figure S4), with JER the most distantly related based on PCA results. Three-dimensional plots did not show clear separation of cattle breeds based on to ROH size, ROH number, and π-out (Additional file 5: Figure S5). New-RED cattle represented the most variable cluster due to high nucleotide diversity and fewer ROH ($P < 0.001$) (Fig. 2.1). There was lower nucleotide diversity and more total ROH in JER compared to HOL and New-RED ($P < 0.001$) (Fig. 2.1). Despite its most distant origin, JER had lower nucleotide diversity (Additional file 4: Figure S4 and Additional file 5: Figure S5), creating several clusters in the three dimensional plot (Additional file 5: Figure S5). However, all Danish cattle breeds were more or less clustered together, with the exception of two New-RED individuals (Additional file 5: Figure S5).

Based on sequence data, New-RED cattle exhibited the fewest ROH and smallest ROH sizes. This is a composite breed with contributions from other red breeds, including Swedish Red, Finnish Ayrshire, and Brown Swiss. Compared with New-RED, these data suggest that the Old-RED breed has been more inbred based on relatively high coverage of long regions of homozygosity (> 3 Mbp) (Additional file 5: Figure S5), probably due to a relatively small breeding population and recent years of close mating.

Furthermore, the sharing of ROH regions was examined among sequenced individuals (Additional file 15: Figure S10B and Additional file 16: Figure S11C). Sharing of ROH regions primarily happened in short rather than long ROH regions, likely a result of combination of inbreeding and selection. Significant correlations were observed between Fst, iHS, and shared ROH regions in bins of 500 kb compared with the whole genome average ($P < 0.001$) (Additional file 15: Figure S10C and Additional file 16: Figure S11D). We also observed that instead of randomly distributed over the genomes (Additional file 17: Figure S12B), there were several obvious ROH-dense peaks distributed shared between individuals across genomes (Additional file 17: Figure S12A). Therefore, the distribution of ROH is not only result of pure demography, but likely the result of selection.

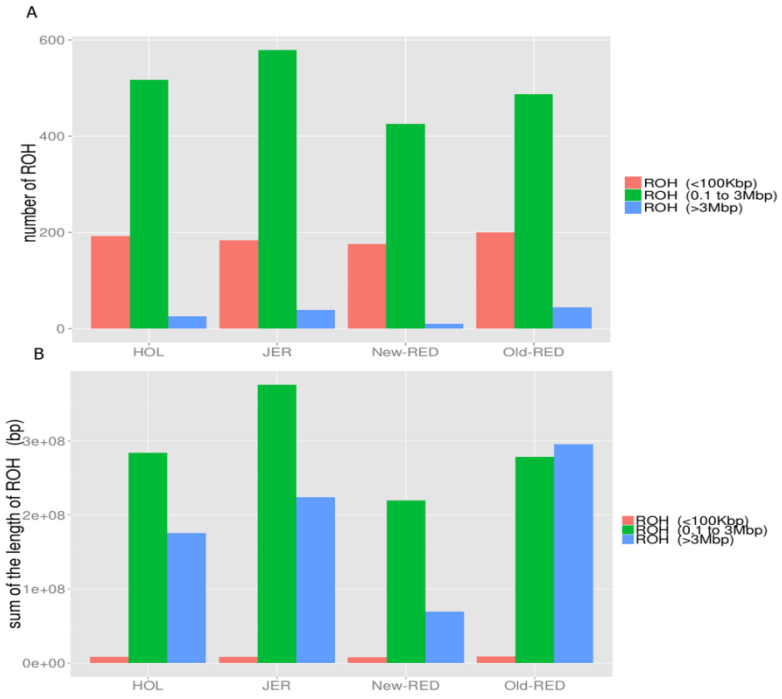


Figure 2.2 Total ROH number and genome proportions. A. The average small (< 100 kbp, Red), medium (0.1 to 3 Mbp, Green), and large (> 3 Mbp, Blue) ROH numbers for the four breeds. B. Average total genome ROH coverage for a given size class within each breed.

2.3.3 Distribution of functional variants in ROH regions

2.2.3.1 Number of deleterious homozygous genotypes in ROH

Additional file 11: Table S2 and Additional file 12: Table S3 show the counts for reference homozygotes (0/0), heterozygotes (0/1), and non-reference homozygotes (1/1) at deleterious and non-deleterious sites, respectively, that were contained within ROH and non-ROH regions (all $g_{d,k}^{i,j}$ and $g_{n,k}^{i,j}$). The number of deleterious non-reference homozygotes was consistent with the ROH coverage in all four breeds. Old-RED and JER showed increased ROH coverage, with a higher number of non-reference deleterious homozygotes in the genome. Non-reference non-deleterious homozygotes also exhibited the same trends as deleterious homozygotes.

Figure 2.3 shows the total number of deleterious nonreference homozygotes (1/1) as a function of the total proportion of the genome covered by ROH ($G_{i,R}$) for all

sequenced individuals. As ROH coverage increased (high $G_{i,R}$ values), a greater number of homozygotes were observed within ROH, which was consistent with findings from the human genome (Szpiech et al., 2013). There was a very strong positive correlation between the number of deleterious homozygotes and the genomic ROH proportion (Pearson r = 0.93, *slope* = 1568.76, *intercept* = −57.63). Similarly, the number of homozygotes outside of ROH decreased with the genomic ROH proportion due to smaller non-ROH regions as ROH coverage increased. As expected, there was a weak negative correlation between deleterious homozygotes outside ROH and the genomic ROH proportions (Pearson r = −0.12, *slope* = −98.67, *intercept* = 693.63). Compared with data from the human genome, the decreased slope for non-ROH regions was much shallower than the increased slope for ROH regions (Szpiech et al., 2013). However, this indicates that the increased deleterious homozygotes in ROH regions exceed deleterious homozygote declines in non-ROH regions in cattle. Similar to the human genome (Szpiech et al., 2013), the fitted lines also predict that, on average, individual non-inbred cattle ($G_{i,R} \approx 0$) carry approximately 694 deleterious homozygous variants. An increased in ROH coverage by 10 % will increase the expected deleterious homozygote numbers in ROH regions by 157 and decrease the expected number of deleterious homozygotes in non-ROH regions by 10, yielding an expected net increase of 147 deleterious homozygotes.

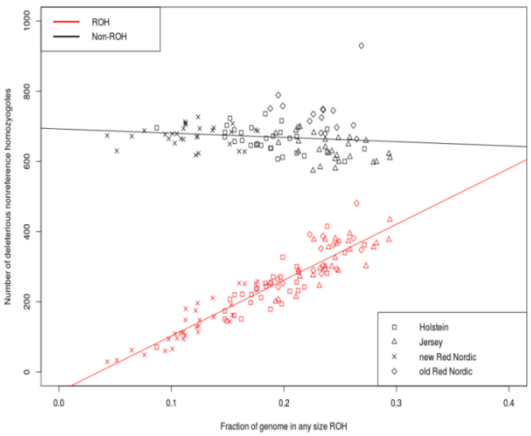


Figure 2.3 Deleterious non-reference homozygotes versus the genome ROH coverage in each individual. Red points represent the number of deleterious homozygotes falling within ROH regions and black points represent the number of deleterious homozygotes falling outside ROH regions.

2.3.4 Deleterious and non-deleterious homozygotes in ROH of any size

Figure 2.4a shows the proportion of deleterious nonreference homozygotes inside and outside ROH regions ($f_{i,R}^{d}$ and $f_{i,R}^{n}$) versus total genomic ROH coverage ($G_{i,R}$). The proportions of non-deleterious and deleterious homozygous genotypes within ROH were strongly positively correlated with total genomic ROH coverage (Pearson *r* = 0.96 for non-deleterious and *r* = 0.99 for deleterious). These high correlations were expected, because as larger proportions of homozygous genotypes occur, ROH coverage in the genome increases, and therefore ROH comprise an increasingly greater proportion of the genome (Szpiech et al., 2013). The $f_{i,R}^{d}$ proportion in genome-wide deleterious homozygotes within ROH consistently exceeded the $f_{i,R}^{n}$ proportion of genome-wide non-deleterious homozygotes within ROH and the increasing slopes differed between deleterious and non-deleterious variants.

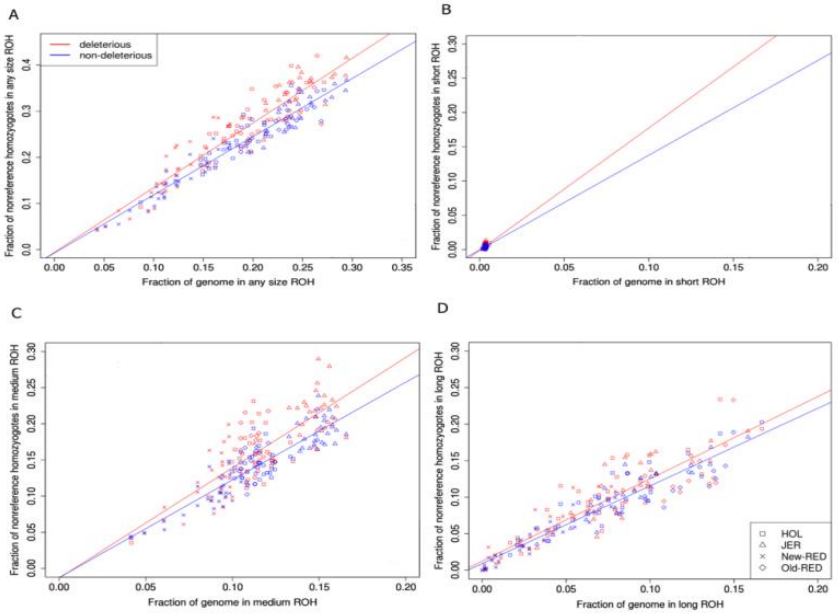


Figure. 2.4 The proportion of all genome-wide non-reference homozygotes falling in ROH regions versus the genome ROH coverage for each individual. A. Any ROH region. B. Short ROH regions. C. Medium regions. D. Long ROH regions. Red points represent deleterious homozygotes, and blue points represent non-deleterious homozygotes

After fitting the two linear regression models, we found β_3 was significant ($P < 0.05$) indicating that the interaction between the two regression slopes (deleterious and non-deleterious) was significant. However, β_2 was not significant ($P = 0.9174$) between the two regression intercepts. This is consistent with previous findings in humans (Szpiech et al., 2013), where deleterious homozygotes showed increased frequency in ROH relative to non-deleterious homozygotes and regression slopes were significantly different between deleterious and non-deleterious homozygotes.

2.3.5 Deleterious and non-deleterious homozygotes by ROH size

Figure 2.4b shows $f_{i,S}^{d}$ and $f_{i,S}^{n}$ versus total genomic coverage for small ROH ($G_{i,S}$) (Additional file 6: Figure S6). The proportion of non-deleterious homozygous genotypes in small ROH, and the proportion of deleterious homozygous genotypes in small ROH were positively correlated with genomic coverage (non-deleterious Pearson $r = 0.32$, deleterious Pearson $r = 0.44$). Figure 2.4c and d showed $f_{i,M}^{d}$ and $f_{i,M}^{n}$ versus total genomic coverage for medium ($G_{i,M}$) and large ($G_{i,L}$) ROH. The regressions for homozygote numbers in medium (non-deleterious $r = 0.88$, and deleterious $r = 0.80$) and large (non-deleterious $r = 0.94$ and deleterious $r = 0.90$) ROH had smaller P-values than in small ROH.

These results show that deleterious homozygotes occur more frequently in ROH than non-deleterious homozygotes. Additionally, when the proportion of deleterious homozygotes within large ROH ($f_{i,L}^{d}$) is compared to the proportion within small ROH ($f_{i,S}^{d}$), there was a substantially higher proportion of genome-wide deleterious homozygotes in small and medium vs. large ROH especially in individuals with moderate to high ROH coverage proportions (Fig. 2.5). Given that ROH coverage (Fig. 2.1) for all individuals across the four breeds differed (as previously mentioned). Therefore, statistical tests for each size group were robust across the breeds and there were different ROH coverage groups across all individuals (Additional file 10: Table S1). Similar trends were observed for each ROH size group, and significantly different degrees of enrichment were observed within each size group.

The intercepts and slopes of deleterious homozygotes and non-deleterious homozygotes were significantly different for large and medium ROH ($\beta2 = 0.02590$, $P < 0.05$; $\beta3 = -0.39931$, $P < 0.001$), but slopes and intercepts were not significantly different between small and medium ROH ($\beta2 = -0.0126778$, $P = 0.433$; $\beta3 = -0.2562209$, $P = 0.948$). These results indicate inbreeding that generates short and

medium ROH increases the proportions of deleterious and non-deleterious homozygotes in ROH regions compared to long ROH.

If deleterious, non-deleterious, and synonymous homozygotes are considered as three separate classes, patterns similar to those observed in the deleterious and non-deleterious homozygotes analysis emerge. Deleterious homozygotes were at highest proportions in short and medium length ROH. There were smaller proportions of synonymous and non-deleterious homozygotes as genomic ROH coverage of small and medium ROH increases (Additional file 7: Figure S7 and Additional file 8: Figure S8). In large ROH, the synonymous homozygote proportion over all homozygotes was higher than deleterious and non-deleterious homozygote proportions.

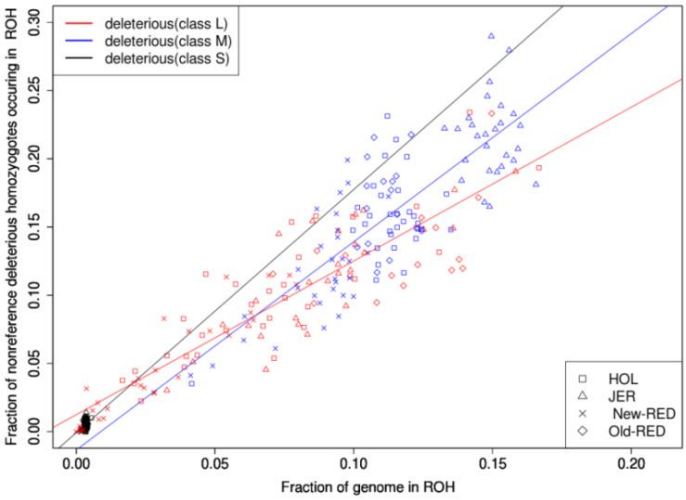


Figure. 2.5 The genome-wide proportion of all non-reference homozygotes falling in different ROH sizes versus genome ROH coverage for each individual. Red, orange, and black points represent deleterious homozygotes in large, medium, and small ROH regions, respectively.

2.3.6 Nonsense variants and ROH

Additional file 13: Table S4 and Additional file 14: Table S5 report nonsense and loss of function nonsense sites, respectively, with counts for reference homozygotes (0/0), heterozygotes (0/1), and non-reference homozygotes (1/1), which fell into ROH and non-ROH regions. Figure 2.6a shows nonsense mutation distribution across all ROH. For low-ROH individuals, the mean proportion of non-

deleterious homozygote variants falling in ROH marginally exceeded the nonsense or loss of function variants. For high-ROH individuals, however, the proportion of non-deleterious homozygotes within ROH was lower than for nonsense homozygotes. When ROH are segregated by size (Fig. 2.6b–d), including individuals with high genomic ROH coverage, the proportion of nonsense homozygotes in medium ROH was greater than that of non-deleterious homozygotes, while the proportion of nonsense homozygotes was slightly lower than non-deleterious homozygotes for large ROH. This is consistent with the finding that high-ROH individuals exhibited an increased proportion of damaging homozygotes (nonsense mutations) in ROH of any size (Fig. 2.6a), primarily driven by medium ROH (Fig. 2.6d).

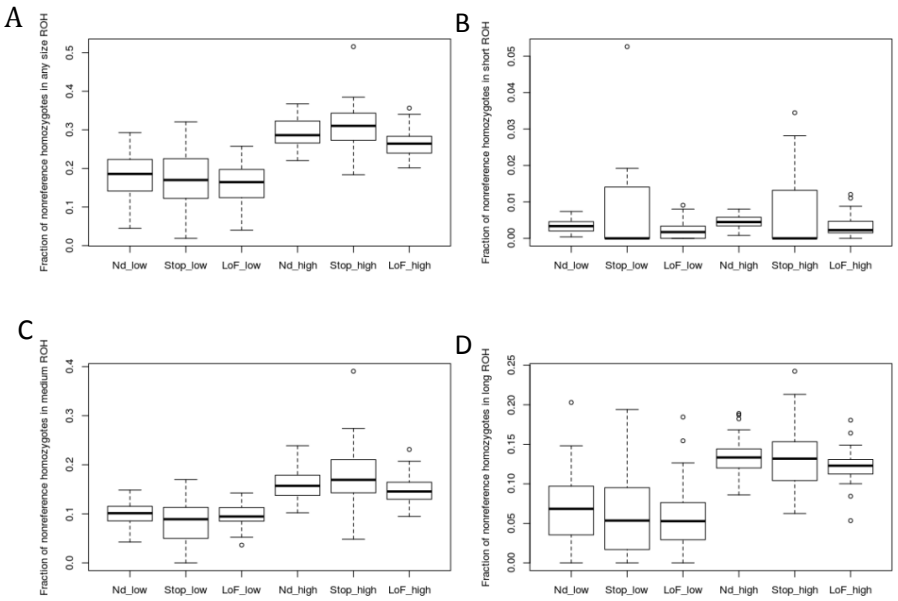


Figure. 2.6 The proportion of all genome-wide non-reference homozygotes falling in ROH regions for non-deleterious variants, nonsense variants, and loss of function nonsense variants versus the genome ROH coverage for individuals in the "low ROH" and "high ROH" groups. A. Any ROH region. B. Short ROH region. C. Medium ROH region. D. Long ROH region.

2.4 Discussion

2.4.1 Cattle genomic ROH patterns

The 104 individuals re-sequenced in the present study are key ancestors from the four Danish cattle breeds. The New-RED is a composite breed with its primary origins in the Old Danish Red, and includes contributions from other red breeds, including Swedish Red Finnish Ayrshire, and Brown Swiss cattle (Andersen et al., 2003). Our results showed ROH size ranged from tens of kb to several Mb and varied among individuals from different cattle breeds. Overall, medium sized ROH were most common. The average proportion of the genome represented by ROH was 19.5 % across all sequenced individuals. However, pedigree records and lower density SNP chips underestimated the inbreeding coefficient for these 104 bulls compared with that generated using ROH from genome sequencing. A previous study in Danish cattle also reported a less than 5% inbreeding coefficient using pedigree information (Sorensen et al., 2005). Among breeds, JER and Old-RED were relatively more inbred compared with HOL based on high ROH genome coverage. Meanwhile, a high number and proportion of long ROH were detected in Old-RED, likely indicating a small population with close mating for some period of time. However, the smallest proportion of ROH was detected in New-RED, presumably due to its outbred origins. These ROH patterns are consistent with the known population history for these four breeds (Andersen et al., 2003). Pemberton et al. (2012) allocated humans from different demographics to their corresponding place of origin by inferring population history from ROH analysis. In our study, we also determined that individuals with similar ROH distributions clustered together due to similar patterns of variation (Additional file 5: Figure S5). In contrast, PCA generated different cluster types in the same sets of individuals based on population structure (Additional file 4: Figure S4).

The fact that nucleotide diversity outside ROH was higher in Old-RED than HOL, JER, and New-RED indicates that the historic genetic diversity of Old-RED might be relatively higher than the other breeds examined due to larger breeding populations for each of several past generations. Furthermore, evidence suggested that Old-RED was the source population of New-RED. Although the newly derived RED populations exhibited the highest nucleotide diversity across the entire genome, the ancient haplotypes were more diverse in the ancestral Old-RED population. Reduced nucleotide diversity outside ROH in New-RED may also be explained by gene flow from HOL to New-RED. In addition, nucleotide diversity levels outside ROH presumably reflected the different origin of Old-RED from HOL and JER.

Bovine major histocompatibility complex (MHC) regions have long been known in the cattle genome (Andersson et al., 1988, Ellis, 1999). Our analysis also detected MHC regions, which have a high degree of nucleotide diversity, on chromosome 23 (Additional file 1: Figure S1). Pemberton et al. (2012) reported a correlation between linkage disequilibrium levels and ROH distribution in the human genome. MHC regions contain recombination hotspots (Andersson et al., 1988, Kauppi et al., 2003) and, therefore, high recombination rates to maintain relatively high levels of genetic diversity, but are subject to over dominant selection (Andersson et al., 1988). ROH are rarely present in MHC regions of the cattle genome, preventing random distribution of ROH (Additional file 1: Figure S1).

Purfield et al. (2012) showed ROH patterns in cattle populations using SNP chip data. Our analysis is the first to use next-generation sequencing data to infer ROH in cattle and indicate that long ROH can only be detected with 50 k or HD SNP chip data (minimum ROH size was 1 Mb), while short ROH are not detectable due to the required density of SNP chip data (Additional file 3: Figure S3). Short ROH regions shared between individuals in our data (Additional file 15: Figure S10b and Additional file 16: Fgure S11c) confirm that short ROH were selected and derived from ancient haplotypes that became fixed in populations (Additional file 9: Figure S9), while long ROH are the result of more recent inbreeding events. Consequently, SNP chip data misses information more relevant to historic inbreeding practices rather than recent inbreeding events. The significant correlation between the sharing of ROH regions and regions putatively under selection (from Fst analysis and iHS testing) (Additional file 15: Figure S10C and Additional file 16: Figure S11D) suggests that some of these short shared ROH are the result of a combination of inbreeding and selection. Instead of ROH regions randomly distributed over the genome, there were dense-ROH peak regions in the actual sharing of ROH regions among individuals (Additional file 17: Figure S12), which supports the hypothesis that the observed ROH patterns are not solely a result of demography (Additional file 17: Figure S12), as random ROH distributions would only be expected as a result of inbreeding.

2.4.1 Distribution of functional variants in the cattle genome

Years of intensive artificial selection in cattle breeding have resulted in reduced genetic diversity in cattle populations as demonstrated by the high proportion of ROH (11.8 – 24.2 %, average 19.5 %) found in this study. However, the results regarding ROH patterns suggest that the distribution and enrichment of putative functional variants in different ROH lengths is more interesting. Szpiech et al.

(2013) suggested that damaging variants are more enriched in human ROH, particularly longer ROH. The observed speed of deleterious homozygote accumulation in ROH far exceeds the accumulated decrease in deleterious homozygotes in non-ROH regions (Fig. 2.3). This is expected, since identity by decent causes homozygosity to increase in ROH regions compared to non-ROH regions. Deleterious variants were expected at a low frequency; therefore, rare occurrences of deleterious alleles were expected in the homozygous state. However, when a stretch of homologous DNA fragments are identical by descent, the probability of deleterious alleles increases (at a rate of p rather than p2, where p is allelic frequency). We also observed a higher allele frequency of deleterious variants in ROH regions compared to non-deleterious variants (Fig. 2.4). This was also expected as increased deleterious variants occur when allele frequencies are extreme, as has been observed in humans (Szpiech et al., 2013). In cattle, variants with 'deleterious' effects on protein structure may be artificially selected more frequently due to economic benefits. One example of this is a myostatin gene mutation in Belgian Blue cattle resulting in a "double muscling" phenotype. Meat from these cattle has a reduced fat content, as the mutation converts feed into increased lean muscle (Kambadur et al., 1997).

Cattle populations have been under strong artificial selection for many generations. The significant correlation between sharing of ROH regions and regions putatively under selection pressure (from Fst analysis and iHS testing) (Additional file 15: Figure S10 and Additional file 16: Figure S11) confirms that some of the shared short ROH regions have been selected and spread throughout the population. Moreover, by randomization of ROH regions over the genomes, Additional file 17: Figure S12B presented the patterns of ROH were only result of inbreeding. However, we do observe several dense-ROH peak regions shared among individuals in our populations, further supporting our belief that ROH patterns are not only the result of pure demography. Therefore, the distribution pattern and abundance of functional variants in different ROH lengths in cattle likely differs from the human population. Artificial selection purges deleterious alleles from regions that frequently occur in ROH, favoring alleles with strong beneficial effects. Specifically, the interaction between inbreeding and artificial selection for particular variants can have a strong effect on the distribution of functional variants. Moreover, potentially deleterious mutations might hitchhike with selected variants. Long-term artificial selection enriches cattle populations with beneficial alleles in short and medium ROH, along with hitchhiked deleterious variants, while variants in long ROH remain neutral.

The proportion of predicted deleterious homozygotes was greater in ROH regions than non-deleterious homozygotes. However, predicted deleterious homozygotes varied by the length of ROH. The rate of change differed between small, medium, and long ROH. Higher rates were observed in short and medium compared to long ROH. The slopes for deleterious and non-deleterious homozygotes were significantly different in short and medium ROH, and the patterns were similar for predicted nonsense (gain or loss of a stop codon) and loss of function (in-frame and frameshift) variants (Fig. 2.6). We also examined patterns of deleterious homozygote frequency in ROH using different length thresholds and saw the similar trends as reported using our original length thresholds (1.small ROH: length < 50 kbp; medium ROH: 50 kbp < =length < =2 Mbp; long ROH: length > 2 Mbp 2. small ROH: length < 150 kbp; medium ROH: 150 kbp < =length < =4 Mbp; long ROH: length > 4 Mbp).

Predicted deleterious variants may be detrimental, however, these allele frequencies cannot increase in long ROH, as inbred individuals harboring a large proportion of long ROH with a high frequency of deleterious alleles will have reduced fitness, leading to decreased survival. Alternatively, predicted deleterious variants may be harmful alleles that were carried into the genome with artificially selected beneficial alleles, and were therefore favored by selection over a number of generations and are reflected in short or medium ROH (Additional file 9: Figure S9, Additional file 15: Figure S10 and Additional file 16: Figure S11). Long ROH are evidence of recent shared ancestry, while short ROH typically reflect more ancient relatedness (Kirin et al., 2010). Long ROH regions have gone through selection for few generations, and will eventually break down into medium and then short ROH. Allelic combinations will likely be recombined within a few generations and disappear due to segregation (Bosse et al., 2012). In contrast, deleterious short or medium ROH variants, which reportedly hitchhike with beneficial alleles, are thought to persist for an extended periods of time and are shared among individuals via gene flow (Additional file 9: Figure S9, Additional file 15: Figure S10 and Additional file 16: Figure S11) (Bosse et al., 2012). Some of these shared short ROH regions were observed to be overlapping with regions under selection based on the Fst analysis and iHS testing (Additional file 15: Figure S10 and Additional file 16: Figure S11). One mechanism for this is when a beneficial mutation occurs in a population and then spreads to the entire population, forming a selective sweep. Artificial selection will favor short or medium ROH regions harboring beneficial mutations that will spread and eventually become fixed in the sampled populations

(Additional file 9: Figure S9, Additional file 15: Figure S10 and Additional file 16: Figure S11). Therefore, we deduce that some predicted deleterious homozygotes in short or medium ROH were deleterious alleles that hitchhiked with beneficial variants and were selected in the population. Alternatively, shorter ROH may represent the interplay between random inbreeding and artificial selection for particular variants. Therefore, some of these shared short ROH tended to be candidate regions for selection. However, homozygosity for certain short and medium ROH regions were not maintained due to the absence of selection for any specific alleles; therefore, variants with deleterious effects will be purged by artificial selection. It should also be noted that the confounding effect of inbreeding with selection in generating long stretch of haplotype homozygosity may influence the robustness of EHH-based tests.

Lohmueller et al. (2008) suggested human populations with decreased genetic diversity supported an excess of recessive deleterious variants, resulting from founder effects in ancient populations during speciation (Ramachandran et al., 2005), with inflation in the frequency of these rare variants in contemporary populations. We observed a higher proportion of deleterious than non-deleterious homozygotes in ROH (Fig. 2.4). Therefore, another possible explanation for these results is a history of population inbreeding and founder events (Additional file 9: Figure S9), with preservation of deleterious variants from ancient populations in contemporary cattle populations represented by our samples. However, artificial selection has been implemented in cattle populations for many years, and regions under selection pressure tend to overlap with short, shared ROH regions (Additional file 15: Figure S10 and Additional file 16: Figure S11). This suggests that these preserved ancient alleles may have carried deleterious alleles via hitchhiking.

Predicting how variation affects gene function has varying degrees of reliability. SIFT scores (Velankar et al., 2013) were used to estimate potential fitness consequences for nonreference alleles in our analysis. Certainty regarding predicted functional effects is based on changes in the primary amino acids and impacts on protein function and biological processes. However, here we examined functional variant distribution in ROH regions instead of exploring the effects of each deleterious variant on fitness. A general pattern was obtained by combining all deleterious or non-deleterious variants into one specific class to explore variant distribution in ROH regions. Furthermore, our observations were confirmed by nonsense and loss of function variant classification. Similar patterns were observed when grouping variants into nonsense and loss of function variants. It should be

noted we only emphasized non-reference homozygotes with substantial effects on the organism and did not determine the impacts of reference homozygotes. There is the potential for deleterious or selected reference homozygotes, and, therefore, these alleles should also be examined. However, reference alleles are not annotated with a SIFT score, preventing their examination in this study.

2.5 Conclusion

We characterized ROH using genome sequence data in four cattle breeds. The genome-wide proportion and distribution patterns of ROH differed among HOL, JER, New-RED, and Old-RED cattle breeds. We observed a significant correlation between the shared short ROH regions and regions putatively under selection. We also showed the distribution of functional variants in different ROH regions and an increased frequency in predicted deleterious homozygotes in short and medium, but not long, ROH, which differs from the human genome. However, the observed pattern and distribution of functional variants is consistent with the population history of the cattle studied, and we suspect that the observed distribution of functional variants is a result of combination of inbreeding and long-term artificial selection in cattle populations. This is supported by the significant correlation between shared short ROH regions and regions putatively under selection. Our findings contribute to the understanding of the effects of inbreeding and probably selection on shaping the distribution of functional variants in the cattle genome.

2.6 Appendix

Supplementary material can be found in the online version of the published paper or can be directly be accessed via
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-1715-x

Zhang, Q., B. Guldbrandtsen, M. Bosse, M. S. Lund, & G. Sahana. 2015. Runs of homozygosity and distribution of functional variants in the cattle genome. BMC genomics, 16(1):542.

2.7 Acknowledgements

Q. Zhang benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”. This research was

supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research.

References

Andersen, B., B. Jensen, A. Nielsen, L. G. Christensen, and T. Liboriussen. 2003. Rød Dansk Malkerace-avlsmæssigt of kulturhistorisk belyst. Danmarks Hordbrugs Forskning.

Andersson, L., A. Lunden, S. Sigurdardottir, C. J. Davies, and L. Rask. 1988. Linkage Relationships in the Bovine Mhc Region - High Recombination Frequency between Class-Ii Subregions. Immunogenetics 27(4):273-280.

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

Bosse, M., H. J. Megens, O. Madsen, Y. Paudel, L. A. Frantz, L. B. Schook, R. P. Crooijmans, and M. A. Groenen. 2012. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genetics 8(11):e1003100.

Brade, W. and E. Brade. 2013. Breeding History of German Holstein Cattle. Ber Landwirtsch 91(2).

Carbone, I., J. L. Jakobek, J. H. Ramirez-Prado, and B. W. Horn. 2007. Recombination, balancing selection and adaptive evolution in the aflatoxin gene cluster of Aspergillus parasiticus. Molecular Ecology 16(20):4401-4417.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics 134(4):1289-1303.

Christensen, L. G. 1989. Cattle-Breeding after 1992. Zuchtungskunde 61(6):428-439.

Ellis, S. A., & Ballingall, K. T. 1999. Cattle MHC: evolution in action? Immunological reviews 167(1):159-168.

Freyer, G., J. Hernandez-Sanchez, and B. G. Cassell. 2005. A note on inbreeding in dairy cattle breeding. Arch Tierzucht 48(2):130-137.

Gautier, M. and R. Vitalis. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28(8):1176-1177.

Gonzalez-Recio, O., E. L. de Maturana, and J. P. Gutierrez. 2007. Inbreeding depression on female fertility and calving ease in Spanish dairy cattle. Journal of Dairy Science 90(12):5744-5752.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges (vol 92, pg 433, 2009). Journal of Dairy Science 92(3):1313-1313.

Hoglund, J. K., G. Sahana, B. Guldbrandtsen, and M. S. Lund. 2014. Validation of associations for female fertility traits in Nordic Holstein, Nordic Red and Jersey dairy cattle. BMC Genetics 15(1):8.

Jansen, S., B. Aigner, H. Pausch, M. Wysocki, S. Eck, A. Benet-Pages, E. Graf, T. Wieland, T. M. Strom, T. Meitinger, and R. Fries. 2013. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics 14(1):466.

Kambadur, R., M. Sharma, T. P. L. Smith, and J. J. Bass. 1997. Mutations in myostatin (GDF8) in double-muscled Belgian blue and Piedmontese cattle. Genome Research 7(9):910-916.

Kauppi, L., A. Sajantila, and A. J. Jeffreys. 2003. Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. Human Molecular Genetics 12(1):33-40.

Kim, E. S., J. B. Cole, H. Huson, 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 US Holstein Cattle. PLoS One 8(11): e80813.

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(11): e13996.

Koenig, S. and H. Simianer. 2006. Approaches to the management of inbreeding and relationship in the German Holstein dairy cattle population. Livestock Science 103(1-2):40-53.

Ku, C. S., N. Naidoo, S. M. Teo, and Y. Pawitan. 2011. Regions of homozygosity and their impact on complex diseases and traits. Human Genetics 129(1):1-15.

Lencz, T., C. Lambert, P. DeRosse, K. E. Burdick, T. V. Morgan, J. M. Kane, R. Kucherlapati, and A. K. Malhotra. 2007. Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. Proceedings of the National Academy of Sciences of the United States of America 104(50):19942-19947.

Leocard, S. 2009. Selective Sweep and the Size of the Hitchhiking Set. Advances in Applied Probability 41(3):731-764.

Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and G. P. D. Proc. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-2079.

Loftus, R. T., D. E. MacHugh, D. G. Bradley, P. M. Sharp, & P. Cunningham. 1994. Evidence for two independent domestications of cattle. Proceedings of the National Academy of Sciences 91(7): 2757-2761.

Lohmueller, K. E., A. Albrechtsen, Y. R. Li, S. Y. Kim, T. Korneliussen, N. Vinckenbosch, G. Tian, E. Huerta-Sanchez, A. F. Feder, N. Grarup, T. Jorgensen, T. Jiang, D. R. Witte, A. Sandbaek, I. Hellmann, T. Lauritzen, T. Hansen, O. Pedersen, J. Wang, and R. Nielsen. 2011. Natural Selection Affects Multiple Aspects of Genetic Variation at Putatively Neutral Sites across the Human Genome. PLoS Genetics 7(10): e1002326.

Lohmueller, K. E., A. R. Indap, S. Schmidt, A. R. Boyko, R. D. Hernandez, M. J. Hubisz, J. J. Sninsky, T. J. White, S. R. Sunyaev, R. Nielsen, A. G. Clark, and C. D. Bustamante. 2008. Proportionally more deleterious genetic variation in European than in African populations. Nature 451(7181):994-U995.

Margolin, S. and J. W. Bartlett. 1945. The Influence of Inbreeding Upon the Weight and Size of Dairy Cattle. Journal of Animal Science 4(1):3-12.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297–1303.

McLaren, W., B. Pritchard, D. Rios, Y. A. Chen, P. Flicek, and F. Cunningham. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26(16):2069-2070.

Miglior, F., B. L. Muir, and B. J. Van Doormaal. 2005. Selection indices in Holstein cattle of various countries. Journal of Dairy Science 88(3):1255-1263.

Miglior, F., B. Szkotnicki, and E. B. Burnside. 1992. Analysis of Levels of Inbreeding and Inbreeding Depression in Jersey Cattle. Journal of Dairy Science 75(4):1112-1118.

Nalls, M. A., R. J. Guerreiro, J. Simon-Sanchez, J. T. Bras, B. J. Traynor, J. R. Gibbs, et al. 2009. Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. Neurogenetics, 10(3), 183-190.

Nei, M. and W. H. Li. 1979. Mathematical-Model for Studying Genetic-Variation in Terms of Restriction Endonucleases. Proceedings of the National Academy of Sciences of the United States of America 76(10):5269-5273.

Nomura, T., Honda, T., & Mukai, F. 2001. Inbreeding and effective population size of Japanese Black cattle. . Journal of Animal Science 79(2):366-370.

Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li. 2012. Genomic patterns of homozygosity in worldwide human populations. American journal of human genetics 91(2):275-292.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38(8):904-909.

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

Purfield D C, B. D. P., McParland S, et al. 2012. Runs of homozygosity and population history in cattle. BMC Genetics 13(1):70.

Pusey, A. and M. Wolf. 1996. Inbreeding avoidance in animals. Trends in Ecology & Evolution 11(5):201-206.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences of the United States of America 102(44):15942-15947.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832-837.

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29(1):308-311.

Sorensen, A. C., M. K. Sorensen, and P. Berg. 2005. Inbreeding in Danish dairy cattle breeds. Journal of Dairy Science 88(5):1865-1872.

Stothard, P., J. W. Choi, U. Basu, J. M. Sumner-Thomson, Y. Meng, X. P. Liao, and S. S. Moore. 2011. Whole genome resequencing of Black Angus and Holstein cattle for SNP and CNV discovery. BMC Genomics 12(1):599.

Szpiech, Z. A., J. Xu, T. J. Pemberton, W. Peng, S. Zollner, N. A. Rosenberg, J. Z. Li. 2013. Long runs of homozygosity are enriched for deleterious variation. American Journal of Human Genetics 93(1):90-102.

Tang, K., K. R. Thornton, and M. Stoneking. 2007. A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biology 5(7):1587-1602.

Velankar, S., J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M. J. Martin, and G. J. Kleywegt. 2013. SIFTS: Structure Integration

with Function, Taxonomy and Sequences resource. Nucleic Acids Research 41(D1):D483-D489.

Weigel, K. 2006. Controlling inbreeding in modern dairy breeding programs. Advances in dairy technology: Proceedings of the Western Canadian Dairy Seminar 18:263-274.

Weir, B. S. and C. C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population-Structure. Evolution 38(6):1358-1370.

Wright, S. 1921. Systems of mating. II. The effects of inbreeding on the genetic composition of a population. Genetics 6(2):124-143.

Yang, J. A., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. American Journal of Human Genetics 88(1):76-82.

Zhan, B., Fadista, J., Thomsen, B., Hedegaard, J., Panitz, F., & Bendixen, C. . 2011. Global assessment of genomic variation in cattle by genome resequencing and high-throughput genotyping. . BMC Genomics 12(1):557.

Zhang, Q. Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics 16(1):1.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10(4):R42.

3

Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds

Qianqian Zhang[1,2], Mario Calus[2], Bernt Guldbrandtsen[1], Mogens S Lund[1] and Goutam Sahana[1]

[1] 1Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, Tjele DK-8830, Denmark; [2] Animal Breeding and Genomics, Wageningen University & Research, Wageningen 6700 AH, the Netherlands.

BMC Genetics (2015) 16:88

Abstract

Background: Levels of inbreeding in cattle populations have increased in the past due to the use of a limited number of bulls for artificial insemination. High levels of inbreeding lead to reduced genetic diversity and inbreeding depression. Various estimators based on different sources, e.g., pedigree or genomic data, have been used to estimate inbreeding coefficients in cattle populations. However, the comparative advantage of using full sequence data to assess inbreeding is unknown. We used pedigree and genomic data at different densities from 50k to full sequence variants to compare how different methods performed for the estimation of inbreeding levels in three different cattle breeds.

Results: Five different estimates for inbreeding were calculated and compared in this study: pedigree based inbreeding coefficient (F_{PED}); run of homozygosity (ROH)-based inbreeding coefficients (F_{ROH}); genomic relationship matrix (GRM)-based inbreeding coefficients (F_{GRM}); inbreeding coefficients based on excess of homozygosity (F_{HOM}) and correlation of uniting gametes (F_{UNI}). Estimates using ROH provided the direct estimated levels of autozygosity in the current populations and are free effects of allele frequencies and incomplete pedigrees which may increase in inaccuracy in estimation of inbreeding. The highest correlations were observed between F_{ROH} estimated from the full sequence variants and the F_{ROH} estimated from 50k SNP (single nucleotide polymorphism) genotypes. The estimator based on the correlation between uniting gametes (F_{UNI}) using full genome sequences was also strongly correlated with F_{ROH} detected from sequence data.

Conclusions: Estimates based on ROH directly reflected levels of homozygosity and were not influenced by allele frequencies, unlike the three other estimates evaluated (F_{GRM}, F_{HOM} and F_{UNI}), which depended on estimated allele frequencies. F_{PED} suffered from limited pedigree depth. Marker density affects ROH estimation. Detecting ROH based on 50k chip data was observed to give estimates similar to ROH from sequence data. In the absence of full sequence data ROH based on 50k can be used to access homozygosity levels in individuals. However, genotypes denser than 50k are required to accurately detect short ROH that are most likely identical by descent (IBD).

Key words: Runs of homozygosity, Polymorphisms, Inbreeding, Cattle, Genome sequencing

3.1 Background

The definition of inbreeding coefficient (F) is the probability that two alleles in an individual are identical by descent (IBD) relative to a base population where all alleles are assumed unrelated (Wright, 1922). Rates of inbreeding have increased as intensive selection was applied to the populations (Margolin and Bartlett, 1945, Miglior et al., 1992, Smith et al., 1998, Nomura et al., 2001, Sorensen et al., 2005, Gonzalez-Recio et al., 2007). Increased levels of inbreeding result in increased probability that animals are homozygous for deleterious alleles (Gonzalez-Recio et al., 2007, Bjelland et al., 2013, Szpiech et al., 2013). Thus, inbred animals suffer from inbreeding depression with reduced fitness, and highly inbred animals may have considerably reduced lifespans (Wright, 1921, Charlesworth and Charlesworth, 1987, Pusey and Wolf, 1996, Smith et al., 1998, González-Recio, 2007, Leroy, 2014). Information on inbreeding is critical in the design of breeding program to control the increase in inbreeding levels and thereby controlling inbreeding depression in the progeny. Pedigree information has been used to calculate the estimated inbreeding coefficient as the expected probability that two alleles at a locus are IBD (Blackwell et al., 1995, Weigel, 2006, Mc Parland et al., 2007). For example, Meuwissen and Luo proposed a method to estimate inbreeding coefficients based on pedigree data of large populations (Meuwissen and Luo, 1992). However, incomplete pedigrees result in erroneous estimates and an underestimation of levels of inbreeding (Cassell et al., 2003). VanRaden (1992) proposed a method to take into account unknown ancestors when estimating inbreeding coefficients, increasing the accuracy of inbreeding level estimates in incomplete pedigrees.

With the availability of Single Nucleotide Polymorphism (SNP) array genotyping technologies, long stretches of homozygous genotypes, known as runs of homozygosity (ROH) can be identified. ROH are believed to reflect an estimate of autozygosity on genomic level and generally identify genomic regions which are IBD (Broman and Weber, 1999, McQuillan et al., 2008). Theoretically, it is expected that ROH can be accurately estimated from the full sequence data, because these estimates do not suffer from sampling such as may be expected when subsets of loci, for instance 50k SNPs, are used (Bosse et al., 2012, Ferencakovic, 2013, Marras et al., 2014). The inbreeding coefficient can be calculated as the proportion of genome covered by ROH and has been shown to be more informative than the inbreeding coefficient estimated from pedigree data or other estimators because

ROH strongly correlate with homozygous mutation load (Keller et al., 2012). ROH have commonly been used to infer population history and to examine the effect of deleterious homozygotes caused by inbreeding in human populations (Charlesworth et al., 1993, Li et al., 2008, McQuillan et al., 2008, Kirin et al., 2010, Ku et al., 2011). Long ROH reflect recent inbreeding, whereas short ROH reflect ancient inbreeding (Kirin et al., 2010). However, only a few studies have evaluated ROH in cattle populations. Ferencakovic et al. examined the effect of SNP density and genotyping errors when estimating autozygosity from high-throughput genomic data (Ferencakovic et al., 2013b). Estimates based on ROH also vary with different densities of genomic data. The minimum length of ROH that can be detected depends on SNP density (Ferencakovic et al., 2013a, Ferencakovic et al., 2013b). Recently, Purfield et al. detected ROH in a cattle population from SNP chip data to infer population history (Purfield et al., 2012). However, to estimate the "true" state of ROH, whole-genome sequences should be used rather than SNP chip data, but, to date, there are only few studies doing this in cattle (MacLeod et al., 2013). With the advent of next-generation sequencing technology, whole genome sequences have become available to examine the fine-scale genetic architecture of the cattle genome. It is now possible to investigate and compare how well different commonly used estimators of inbreeding level correlate with ROH estimated using next-generation sequence (NGS) data.

In recent years, widespread availability of genotype data enabled computation of inbreeding from the diagonals of genomic relationship matrices, i.e., the "GRM" method (F_{GRM}), as a by-product of genomic selection. Similarly, using the genotypes, the inbreeding coefficient can be computed based on excess of homozygosity following Wright (1948) (F_{HOM}) and based on correlation between uniting gametes following Wright (1922) (F_{UNI}). The objective of the present study was to compare different estimators for inbreeding coefficients calculated from pedigree, 50k SNP chip genotypes and full sequence data with estimates based on ROH, for three different dairy cattle breeds.

3.2 Methods

3.2.1 SNP genotyping and sequencing

A total of 89 bulls with a high genetic contribution to current Danish dairy cattle populations were selected for whole-genome resequencing. These included 32 Holstein (HOL), 27 Jersey (JER), and 30 Danish Red Cattle (RDC) bulls. RDC cattle are

a composite breed with contributions from different red breeds, including Swedish Red, Finnish Ayrshire, and Brown Swiss (Andersen et al., 2003). Only bi-allelic variants SNPs with a phred-scaled quality score (Ewing et al., 1998) higher than 100 were kept for analysis to ensure the quality of variants. Genotypes were extracted from whole-genome sequence (WGS) data using GATK (McKenna et al., 2010) and a perl script. The sequence variants with read depth lower than 7 or higher than 30 were filtered out. In addition, 85 of the sequenced animals were genotyped with the Illumina 50k SNP assay (BovineSNP50 BeadChip version 1 or 2, Illumina, San Diego, CA). SNP genotyping and quality control were as described by Hoglund et al. (2014). Among the whole genome sequenced animals, 4 animals were not genotyped with the 50k SNP chip. Their genotypes for the SNPs on the 50k chip were extracted from their whole-genome sequences. The quality of genotype calls from SNP chips is expected to be higher than that of whole-genome sequences; therefore, only sequence variants with a high quality score (phred score > 100) were included. The corresponding corrections for reverse strand calls in the sequence data were converted to Illumina calls by correcting locus calling from reverse strands in Illumina calls to maintain consistency of allele encoding between Illumina calls and sequence data. The concordance between the SNP chip and sequence data was ~97 %.

3.2.2 Estimation of inbreeding

3.2.2.1 Using pedigree records (F_{PED})

Inbreeding coefficients for the 89 bulls were estimated using pedigree records (F_{PED}). The average pedigree depth was ~8 generations ranging from 3 to 13. Average pedigree depth was 7, 8 and 9 for HOL, JER, and RDC, respectively. The method proposed by VanRaden (1992) was used to compute inbreeding coefficients, which replaces unknown inbreeding coefficients by average inbreeding coefficients in the same generations. Inbreeding coefficients were calculated using the following formula (Quaas, 1976):

$$\mathbf{A}_{ii} = \sum_{j=1}^{i} \mathbf{L}_{ij}^{2}\, \mathbf{D}_{jj}$$

where $\mathbf{A_{ii}}$ is the i^{th} diagonal element of the **A** matrix (pedigree relationship matrix), which is equal to the inbreeding coefficient of the i^{th} animal plus 1. **L** is a lower triangular matrix containing the fraction of the genes that animals derive from their ancestors, and **D** is a diagonal matrix containing the within family additive genetic variances of animals (Meuwissen and Luo, 1992). The computation for matrix

elements $\mathbf{L_{ij}}$ and $\mathbf{D_{jj}}$ follows the rule of computation of the $\mathbf{A}$ matrix (Meuwissen and Luo, 1992). The detailed decomposition for computing $\mathbf{A}_{ii}$ is explained by Meuwissen and Luo (Meuwissen and Luo, 1992). The analysis was conducted using Relax2 software (Strandén and Vuori, 2006).

3.2.2.2 Using genotypes (F_{ROH}, F_{GRM}, F_{HOM}, F_{UNI})

Sequence data ROH were detected from sequence data using all bi-allelic variants according to the method of Bosse et al. (Bosse et al., 2012). This method was used to compute ROH for sequence data instead of PLINK because not all short ROH can be detected using PLINK for sequence data (the sliding window size in PLINK is fixed; therefore, ROH shorter than a certain length cannot be detected). The measure of homozygosity based on ROH (F_{ROH}) from genomic data is defined as the total length of genome covered by ROH divided by the overall length of genome covered by SNPs or sequences as follows (McQuillan et al., 2008):

$$F_{ROH} = \frac{L_{ROH}}{L_{AUTO}}$$

where L_{ROH} is the sum of ROH lengths and L_{AUTO} is the total length of autosomes covered by reads. The inbreeding coefficient was calculated by extracting ROH from sequence data. Three ROH estimates based on lengths were calculated from sequence data. The ROH was calculated separately by summing the ROH in different length classes: 1) based on all ROH; 2) ROH >1 Mbp; 3) ROH >3 Mbp. In addition, three other estimates of inbreeding coefficients were calculated using sequence data (F_{GRM}, F_{HOM}, F_{UNI}). The F_{GRM} estimate was calculated following VanRaden (2008) based on the variance of the additive genotypes. F_{GRM} was derived from

$$F_{GRM} = \frac{[x_i - E(x_i)]^2}{h_i} - 1 = \frac{(x_i - 2\hat{p_i})^2}{h_i} - 1$$

where p_i is the observed fraction of the first allele at locus i, $h_i = 2p_i(1 - p_i)$ and x_i is the number of copies of the reference allele (i.e., the allele whose homozygous genotype was coded as "0") for the i^{th} SNP (Yang et al., 2011). This was equivalent to estimating an individual's relationship to itself (diagonal of the SNP-derived GRM). The F_{HOM} estimate was calculated based on the excess of homozygosity following Wright (1948) (Wright, 1948):

$$F_{HOM} = \frac{[O(\#HOM) - E(\#HOM)]^2}{1 - E(\#HOM)} = 1 - \frac{x_i(2 - x_i)}{h_i}$$

where $O(\#HOM)$ and $E(\#HOM)$ are the observed and expected numbers of homozygous genotypes in the sample, respectively (Yang et al., 2011). The F_{UNI}

estimate was calculated based on the correlation between uniting gametes following Wright (1922) (Wright, 1922):

$$F_{UNI} = \frac{x_i^2 - (1 + 2p_i)x_i + 2p_i^2}{h_i}$$

where h_i and x_i are the same as for F_{GRM} (Yang et al., 2011). The calculations for these three estimates F_{GRM}, F_{HOM} and F_{UNI} were computed using the option –ibc from GCTA software (Yang et al., 2011).

50k SNP chip ROH were detected from 50k SNP chip data using the software PLINK with adjusted parameters (–homozyg-density 1000, –homozyg-window-het 1, –homozyg-kb 10, –homozyg-window-snp 20) (Purcell et al., 2007, Bosse et al., 2012). These settings for PLINK to detect ROH in SNP data were chosen to make the detected ROH in SNP chip data and sequence data as similar as possible to enable comparisons of results when using different types of data. Genomic estimates of the inbreeding coefficient based on all ROH (F_{ROH}) were calculated using the same formula as was used for the sequence data. The other three types of estimates (F_{GRM}, F_{HOM}, F_{UNI}) were also calculated for genotypes extracted from 50k SNP chip data using the same methods as for sequence data.

Pearson's correlation coefficients were calculated between estimates of inbreeding coefficients from each of pedigree records, 50k SNP genotypes, and whole-genome sequence variants. All correlations between different inbreeding coefficient estimators were tested within breed to determine whether they were significantly different from 0 using the R (http://www.r-project.org/) cor and cor.test functions.

3.2.3 Impact of allele frequencies on estimators of inbreeding

As some estimators explicitly use allele frequencies to compute inbreeding coefficients, it is important to investigate how varying allele frequencies affect estimated inbreeding coefficients. Here, we investigated how the three different estimators change across the whole range of allele frequencies. For each genotype x_i (homozygous for the reference allele; heterozygous for the reference and non-reference allele; homozygous for the non-reference allele), the values can be written as a function of allele frequency p_i, as shown in Table 3.1.

Table 3.1 Formula for calculating three estimators (F_{GRM}, F_{HOM} and F_{UNI}) for each genotype. homozygous for reference allele; heterozygous for reference and non-reference allele; homozygous for non-reference allele.

	F_{GRM}	F_{HOM}	F_{UNI}
$x_i = 0$	$F_{GRM} = \frac{3p_i - 1}{1 - p_i}$	$F_{HOM} = 1$	$F_{UNI} = \frac{p_i}{1 - p_i}$
$x_i = 1$	$F_{GRM} = \frac{6{p_i}^2 - 6p_i + 1}{2p_i(1 - p_i)}$	$F_{HOM} = 1 - \frac{1}{2p_i(1 - p_i)}$	$F_{UNI} = -1$
$x_i = 2$	$F_{GRM} = \frac{2 - 3p_i}{p_i}$	$F_{HOM} = 1$	$F_{UNI} = \frac{1 - p_i}{p_i}$

x_i is the number of reference allele

3.3 Results

We used five different approaches (F_{PED}, F_{GRM}, F_{HOM}, F_{UNI}, F_{ROH}) to estimate inbreeding coefficients using information from three different sources: pedigree, whole genome sequence and 50k SNP chip genotype data. There were total 11 estimates of inbreeding coefficients for each animal (Table 3.2). The average inbreeding coefficients estimated using different approaches and different data sets are presented in Table 3.2. The F_{PED} and F_{ROH} estimated from 50k data for HOL and JER are significantly higher than for RDC ($p < 0.05$). For inbreeding coefficients estimated from sequence data, F_{ROH}, $F_{ROH>1Mb}$, $F_{ROH>3Mb}$, F_{HOM} and F_{UNI} differed significantly among breeds, being highest in JER and lowest in RDC. The mean F_{ROH} for 50k SNP chip data (0.066), and sequence data (0.19) are significantly higher than F_{PED} (0.016) ($p < 0.01$).

F_{ROH} estimated from sequence data is a direct and accurate estimate of the levels of homozygosity. It mostly reflects regions which were IBD on the genome; therefore, we limited our comparisons to comparing between F_{ROH} from sequence data with other estimates of F. High correlations were observed between F_{ROH} estimated from the 50k and sequence data with $F_{ROH>1Mb}$ and $F_{ROH>3Mb}$ from the sequence data for all three breeds (Tables 3.3, 3.4 and 3.5). The correlation between F_{ROH} estimated from 50k data and $F_{ROH>3Mb}$ was higher than F_{ROH} estimated from 50k data and $F_{ROH>1Mb}$ in JER and RDC (Tables 3.4 and 3.5). F_{ROH} was consistently positively correlated with F_{HOM} and F_{UNI}, when both were computed from either 50k or sequence data in all three breeds (Tables 3.3, 3.4 and 3.5). A high correlation was found between F_{ROH} and F_{UNI}, when both were computed from either 50k or sequence data in all three breeds (Tables 3.3, 3.4 and 3.5). However, for different breeds, F_{HOM} and F_{UNI} were correlated differently across different densities of

Table 3.2 Estimated mean (min-max) of pedigree-based inbreeding coefficient (F_{PED}), GRM-based inbreeding coefficient (F_{GRM}), inbreeding coefficients based on excess of homozygosity (F_{HOM}), inbreeding coefficients based on correlation between uniting gametes (F_{UNI}), ROH-based inbreeding coefficients (F_{ROH}).

F_{ROH} greater than 1 Mb, 3 Mb derived from sequence data were reported.

		Mean			Range		
Inbreeding coefficients		HOL	JER	RDC	HOL	JER	RDC
50k SNP chip data	F_{PED}	0.036^{A}	0.018^{B}	0.003^{C}	0-0.100	0-0.060	0-0.013
	F_{ROH}	0.066^{A}	0.070^{A}	0.038^{B}	0.011-0.160	0.015-0.140	0.006-0.088
	F_{GRM}	0.023^{A}	-0.062^{A}	0.345^{B}	-0.162-0.683	-0.365-0.351	-0.055-0.653
	F_{HOM}	-0.008^{A}	-0.001^{A}	-0.234^{B}	-0.420-0.185	-0.227-0.147	-0.403-(-0.021)
	F_{UNI}	0.013^{A}	-0.031^{B}	0.057^{C}	-0.076-0.274	-0.121-0.063	-0.048-0.177
Sequence data	F_{ROH}	0.187^{A}	0.242^{B}	0.118^{C}	0.087-0.271	0.193-0.294	0.043-0.177
	$F_{ROH> 1Mb}$	0.113^{A}	0.162^{B}	0.055^{C}	0.060-0.205	0.104-0.225	0.009-0.110
	$F_{ROH> 3Mb}$	0.070^{A}	0.089^{B}	0.027^{C}	0.017-0.167	0.033-0.158	0-0.079
	F_{GRM}	-0.108^{A}	-0.122^{A}	0.014^{B}	-0.189-0.031	-0.179-(-0.031)	-0.244-0.34
	F_{HOM}	0.069^{A}	0.145^{B}	-0.123^{C}	-0.082-0.208	0.053-0.231	-0.408-0.061
	F_{UNI}	0.028^{A}	0.059^{B}	-0.007^{C}	-0.031-0.087	0.024-0.108	-0.054-0.055

HOL Holstein, JER Jersey, RDC Danish Red cattle. Significantly different means within each breed are indicated by a different superscript letter, *P*-values < 0.05

Table 3.3 Correlation coefficients between different estimates for inbreeding from different data sets for HOL.

Correlation		F_{PED}	50k SNP chip data				Sequence data					
			F_{ROH}	F_{GRM}	F_{HOM}	F_{UNI}	F_{ROH}	$F_{ROH> 1Mb}$	$F_{ROH> 3Mb}$	F_{GRM}	F_{HOM}	F_{UNI}
F_{PED}		1	0.82**	-0.20	0.58**	0.20	0.73**	0.83**	0.84**	-0.26	0.78**	0.68**
50k SNP chip data	F_{ROH}		1	-0.23	0.61**	0.15	0.87**	0.96**	0.96**	0.03	0.70**	0.88**
	F_{GRM}			1	-0.83**	0.87**	-0.10	-0.13	-0.16	0.36*	-0.31	-0.0005
	F_{HOM}				1	-0.44*	0.50**	0.58**	0.67**	-0.38*	0.66**	0.41*
	F_{UNI}					1	0.27	0.29	0.27	0.21	0.11	0.35
Sequence data	F_{ROH}						1	0.96**	0.91**	0.09	0.71**	0.95**
	$F_{ROH> 1Mb}$							1	0.98**	0.01	0.77**	0.94**
	$F_{ROH> 3Mb}$								1	-0.32	0.77**	0.90**
	F_{GRM}									1	-0.61**	0.29
	F_{HOM}										1	0.58**
	F_{UNI}											1

*: significantly different from 0 at $p<0.05$; **: significantly different from 0 at $p<0.01$. F_{PED} is the inbreeding coefficient estimated from pedigree data. F_{ROH} is inbreeding coefficient estimated based on ROH for 50k data and for sequence data F_{ROH}> 1Mb and $F_{ROH> 3Mb}$ are also reported. F_{GRM} is GRM-based inbreeding coefficient estimated from 50k and sequence data. F_{HOM} is inbreeding coefficient estimated based on excess of homozygosity for 50k and sequence data. F_{UNI} is the inbreeding coefficient estimated based on correlation of uniting gametes for 50k and sequence data.

Table 3.4 Correlation coefficients between different estimates for inbreeding from different data sets for JER.

Correlation		F_{PED}	50k SNP chip data				Sequence data					
			F_{ROH}	F_{GRM}	F_{HOM}	F_{UNI}	F_{ROH}	$F_{ROH> 1Mb}$	$F_{ROH> 3Mb}$	F_{GRM}	F_{HOM}	F_{UNI}
F_{PED}		1	0.47*	-0.18	0.46*	0.25	0.46*	0.52*	0.53*	-0.21	0.60**	0.43*
50k SNP chip data	F_{ROH}		1	0.36	0.06	0.79**	0.92**	0.93**	0.96**	0.29	0.67**	0.96**
	F_{GRM}			1	-0.89**	0.80**	0.19	0.16	0.22	0.86**	-0.34	0.44*
	F_{HOM}				1	-0.43*	0.24	0.28	0.21	-0.76**	0.66**	-0.01
	F_{UNI}					1	0.67**	0.67**	0.71**	0.69**	0.20	0.84**
Sequence data	F_{ROH}						1	0.99**	0.96**	0.14	0.76**	0.92**
	$F_{ROH> 1Mb}$							1	0.97**	0.094	0.80**	0.91**
	$F_{ROH> 3Mb}$								1	0.20	0.74**	0.95**
	F_{GRM}									1	-0.48*	0.42*
	F_{HOM}										1	0.60**
	F_{UNI}											1

*: significantly different from 0 at $p<0.05$; **: significantly different from 0 at $p<0.01$. F_{PED} is the inbreeding coefficient estimated from pedigree data. F_{ROH} is inbreeding coefficient estimated based on ROH for 50k data and for sequence data $F_{ROH> 1Mb}$ and $F_{ROH> 3Mb}$ are also reported. F_{GRM} is GRM-based inbreeding coefficient estimated from 50k and sequence data. F_{HOM} is inbreeding coefficient estimated based on excess of homozygosity for 50k and sequence data. F_{UNI} is the inbreeding coefficient estimated based on correlation of uniting gametes for 50k and sequence data.

Table 3.5 Correlation coefficients between different estimates for inbreeding from different data sets for RDC.

Correlation		F_{PED}	50k SNP chip data				Sequence data					
			F_{ROH}	F_{GRM}	F_{HOM}	F_{UNI}	F_{ROH}	$F_{ROH>1Mb}$	$F_{ROH>3Mb}$	F_{GRM}	F_{HOM}	F_{UNI}
F_{PED}		1	0.54**	0.36*	-0.31	0.45*	0.49**	0.54**	0.51**	-0.21	0.37*	0.32
50k SNP chip data	F_{ROH}		1	0.41*	0.35	0.80**	0.85**	0.96**	0.98**	0.08	0.21	0.77**
	F_{GRM}			1	-0.66**	0.82**	0.22	0.34	0.38*	-0.36	0.43*	0.05
	F_{HOM}				1	-0.10	0.40*	0.40*	0.38*	0.38*	-0.23	0.52
	F_{UNI}					1	0.60**	0.76**	0.79**	-0.20	0.40*	0.46*
Sequence data	F_{ROH}						1	0.93**	0.87**	0.003	0.31	0.81**
	$F_{ROH>1Mb}$							1	0.97**	0.010	0.29	0.79**
	$F_{ROH>3Mb}$								1	0.038	0.25	0.76**
	F_{GRM}									1	-0.95**	0.54**
	F_{HOM}										1	-0.24
	F_{UNI}											1

*: significantly different from 0 at $p<0.05$; **: significantly different from 0 at $p<0.01$. F_{PED} is the inbreeding coefficient estimated from pedigree data. F_{ROH} is inbreeding coefficient estimated based on ROH for 50k data and for sequence data $F_{ROH>1Mb}$ and $F_{ROH>3Mb}$ are also reported. F_{GRM} is GRM-based inbreeding coefficient estimated from 50k and sequence data. FHOM is inbreeding coefficient estimated based on excess of homozygosity for 50k and sequence data. F_{UNI} is the inbreeding coefficient estimated based on correlation of uniting gametes for 50k and sequence data.

genomic data. For HOL and RDC, the higher the density of genomic data used for F_{UNI}, the higher the correlation was between F_{UNI} and F_{ROH} from sequence data (Tables 3.3 and 3.5). For HOL, the correlation between F_{UNI} and F_{ROH} from sequence data (0.95) was still higher than the correlation between F_{ROH} estimated from 50k SNP chip data and sequence data (0.87) (Table 3.3). In contrast to JER, F_{HOM} and F_{UNI} were most highly correlated with F_{ROH} estimated from sequence data (Table 3.5).

F_{PED} was mostly intermediately correlated with F_{HOM} and F_{ROH} estimated from 50k and sequence data. The highest correlation between F_{PED} and F_{ROH} estimated from 50k and sequence data was found in HOL (Table 3.3). The strongest correlation among estimators of F_{ROH} (F_{ROH} from 50k or sequence data or $F_{ROH>3Mb}$ or $F_{ROH>1Mb}$ from sequence data) and F_{PED} was observed between F_{PED} and $F_{ROH>3Mb}$ from sequence data in HOL (Table 3.3). A moderate correlation was found between F_{PED} and F_{ROH} estimated from 50k and sequence data for JER and RDC (Tables 3.4 and 3.5).

The estimate F_{GRM} from both 50k and sequence data and F_{PED} had a correlation close to zero in all three breeds and the values were often negative (Tables 3.3, 3.4 and 3.5). At the same time, F_{GRM} estimated from 50k and sequence data generally showed a low correlation with other estimates except between two estimates F_{GRM} estimated from 50k and sequence data in HOL and JER, and between F_{GRM} and F_{UNI} estimated from 50k data (Tables 3.3 and 3.4).

3.4 Discussion

Pedigree has been used to estimate inbreeding coefficients in animal breeding for over 50 years (Wright, 1922, Meuwissen and Luo, 1992). Recently, researchers have utilized runs of homozygosity (ROH) estimated from medium density genotype data such as 50k SNP chip data to estimate inbreeding coefficients in livestock populations (Bosse et al., 2012, Ferencakovic et al., 2013a, Ferencakovic et al., 2013b, Marras et al., 2014). ROH were initially used to explore regions of inbreeding in the genome and further investigate the fitness effect of these regions on different traits (Charlesworth and Charlesworth, 1987, Gonzalez-Recio et al., 2007, Bjelland et al., 2013, Pryce et al., 2014). Population subdivision and either inbreeding or inbreeding avoidance affects the whole genome composition, whereas selection and assortative mating will affect only those loci associated with particular phenotypes. However, we observed that inbreeding coefficient F_{ROH}

estimated from sequence data were relatively higher for chromosome 1 and 10 for all four breeds (Fig. 3.1). This is most likely because the local recombination rate is relatively lower than average, which results in high levels of homozygosity on average (Arias et al., 2009, Bosse et al., 2012).

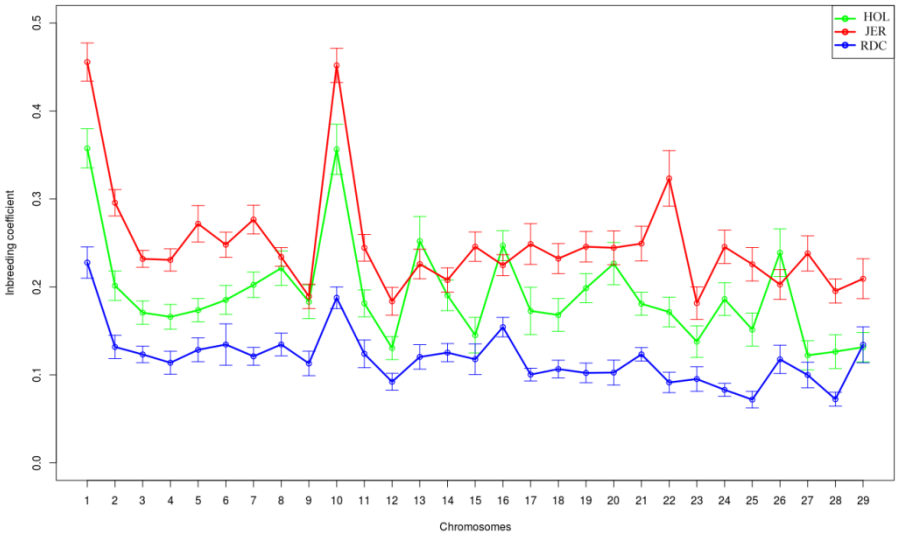


Figure 3.1 Distribution of inbreeding coefficients F_{ROH} estimated from sequence data using ROH for each chromosome in three breeds. Inbreeding coefficients F_{ROH} estimated from sequence data versus chromosomes 1–29 in HOL, JER and RDC. Standard error bars were computed among individuals within HOL, JER and RDC.

Our study is the first to calculate inbreeding coefficient based on ROH from full sequence data in cattle. The objective of this study was to compare estimates of inbreeding calculated from different methods and different data sources (pedigree, 50k SNP chip genotypes and full sequence data).

The pedigree-based inbreeding coefficient, F_{PED}, was moderately correlated with F_{HOM} and F_{ROH} in all breeds. These moderate correlations (~0.47 to 0.56) may be partly explained by the relatively shallow depth of the pedigree records (~8–9) for these bulls. Another difference between F_{ROH} and F_{PED} is that short ROH capture ancient inbreeding while long ROH capture recent inbreeding whereas pedigree captures only relatively recent inbreeding. Pedigree accounts only for inbreeding that occurred since pedigree recording began. Therefore, after excluding ROH smaller than 1 or 3 Mbp, the correlation between F_{PED} and F_{ROH} from sequence data

increased slightly for all breeds. We should also point out that a very long stretch of homozygosity using marker data might not actually be completely homozygous and therefore, higher density data was suggested to be used to detect selective sweeps through runs of homozygosity (Ramey et al., 2013). Sorensen et al. (2005) has estimated inbreeding in Danish Dairy Cattle Breeds and our estimates F_{PED} are lower than theirs. This is because our sampled animals for sequencing are founder and older animals compare to the other study where they used all animals (Sorensen et al., 2005).

Estimates of inbreeding coefficients differed with methods. Inbreeding coefficients estimates from methods using allele frequencies, i.e., F_{GRM}, F_{HOM} and F_{UNI}, showed considerable variation across data type and breeds. These estimators were sensitive to allele frequencies compared to ROH estimators, especially for populations with divergent allele frequencies (e.g., Fig. 3.2; RDC population). The estimates of genomic inbreeding coefficients are dependent on the allele frequencies in the base population (VanRaden, 2008).

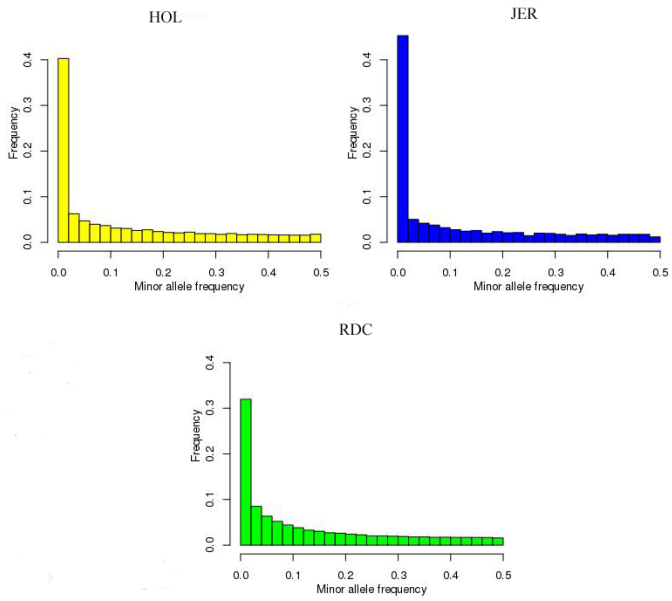


Figure 3.2 Minor allele frequency distribution for HOL, JER, and RDC bulls from sequence data. Minor allele frequency in HOL (yellow), JER (blue), and RDC (green) bulls against the minor allele frequency among all loci.

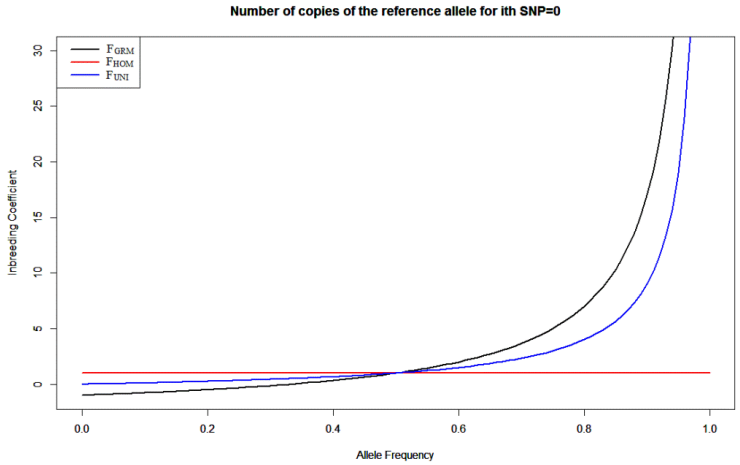


Figure 3.3 Inbreeding coefficients F_{GRM}, F_{HOM} and F_{UNI} against the reference allele frequency changing from 0 to 1 when the number of copies of reference alleles for the i^{th} SNP is 0. Black line represents F_{GRM}; red line represents F_{HOM} and blue line represents F_{UNI}.

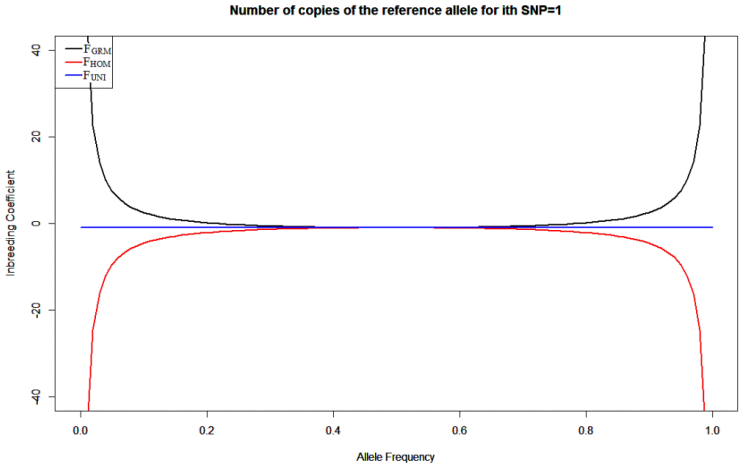


Figure 3.4 Inbreeding coefficients F_{GRM}, F_{HOM} and F_{UNI} against the reference allele frequency changing from 0 to 1 when the number of copies of reference alleles for the i^{th} SNP is 1. Black line represents F_{GRM}; red line represents F_{HOM} and blue line represents F_{UNI}.

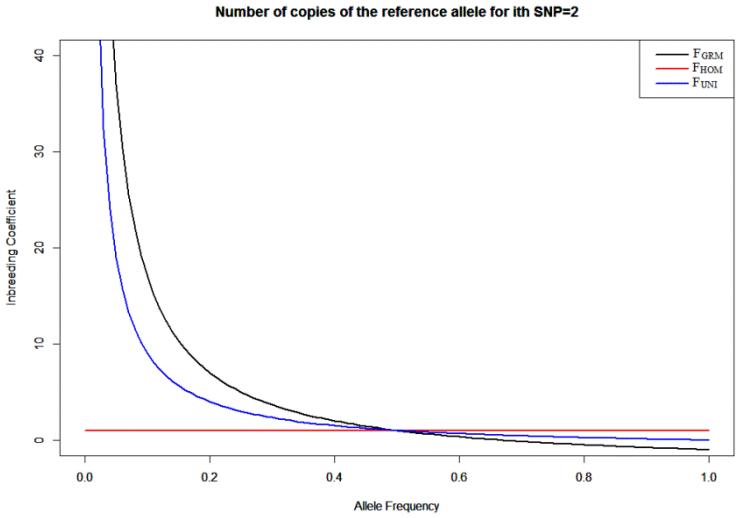


Figure 3.5 Inbreeding coefficients F_{GRM}, F_{HOM} and F_{UNI} against the reference allele frequency changing from 0 to 1 when the number of copies of reference alleles for the i^{th} SNP is 2. Black line represents F_{GRM}; red line represents FHOM and blue line represents F_{UNI}.

In order to explore the reasons about the various correlations between inbreeding coefficients estimates using allele frequencies, F_{GRM}, F_{HOM} and F_{UNI} were plotted against the allele frequency changing from 0 to 1 when the number of copies of reference alleles for i^{th} SNP is 0, 1 or 2 (Figs. 3.3, 3.4 and 3.5). When a locus is homozygous for either the reference alleles or the non-reference alleles with the allele frequency ranging from 0 to 1, F_{GRM} ranged from -1 to infinity, F_{HOM} has a constant value of 1 and F_{UNI} ranged from 0 to infinity (Figs. 3.3 and 3.5). F_{HOM} gave constant estimates for homozygous genotypes, regardless of the allele frequency (Figs. 3.3 and 3.5). When the allele frequency of the non-reference alleles is smaller than 0.2 or larger than 0.8, F_{GRM} was less than 0 (Figs. 3.3 and 3.5). When the allele frequency of the non-reference allele was between 0.2 and 0.5 or when the allele frequency of the reference allele was between 0.5 and 0.8, F_{GRM} become positive and ranges from 0 to 1 (Figs. 3.3 and 3.5).

For a heterozygous locus with an allele frequency ranging from 0 to 1, F_{GRM} and F_{HOM} ranged from minus infinity to plus infinity, and F_{UNI} has a constant value of 0 (Fig. 3.4). If the allele frequency was smaller than 0.2 or larger than 0.8 F_{GRM} become very large positive whereas F_{HOM} become a large negative. F_{HOM} was always negative, and F_{GRM} was always positive (Fig. 3.4). Thus, when a population has a high level of heterozygosity and some rare alleles with small frequency, F_{GRM} would

yield large positive inbreeding coefficients, which can be misleading. This result explains why F_{GRM} was positive in the RDC breed (Table 3.2): this population had a higher level of heterozygosity than HOL and JER. F_{UNI} gave a stable value of 0 when the locus was heterozygous and therefore was robust to allele frequency (Fig. 3.4).

The correlation between the three estimators F_{GRM}, F_{HOM} and F_{UNI} was computed for each of the three genotypes (i.e., homozygotes for allele 1, homozygotes for allele 2 and heterozygotes) for comparison between F_{GRM}, F_{HOM} and F_{UNI} when the allele frequency was varied between 0 and 1 (Fig. 3.6). Correlations reached the maximal value (i.e., 1) when the allele frequencies were 0.5. When the allele frequencies were extremely high or low, correlations between estimators became low, especially the correlation between F_{GRM} and F_{HOM} (0.27). The correlation plot (Fig. 3.6) reflected a similar result as those in Figs. 3.3, 3.4 and 3.5. Therefore, when computing inbreeding coefficients using allele frequencies, populations with different allele frequencies might have very different inbreeding coefficients and the correlations between those inbreeding coefficients might be very low, with different allele frequencies.

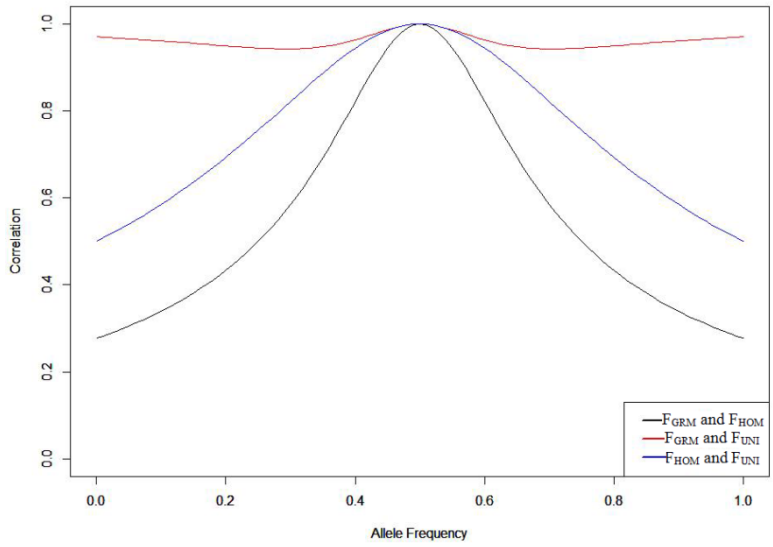


Figure 3.6 Correlations between F_{GRM} and F_{HOM}, F_{GRM} and F_{UNI}, and F_{HOM} and F_{UNI} when reference allele frequency changes between 0 and 1. Black line represents correlation between F_{GRM} and F_{HOM} against reference allele frequencies; red line represents correlation

between F_{GRM} and F_{UNI} against reference allele frequencies and blue line represents correlation between F_{HOM} and F_{UNI} against reference allele frequencies.

The comparison between F_{GRM} and other estimators showed a very low correlation and F_{GRM} was mostly negatively correlated with other estimators. F_{HOM} based on excess of homozygosity was positively correlated with other estimators and was relatively highly correlated with F_{ROH} detected from 50k and sequence data. F_{UNI} based on correlations between uniting gametes estimated from 50k data generally was negatively correlated with other estimators. However, with increasing marker density, the correlation between F_{UNI} and other estimators became positive for the HOL and RDC populations. Surprisingly, when using sequence data, F_{UNI} was highly correlated with other estimators, especially F_{ROH}, detected from sequence data (~0.95) for HOL. This correlation may have resulted from the nature of the estimators: F_{ROH} uses only runs of homozygosity, whereas the other estimators (to some extent) capture all of the homozygosity. This high correlation for F_{UNI} and F_{ROH} compared with low correlation between F_{GRM} and F_{ROH} might also be explained by the algorithms: F_{GRM} = (1 + F)-1 and F is the correlation between uniting gametes. This estimator has only sampling on the F-term, whereas in the F_{GRM} estimator there is also sampling variance on the "1", which creates additional sampling variance.

It is known that RDC is an admixed breed with introgressed haplotypes from Old Danish Red, Holstein and Brown Swiss breeds. HOL and JER are relatively pure breeds and more inbred than RDC (Zhang et al., 2015). Therefore, minor allele frequencies tend to be lower in HOL and JER breeds than in RDC. F_{GRM} is negatively correlated with other estimators for all three breeds. F_{HOM} becomes negative for RDC, which is likely due to the admixture present in RDC. Therefore, it appears that F_{GRM} tends to be less accurate for populations with a low minor allele frequency and that F_{HOM} tends to be less accurate for populations with a higher level of heterozygosity. This argument is supported by our results that the three inbreeding estimators F_{GRM}, F_{HOM} and F_{UNI} were most closely correlated with each other when the allele frequency is approximately 0.5 (Figs. 3.3, 3.4 and 3.5). Therefore, the three estimators F_{GRM}, F_{HOM} and F_{UNI} depend strongly on the estimation of allele frequencies in the population, unlike F_{ROH}. However, here we only took one locus as an example to study the impact of allele frequencies on three estimators F_{GRM}, F_{HOM} and F_{UNI}.

3.5 Conclusion

In this study, we compared different estimators of inbreeding coefficient with different types of data (pedigree, 50k SNP chip genotypes and full sequence data). Methods based on GRM, excess of homozygosity and the correlation between uniting gametes were observed to be sensitive to allele frequencies in the base population. The estimator based on pedigree data was moderately correlated with estimators based on ROH when a pedigree is relatively complete. Estimators based on ROH from SNP chip genotypes and full sequence directly reflect homozygosity on the genome, and have the advantage of not being affected by estimates of allele frequency or incompleteness of the pedigree. Inbreeding estimated from ROH was shown to be affected by the marker density used. Using sequence data, we obtained a full picture of the distribution of ROH on the genome, including short and medium length ROH that reflect ancient inbreeding regions which are possibly IBD. Detecting ROH based on high-density or 50k chip data was shown to give estimates most closely related to ROH from sequence data. However, more than 50k genotypes are required to accurately detect short ROH that are most likely identical by descent (IBD).

3.6 Appendix

Supplementary material can be found in the online version of the published paper or can be directly be accessed via
https://bmcgenet.biomedcentral.com/articles/10.1186/s12863-015-0227-7

Zhang, Q., M. P. Calus, B. Guldbrandtsen, M. S. Lund, & G. Sahana 2015. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics, 16(1), 88.

3.7 Acknowledgement

Q. Zhang benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”. This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research.

References

Andersen, B., B. Jensen, A. Nielsen, L. G. Christensen, and T. Liboriussen. 2003. Rød Dansk Malkerace-avlsmæssigt of kulturhistorisk belyst. Danmarks HordbrugsForskning.

Arias, J. A., M. Keehan, P. Fisher, W. Coppieters, and R. Spelman. 2009. A high density linkage map of the bovine genome. BMC Genetics 10(1):18.

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

Blackwell, B. F., P. D. Doerr, J. M. Reed, and J. R. Walter. 1995. Inbreeding Rate and Effective Population-Size - a Comparison of Estimates from Pedigree Analysis and a Demographic-Model. Biological Conservation 72(3):407-407.

Bosse, M., H. J. Megens, O. Madsen, Y. Paudel, L. A. Frantz, L. B. Schook, R. P. Crooijmans, and M. A. Groenen. 2012. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genetics 8(11):e1003100.

Broman, K. W. and J. L. Weber. 1999. Long homozygous chromosomal segments in reference families from the Centre d'Etude du Polymorphisme Humain. American Journal of Human Genetics 65(6):1493-1500.

Cassell, B. G., V. Adamec, and R. E. Pearson. 2003. Effect of incomplete pedigrees on estimates of inbreeding and inbreeding depression for days to first service and summit milk yield in Holsteins and Jerseys. Journal of Dairy Science 86(9):2967-2976.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. Genetics 134(4):1289-1303.

Charlesworth, D. and B. Charlesworth. 1987. Inbreeding Depression and Its Evolutionary Consequences. Annual Review of Ecology and Systematics 18:237-268.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8(3):175-185.

Ferencakovic, M., E. Hamzic, B. Gredler, T. R. Solberg, G. Klemetsdal, I. Curik, and J. Solkner. 2013. Estimates of autozygosity derived from runs of homozygosity: empirical evidence from selected cattle populations. Journal of Animal Breeding and Genetics 130(4):286-293.

Ferencakovic, M., Solkner, J., & Curik, I. . 2013. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genetics Selection Evolution 45(1):42.

Gonzalez-Recio, O., E. L. de Maturana, and J. P. Gutierrez. 2007. Inbreeding depression on female fertility and calving ease in Spanish dairy cattle. Journal of Dairy Science 90(12):5744-5752.

González-Recio, O., López de Maturana, E., & Gutiérrez, J. P. 2007. Inbreeding depression on female fertility and calving ease in Spanish dairy cattle. Journal of Dairy Science 90(12):5744-5752.

Hoglund, J. K., G. Sahana, B. Guldbrandtsen, and M. S. Lund. 2014. Validation of associations for female fertility traits in Nordic Holstein, Nordic Red and Jersey dairy cattle. BMC Genetics 15(1):8.

Keller, M. C., P. M. Visscher, and M. E. Goddard. 2012. Quantification of Inbreeding Due to Distant Ancestors and Its Detection Using Dense Single Nucleotide Polymorphism Data. Genetics 190(1):283-283.

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(11): e13996.

Ku, C. S., N. Naidoo, S. M. Teo, and Y. Pawitan. 2011. Regions of homozygosity and their impact on complex diseases and traits. Human Genetics 129(1):1-15.

Leroy, G. 2014. Inbreeding depression in livestock species: review and meta-analysis. Animal genetics 45(5):618-628.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, S. Ramachandran, H. M. Cann, G. S. Barsh, M. Feldman, L. L. Cavalli-Sforza, and R. M. Myers. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319(5866):1100-1104.

MacLeod, I. M., D. M. Larkin, H. A. Lewin, B. J. Hayes, and M. E. Goddard. 2013. Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. Molecular Biology Evolution 30(9):2209-2223.

Margolin, S. and J. W. Bartlett. 1945. The Influence of Inbreeding Upon the Weight and Size of Dairy Cattle. Journal of Animal Science 4(1):3-12.

Marras, G., G. Gaspa, S. Sorbolini, C. Dimauro, P. Ajmone-Marsan, A. Valentini, J. L. Williams, and N. P. Macciotta. 2014. Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. Animal genetics 46(2),:110-121.

Mc Parland, S., J. F. Kearney, M. Rath, and D. P. Berry. 2007. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. Journal of Animal Science 85(2):322-331.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297–1303.

McQuillan, R., A. L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, L. Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, O. Polasek, 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. American Journal of Human Genetics 83(3):359-372.

Meuwissen, T. H. E. and Z. Luo. 1992. Computing Inbreeding Coefficients in Large Populations. Genetics Selection Evolution 24(4):305-313.

Miglior, F., B. Szkotnicki, and E. B. Burnside. 1992. Analysis of Levels of Inbreeding and Inbreeding Depression in Jersey Cattle. Journal of Dairy Science 75(4):1112-1118.

Nomura, T., T. Honda, and F. Mukai. 2001. Inbreeding and effective population size of Japanese Black cattle. Journal of Animal Science 79(2):366-370.

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

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

Purfield D C, B. D. P., McParland S, et al. 2012. Runs of homozygosity and population history in cattle. BMC Genetics 13(1):70.

Pusey, A. and M. Wolf. 1996. Inbreeding avoidance in animals. Trends in Ecology & Evolution 11(5):201-206.

Quaas, R. L. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. Biometrics:949-953.

Ramey, H. R., J. E. Decker, S. D. McKay, M. M. Rolf, R. D. Schnabel, and J. F. Taylor. 2013. Detection of selective sweeps in cattle using genome-wide SNP data. BMC Genomics 14(1):382.

Smith, L. A., B. G. Cassell, and R. E. Pearson. 1998. The effects of inbreeding on the lifetime performance of dairy cattle. Journal of Dairy Science 81(10):2729-2737.

Sorensen, A. C., M. K. Sorensen, and P. Berg. 2005. Inbreeding in Danish dairy cattle breeds. Journal of Dairy Science 88(5):1865-1872.

Strandén, I. and K. Vuori. 2006. Relax2 : pedigree analyses program. Proceedings of the 8th WCGALP 13-18 Aug. 2006, Belo Horizonte, MG, Brazil.

Szpiech, Z. A., J. Xu, T. J. Pemberton, W. Peng, S. Zollner, N. A. Rosenberg, J. Z. Li. 2013. Long runs of homozygosity are enriched for deleterious variation. American Journal of Human Genetics 93(1):90-102.

VanRaden, M. 1992. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. Journal of Dairy Science 75(11):3136-3144.

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

Weigel, K. 2006. Controlling inbreeding in modern dairy breeding programs. In Advances in dairy technology: proceedings of the Western Canadian Dairy Seminar 18:263-274.

Wright, S. 1921. Systems of mating. II. The effects of inbreeding on the genetic composition of a population. Genetics 6(2):124-143.

Wright, S. 1922. Coefficients of inbreeding and relationship. The American Naturalist 56:330-338.

Wright, S. 1948. Genetics of Populations. Encyclopaedia Britannica 10:111-A-D-112.

Yang, J. A., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. American journal of human genetics 88(1):76-82.

Zhang, Q., B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. 2015. Runs of homozygosity and distribution of functional variants in the cattle genome. BMC Genomics 16(1):542.

4

Detection of introgressed genomic regions in modern dairy breeds: A case study of the hybrid Modern Danish Red cattle

Qianqian Zhang[1,2], Mario Calus[2], Mirte Bosse[2], Goutam Sahana[1], Mogens S Lund[1] and Bernt Guldbrandtsen[1]

[1] 1Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, Tjele DK-8830, Denmark; [2] Animal Breeding and Genomics, Wageningen University & Research, Wageningen 6700 AH, The Netherlands.

Abstract

Different breeding strategies including introgression and artificial selection have generated numerous desirable phenotypes and superior performance in domestic animals. In dairy cattle, high yielding breeds are introduced to the local breed to improve milk production and performance in dairy cattle due to the change in production subsidy regimen. With availability of whole genome sequencing, we are able to detect genomic signals of introgression from high yielding breeds. The Modern Danish Red Dairy Cattle is studied as an example of a composite breed. It originates from crossing of Traditional Danish Red Dairy Cattle with Holstein and Brown Swiss known for high milk production. We demonstrated that the genomes of hybrid Modern Danish Red cattle are heavily influenced by contributions from Holstein and Brown Swiss. Some of haplotypes derived from Holstein and Brown Swiss are under selection and presumably have spread due to selection. Genes previously identified as affecting milk production, and protein and fat content (*BOLA* and *THRSP* for Holstein introgressed haplotypes; *ITPR2*, *BCAT1*, *ZNF215* for Brown Swiss introgressed haplotypes) were found to overlap haplotypes introgressed from both Holstein and Brown Swiss. Some of these regions introgressed from Holstein or Brown Swiss also contain genes and QTLs associated with calving traits, body confirmation, feed efficiency, carcass, and fertility traits. Some of these introgressed regions are also under selection in Modern Danish Red Dairy Cattle. A combination of selection, recombination and drift has resulted in the genomic footprints of introgression in the genome of hybrid Modern Danish Red Dairy Cattle. Introgressed haplotypes from Holstein or Brown Swiss have made a major contribution to the genome of the hybrid Modern Danish Red Dairy Cattle. Our findings contribute to an understanding of genomic consequences of selective introgression into a modern dairy cattle breed.

Key words: Introgression, Dairy cattle, Signature of selection, Modern Danish dairy cattle, Holstein cattle, Brown Swiss cattle

4.1 Background

Artificial selection and breeding have enabled generating numerous desirable phenotypes in domestic animals such as cattle (Hartwig et al., 2015, Buzanskas et al., 2017, Davis et al., 2017) , pigs (Bosse et al., 2014a, Ai et al., 2015, Bosse et al., 2015) and dogs (Galov et al., 2015, vonHoldt et al., 2016). Strategies including crossbreeding and introgression have been very successful in improving productivity and performance in domestic animals. For example, Chinese pig breeds have been imported to Europe to improve the productivity of European pigs in the late eighteenth and early nineteenth centuries (White, 2011). The fertility related traits have been largely improved by the crossbreeding and introgression from Asian pigs (Merks et al., 2012, Bosse et al., 2014a, Bosse et al., 2015). Similarly, in dairy cattle, crossbreeding and following introgression between local breeds with other breeds has been applied in order to achieve better productivity and performance (Davis et al., 2017). The genetic architecture of modern domestic animals including dairy cattle is shaped by the interplay between different forces including the intentional introduction of favorable alleles from other breeds, subsequent selection for favorable introgressed alleles and demographic processes.

Since the introduction of scientific theory in animal breeding, the main breeding goal in cattle has been to improve milk yield, fat, and protein content, even for dual-purpose breeds in intensive farming systems especially in Europe (Hartwig et al., 2015). By crossing of high yielding breeds into local breeds, the productivity of local breeds could rapidly be increased at the expense of the genetic distinctiveness of the local breeds. The high productivity of these admixed breeds was further improved by intense selection, resulting in increased frequencies or even fixation of favorable alleles. Many of the alleles thus spreading in the population will have been of introgressed origin. With the great success of cattle breeding including crossbreeding, and because of the availability of large scale genomic data sets, analysis of admixed local cattle breed represents an appealing model to identify the genomic regions affecting traits of interest from other breeds. We expect that these introgressed genomic regions play an important role in improving productivity and performance in the crossbreeds. We hypothesize that: 1) the genome-wide introgressed regions are non-randomly distributed across the genome with respect to their genomic locations, 2) these detected introgressed regions affect production traits, and 3) they are or have been under selection.

To test these hypotheses and identify specific genomic regions and genes of interest involved in the important production traits from high-yields breeds such as Holstein (HOL) and Brown Swiss (BSW), we use the hybrid Modern Danish Red Dairy Cattle (mRDC) breed as an example. Our analyses illustrate the patterns of introgressed and selected haplotypes in an admixed local breed. The hybrid mRDC originates from traditional Danish Red Dairy Cattle (tRDC). In recent decades Holstein and Brown Swiss have been used extensively to improve the milk yield of the breed (Kantanen et al., 2000). Years of crossbreeding and selection have led to the differentiation between mRDC and tRDC. The introgression of and selection for haplotypes from HOL and BSW has probably made a significant contribution to the increased milk production level of mRDC. The objective of this study is to examine genomic patterns of introgression from high-yielding breeds (HOL and BSW) in the hybrid mRDC, and unravel the consequences of introgression and artificial selection at the genome level using the whole genome sequencing data.

4.2 Methods

4.2.1 SNP genotyping, sequencing, variant calling, and quality control

Whole genome sequence data were available for 213 animals from 7 breeds (84 Holstein: HOL; 18 Brown Swiss: BSW; 20 traditional Red Danish Dairy Dairy Cattle: tRDC; 30 modern Danish Red Dairy cattle: mRDC; 27 Jersey: JER; 16 Swedish Red Dairy Cattle: RDCSWE; 18 Finnish Red Dairy Cattle: RDCFIN). All individuals' genomes were sequenced to ~10× of depth or deeper using Illumina paired-end sequencing. Reads were aligned to the cattle genome assembly UMD3.1 (Zimin et al., 2009) using bwa (Li and Durbin, 2009). Aligned sequences were converted to raw BAM files using samtools (Li et al., 2009). Duplicate reads were marked using the samtools rmdup option (Li et al., 2009). The Genome Analysis Toolkit (McKenna et al., 2010) was used for local realignment around insertion/deletion (indels) regions, and recalibration following the 1000 Bull Genome Project guidelines (Daetwyler et al., 2014) incorporating information from dbSNP (McKenna et al., 2010). Finally, variants were called using the Genome Analysis Toolkit's Unified Genotyper (McKenna et al., 2010), which simultaneously calls short indels and SNPs. Indels were excluded in further analyses.

Illumina Bovine HD Beadarray data were available for 1,889 animals from 8 breeds. (135 tRDC, 245 mRDC, 158 HOL, 382 RDCNOR, 246 RDCSWE, 243 RDCFIN, 60 Danish Shorthorn, 420 JER). Subsets of markers on the Illumina HD chip detected in the

1000 Bull Genomes Project Run 4 data (Daetwyler et al., 2014) were extracted for 123 Brown Swiss bulls. Only 302,319 markers appearing in all data sets were retained for further analysis.

4.2.2 Population structure

4.2.2.1 Principle component analysis

To get an overview of population structure of the genotyped animals from different breeds, the full Illumina HD dataset (including BSW data) was used for principal component estimation using the program smartpca from the Eigensoft package (Price et al., 2006).

4.2.2.2 Admixture analysis

The sequence and SNP chip variants were pruned to remove markers with pairwise Linkage disequilibrium (LD) greater than 0.1 with any other SNP within a 50 SNP sliding window (advancing by 10 SNPs at a time), because the program Admixture does not take LD into consideration. The remaining markers (consisting of 107,915 SNPs) were used in the admixture analysis. These were analyzed using the program Admixture by Pritchard et al. (2000). The analysis was done with K between 2 and 20. Cross validation was done, and a K value with a low cross validation error was chosen. The log likelihood curve (Supplementary figure 1) shows a progressive improvement with higher numbers of components and no evidence of any leveling off until 11 and after 11, the log likelihood remains roughly constant. The results shown are for K=11 at which the phenotypically distinct Danish Shorthorn breed was assigned a separate component by Admixture.

4.2.3 Introgression mapping

HOL, BSW and tRDC have made large genetic contributions to the mRDC. Therefore, 30 sequenced mRDC, 15 tRDC, 20 BSW and 32 HOL were selected for introgression mapping analysis. The identity-by-descent (IBD) regions comparing mRDC and tRDC were used as a reference to map the introgression regions from HOL and BSW using a pairwise comparison between these breeds. Following the method for introgression mapping from Bosse et al. (2014b), sequences for 29 autosomes were first phased separately by Beagle fastIBD (V. 3.3.2) (Browning and Browning, 2007). Pairwise comparisons for detecting IBD were performed between mRDC and tRDC; mRDC and HOL; mRDC and BSW. As suggested in the Beagle documentation (Browning and Browning, 2007), 10 independent runs for phasing and pairwise IBD detection were performed. The identified IBD segments were

combined from 10 runs and the threshold parameter compromising between power and false-discovery rate was 10^{-10} for identifying the true shared IBD as suggested by Browning and Browning (2007). We defined the IBD score as the proportion of the number of recorded true IBD haplotype segments over the total number of pairwise comparisons using a window of 10 kbp. The IBD score was calculated for each pairwise comparison using a custom perl script. To quantify the relative proportion of introgressed genome from HOL or BSW, we calculated the relative IBD score (rIBD) as follows: IBD score (mRDC & tRDC) - IBD score (mRDC & HOL) or IBD score (mRDC & BSW). Thus, the rIBD has values in the range of -1 to 1. rIBD=1 signifies that all haplotypes in the target breed originate from the first source breed, while rIBD=-1 signifies 100% from the second source breed. The z transformed relative IBD score (Z_{rIBD}) was calculated as the deviation of rIBD from the mean rIBD in units of standard deviations. The threshold for extreme rIBD introgressed from HOL or BSW was set to 3 times the square root of the robust variance estimate of the rIBD scores across the whole genome from the mean in the both tails of the distribution.

4.2.4 GO-enrichment analysis

All annotated genes in Bos taurus genome (UMD3.1) were extracted from Ensembl (Yates et al., 2016). GO-enrichment analysis was performed for genes overlapping the top 2.28% ($\delta>2$ for ZrIBD) of regions with an overrepresentation of HOL or BSW haplotypes in mRDC. The DAVID 6.8 Beta (Jiao et al., 2012) was used to identify overrepresented biological process-related GO terms. Significance levels were adjusted based on the Benjamini-Hochberg for multiple comparisons (Benjamini and Hochberg, 1995).

4.2.5 Detection of signature of selection

4.2.5.1 Fst analysis

The genetic differentiation between individuals from tRDC and mRDC was measured by pairwise Fst analysis following Weir and Cockerham (1984). Pairwise Fst was computed with Genepop 4.2 in bins of 10 kb over the full length of the genome (Weir and Cockerham, 1984). The correlations between the Fst and rIBD scores for HOL and BSW introgression for the same bins of 10 kb were calculated.

4.2.5.2 Extended haplotype homozygosity tests

The extended haplotype homozygosity tests were applied between the breeds for the sequenced individuals as a second test for signatures of selection. The genome-

wide scan for integrated haplotype score (iHS) for mRDC was performed using the R package rehh (Sabeti et al., 2002, Gautier and Vitalis, 2012). The significance levels (the corresponding p-values, assuming iHS are normally distributed under the neutral hypothesis) within mRDC were averaged for a bin of 10 kb and were correlated with rIBD score for HOL or BSW introgression for the same bin of 10 kb by Pearson's correlation. The genomic regions with p less than 2 were extracted as significant regions in iHS test.

4.2.5.3 Sharing of runs of homozygosity (ROH)

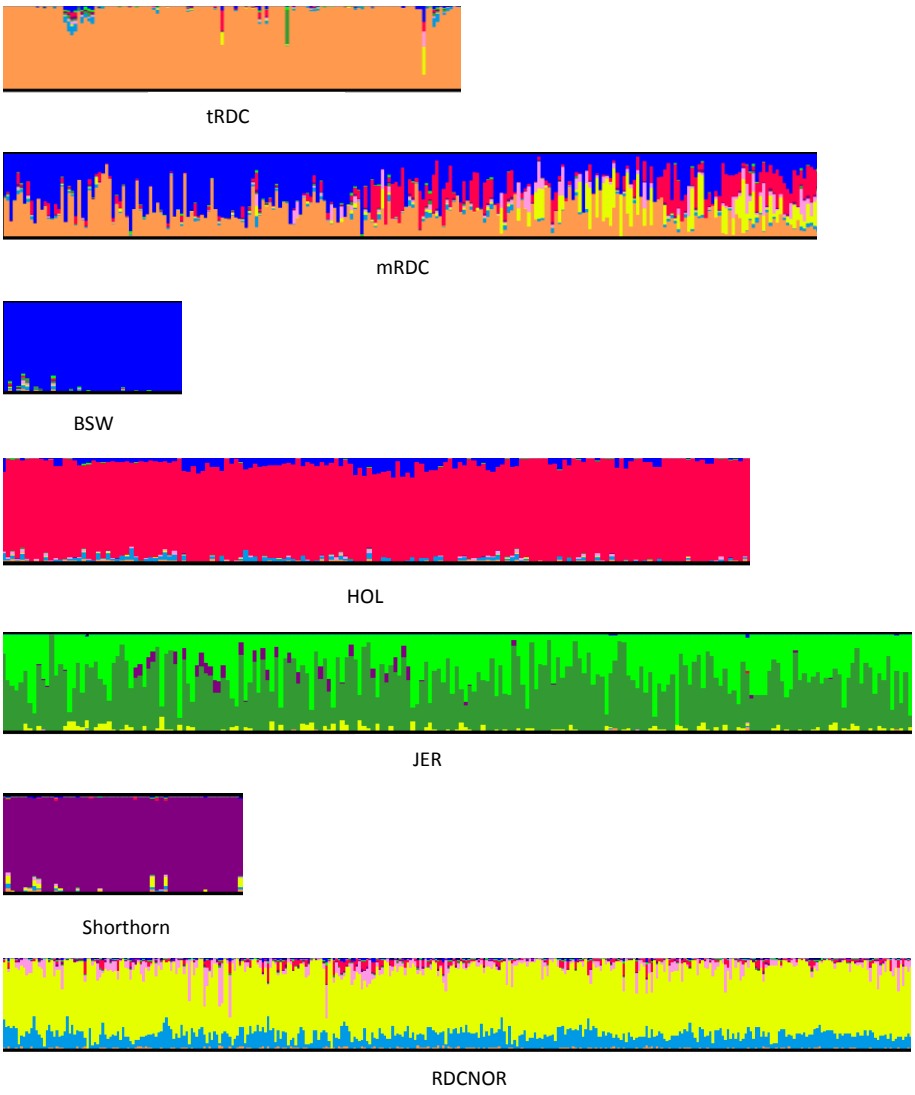
Runs of homozygosity (ROH) were computed for the sequenced animals to detect shared short ROH among individuals. For a description of procedures for calculation of the nucleotide diversity and for detection of ROH, see (Zhang et al., 2015a, Zhang et al., 2015b). The sharing of ROH regions was calculated as the number of individuals sharing the same ROH region on a particular segment using a window of 10 kb bin across the whole genome in mRDC. The sharing of ROH regions was correlated with rIBD score for HOL or BSW introgression for the same bin of 10 kb by Pearson's correlation.

4.3 Results and Discussion

4.3.1 Population structure and evidence of introgression

Population structure of the sampled cattle was analyzed using Admixture and PCA shown in figure 4.1 and 4.2. We observed that mRDC had contribution from tRDC, HOL, BSW, Shorthorn, RDCNOR, RDCSWE and RDCFIN in the Admixture analysis (Fig. 4.1). Figure 4.1 clearly demonstrates the hybrid nature of mRDC cattle which was consistent with recorded pedigree information of introgression history of mRDC. The largest contribution in mRDC comes from tRDC, the recipient population. It is notable that BSW and HOL are two mainstream breeds, each contributing heavily to the genomes of extant mRDC individuals. Red cattle breeds in other Nordic countries other than Denmark (RDCNOR, RDCSWE and RDCFIN) also have contributed to mRDC. In the PCA analysis, Principal Component 1 (PC1, 7.0 % of variance, supplementary Figure S2) separated Jersey from all the other breeds. PC2 and PC3 separate HOL, BSW, tRDC, RDCNOR, RDCFIN, RDCSWE and Shorthorn accordingly. mRDC, however, was dispersed between the other breeds demonstrating admixture of the other breeds which have contribution to mRDC (Fig. 4.2). PC2 (4.3 % of variance) separated northern RDC in Norway, Sweden and Finland (RDCNOR, RDCSWE and RDCFIN) from tRDC. PC3 (2.6 % of variance)

distinguished BSW from the other breeds. Plotting PC2 against PC3 (Fig. 4.2a) individuals belonging to mRDC were spread out between tRDC, BSW and northern RDC. PC4 (2.0 % of variance) separate Holstein, Danish Shorthorn and SDM1965 from the other breeds (Fig. 4.2b). These results support that mRDC indeed is a composite breed and can be used to study the introgression from HOL and BSW.



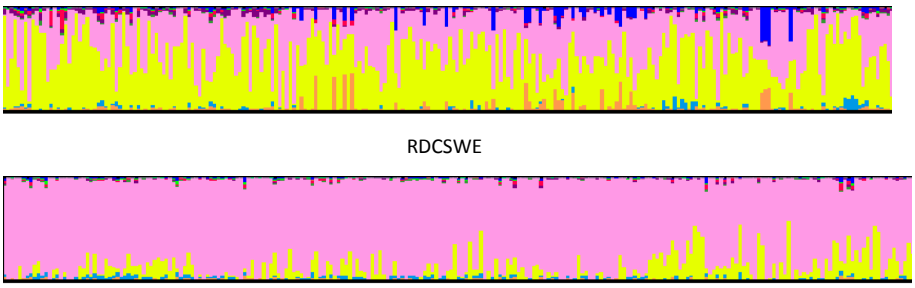


Figure 4.1 Admixture analysis of different cattle breeds with k=9. BSW – Brown Swiss, HOL – Holstein, tRDC – traditional Red Dairy Cattle, mRDC – modern Red Dairy cattle, JER – Jersey cattle, RDCNOR – Norwegian Red Dairy Cattle, RDCSWE – Swedish Red Dairy Cattle, RDCFIN – Finish Red Dairy Cattle .

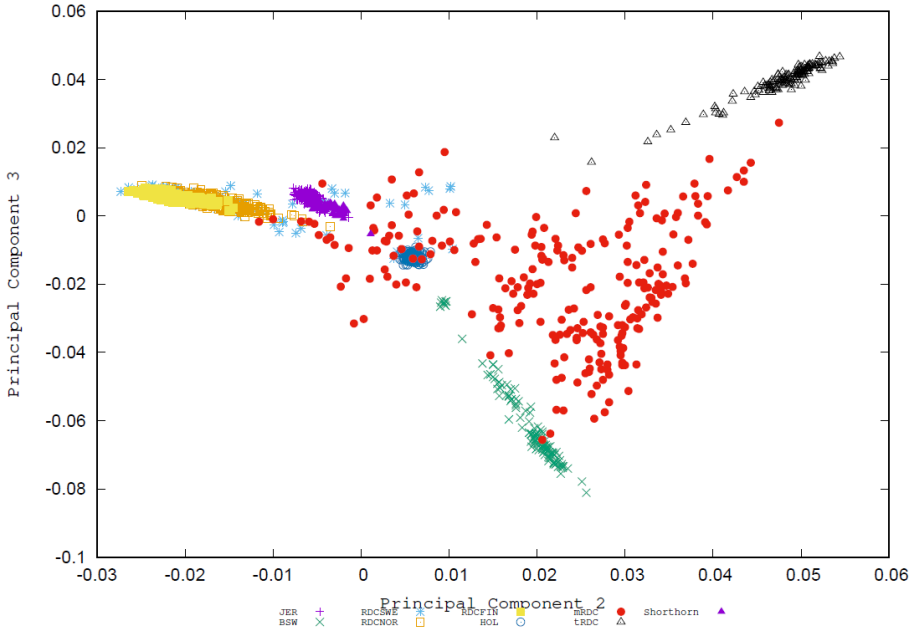


Figure 4.2 Principal component analysis (PCA) plots among different cattle breeds (Principal component 2 v.s. principal component 3). BSW – Brown Swiss, HOL – Holstein, tRDC – traditional Red Dairy Cattle, mRDC – modern Red Dairy cattle, JER – Jersey cattle, RDCNOR – Norwegian Red Dairy Cattle, RDCSWE – Swedish Red Dairy Cattle, RDCFIN – Finish Red Dairy Cattle.

4.3.2 Introgression mapping

In order to infer whether a region was introgressed from different breeds in multiple individuals, the frequencies of all mRDC haplotypes that were of HOL, BSW or tRDC origin were calculated in all replicates using different seeds for IBD detection in Beagle using a sliding window of the bin of 10,000 bases in the whole genome. The relative fractions of HOL or BSW haplotypes versus tRDC haplotypes in the mRDC group were calculated as relative IBD (rIBD) scores. Shared haplotypes were observed between mRDC on one hand, and HOL, BSW and tRDC on the other hand (Fig. 4.3 and 4.4), in agreement with the result observed from population structure analysis that shows HOL, BSW and tRDC have contribution to mRDC. The HOL haplotype or BSW haplotype frequency in mRDC population, for a given locus, (i.e. rIBD score) ranged from 0.73 to -0.74 and from 0.81 to -0.92, where 1 indicates that all haplotypes are HOL haplotype or BSW haplotype and none are tRDC haplotype, and -1 indicates that all haplotypes are tRDC-like. The rIBD scores averaged across the genome were negative (-0.06 for HOL introgression and -0.09 for BSW introgression) (Figures 4.3b and 4.4b), showing that the majority of the genome displays more similarity with the tRDC than with either HOL or BSW. However, every chromosome contained genomic regions where the signal for HOL or BSW haplotype was stronger than tRDC.

We observed that the distribution of rIBD from the comparison between mRDC and HOL or BSW for IBD haplotypes resembled a normal distribution (Fig. 4.3b and 4.4b). By taking rIBD values greater than the threshold defined in Method part 3, we are able to identify the regions which were likely to be HOL origin or BSW origin. Across the whole genome, many known genes and QTLs are located within the regions that are likely of HOL origin. The genes and QTLs are associated with many important traits including milk-related traits such as milk yield, protein, fat yield and percentage (*BOLA* and *THRSP*) (Bennewitz et al., 2003, Fontanesi et al., 2014), calving traits (*SYT3*, *RPS9*, and *ABCC1*) (Kolbehdari et al., 2008, Cole et al., 2011, Moore et al., 2016), feed efficiency related traits (*ATP6V1B2*, *CCKBR*) (Abo-Ismail et al., 2013), carcass traits (*RPTOR*, *INTS4*) (Sasago et al., 2017), despite that there is no GO term significantly enriched in the gene list. The longest introgressed region (defined as the region with rIBD>0) from HOL to mRDC is on chromosome 18 (56,320,000-61,340,000 bp) (Fig. 4.1a). This region was found to be associated with calving traits in Holstein population (Cole et al., 2011, Mao et al., 2015). It was shown that the recombination rate of this region on chromosome 18 was low (Weng et al., 2014). The long genomic regions showing signal of introgression tend to locate in the regions with low recombination rate. Moreover, the supposedly introgressed haplotype includes numerous annotated genes due to genetic

hitchhiking and short time of introgression. The region with highest rIBD score is located on chromosome 4 (120,540,000-120,800,000 bp), which is on the downstream of gene *VIPR2*. Gene *VIPR2* was proposed to be a candidate gene affecting fat percentage and playing an important role in milk synthesis (Capomaccio et al., 2015).

Similarly, many known genes and QTLs are enriched in the regions where mRDC shared haplotypes with BSW, although there is still no GO term significantly enriched in the gene list. These genes and QTLs are mainly affecting bovine growth and body conformation traits such as stature (*NCAPG*, *LCORL*, *ZNF215*, *PPP2R1A*) (Magee et al., 2010, Lindholm-Perry et al., 2011, Cole et al., 2014, Sahana et al., 2015), milk composition including fat and protein percentage and yield (*ITPR2*, *BCAT1*, *ZNF215*) (Magee et al., 2010, Pimentel et al., 2011, Fang et al., 2014), fertility (*EIF4G3*, *TGFA*, *APBB1*) (Hering et al., 2014, Hoglund et al., 2015, Ortega et al., 2016), and feed efficiency related traits (*CLMP*, *CCKBR*) (Abo-Ismail et al., 2013, Serao et al., 2013), although there is no GO term significantly enriched still. While the longest region introgressed from HOL in mRDC was ~4 Mb, there is no introgressed region longer than 1.5 Mb from BSW in mRDC. The highest peak of rIBD signal is observed on chromosome 17 (35,630,000-35,990,000 bp). This region locates on the downstream of *IL2* gene and there are three unannotated genes in this region. It was shown that *IL2* gene was associated with milk yield and lactation persistency (Prakash et al., 2011).

4.3.3 Regions of introgression and evidence for selection

We have shown that HOL or BSW haplotypes that showed introgression in mRDC often originate from genomic regions harboring genes which are associated with milk production, calving traits, feed efficiency, fertility and body conformation. The genomic regions showing signals of introgression from HOL or BSW are probably a result of combination of drift and selection and the length of the haplotype is affected by the local recombination rate. Some of the introgressed haplotypes, however, will be eliminated by purifying selection due to selective disadvantage while some will be selected due to selective advantage. We used three independent methods, iHS, Fst, and sharing of ROH among individuals, to identify the regions which are under selection. iHS could identify the regions which are showing extended homozygosity in mRDC with signal of selection due to hitchhiking. The sharing of ROH among individuals could differentiate the genomic regions which are already fixed during selection process in mRDC. The Fst statistics

reflects the genomic regions which show highly differentiation between mRDC and tRDC.

We observed significant correlations between Fst and positive rIBD scores from both HOL and BSW (p<0.001), supporting that at least some of the regions in mRDC showing differentiation from tRDC are introgressed from HOL or BSW (the regions fall in the top right corner of Figure 4.5b and 4.6b) (Fig. 4.5 and 4.6). The longest region introgressed from HOL in mRDC also showed high Fst values (> 2 times s.d. of overall Fst mean=0.155) in some of the 10 kbp bins. There are genes overlapping with these bins with high Fst values in this region such as *SIGLEC1*, *VSIG10L* and *IGLON5* associated with disease and immune response in human (Sabater et al., 2014, Fecteau et al., 2016, Komohara et al., 2017), while there are QTLs in this region associated with calving traits in cattle. The region on chromosome 19 (52370000- 52380000 bp) with Fst of 0.748 show highly differentiation between tRDC and mRDC which overlaps with HOL introgressed haplotype in mRDC and gene *RPTOR* associated with carcass traits in cattle (Sasago et al., 2017) locate in this region. Moreover, the region on chromosome 29 (18210000-18220000 bp) has a Fst of 0.164, which also locates with the HOL haplotypes introgressed in mRDC and overlaps with gene *INTS4* associated with carcass traits (Sasago et al., 2017). Similarly, we found that the region on chromosome 6 (38,730,000- 38,780,000 bp) with an average Fst value of 0.251 overlaps with BSW haplotypes introgressed in mRDC and gene *NCAPG* associated with body confirmation such as stature, and feed efficiency (Lindholm-Perry et al., 2011, Sahana et al., 2015) locates in this region.

As shown in Fig. 4.7 and 4.8, we compare between the introgressed regions from HOL or BSW in mRDC and the significant regions from iHS test. There is no positive correlation between rIBD score for introgressed regions from HOL in mRDC and iHS test significance. However, we do observe that the positive correlation between rIBD score for introgressed region from BSW in mRDC and iHS test signicifance. For example, significant region in iHS test (85,460,000 - 85,470,000 bp; p > 2) on chromosome 5 overlap with introgressed region from BSW in mRDC. Gene *BCAT1* locates in this region and associated with milk yield, protein and fat percentage in milk (Pimentel et al., 2011). There is one region showing high significance in iHS test (46,500,000- 46,600,000; p > 2) on chromosome 15 also overlapping with introgressed region from BSW in mRDC. Gene *OR6A2* locates in this significant region from iHS test, which is a gene in the family of olfactory receptor genes (Kurland et al., 2010). Meanwhile, this region also locates in the downstream of

gene *ZNF215* affecting body confirmation and milk composition (Magee et al., 2010), upstream of gene *APBB1* affecting fertility (Ortega et al., 2016). The genomic regions showing signals both from iHS test and introgression mapping, e.g. gene *OR6A2* and *BCAT1*, are introgressed from BSW, and probably under selection but not yet fixed in the population due to low pressure of selection.

For the comparison between overlapping of ROH regions and introgressed regions from HOL or BSW, we observed significant correlation (p<0.001) between the number of individuals of overlapping of ROH regions and rIBD score from HOL but not from BSW, which is opposite to comparison between iHS test and rIBD score that significant positive correlation was observed for BSW but not HOL (Fig. 4.9 and 4.10). The short ROHs sharing between multiple individuals indicate selection in the population and probably are already fixed. We could probably, therefore, infer that HOL introgression in mRDC is probably long time back, and some favorable alleles have been fixed already. The fixation of these regions is also subject to selection pressure. Similarly, we set the threshold of the regions where the number of individuals sharing ROH regions more than twice of mean plus twice of standard deviation as showing significantly enriched ROH regions. Interestingly, many small regions, which show high enrichment of ROH regions, overlap with the longest introgressed region from HOL in mRDC e.g. the region where gene *SYT3* locates, which affecting calving traits (Kolbehdari et al., 2008, Cole et al., 2011). Similarly, we also observed that the region introgressed from HOL in mRDC, where gene *THRSP* and *INTS4* locates, shows highly enrichment of ROH region among individuals. Studies showed that gene *THRSP* is associated with milk composition and involved in the regulation of mammary synthesis of milk fat (Fontanesi et al., 2014) and gene *INTS4* is associated with myristic acid content affecting carcass traits (Sasago et al., 2017). For the genomic regions introgressed from BSW in mRDC, we observed two regions, where gene *TGFA* associated with sperm motility (Hering et al., 2014), and gene *PPP2R1A* associated with body weight (Cole et al., 2014) locate, show high level of sharing of ROH regions between individuals.

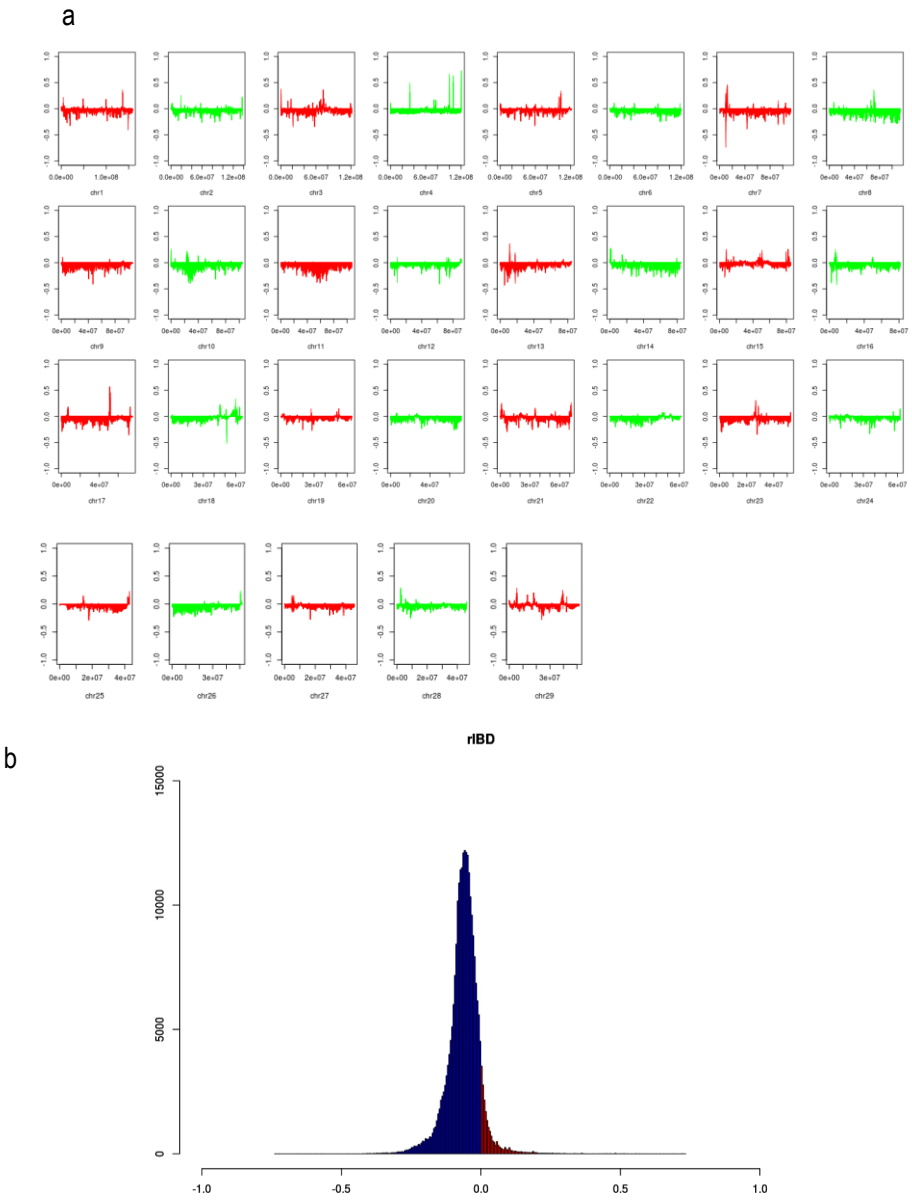


Figure 4.3 Genome-wide pattern of relative identity-by-descent (rIBD) score showing introgressed haplotypes from Holstein (HOL) in modern Danish Red dairy cattle (mRDC). a. the rIBD score for all 29 autosomes: the positive scores show the signals where it is more HOL-like whereas the negative scores show the signals where it is more traditional Danish red cattle (tRDC)-like. b. the distribution of rIBD score: the positive scores show the signals where it is more HOL-like whereas the negative scores show the signals where it is more tRDC-like.

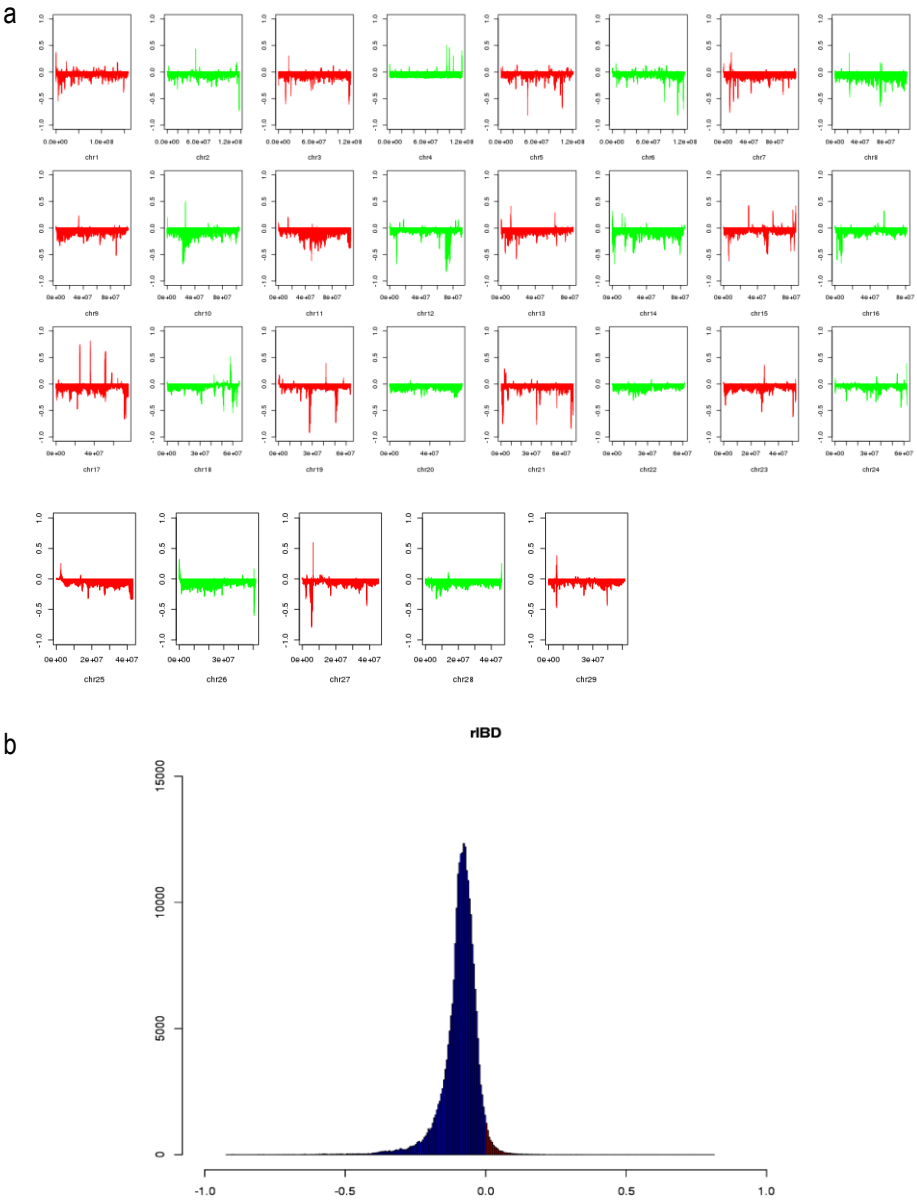


Figure 4.4 Genome-wide pattern of relative identity-by-descent (rIBD) score showing introgressed haplotypes from Brown Swiss (BSW) in modern Danish Red dairy cattle. a. the rIBD score for all 29 autosomes: the positive scores show the signals where it is more BSW-like whereas the negative scores show the signals where it is more traditional Danish red cattle (tRDC)-like. b. the distribution of rIBD score: the positive scores show the signals

where it is more HOL-like whereas the negative scores show the signals where it is more tRDC-like.

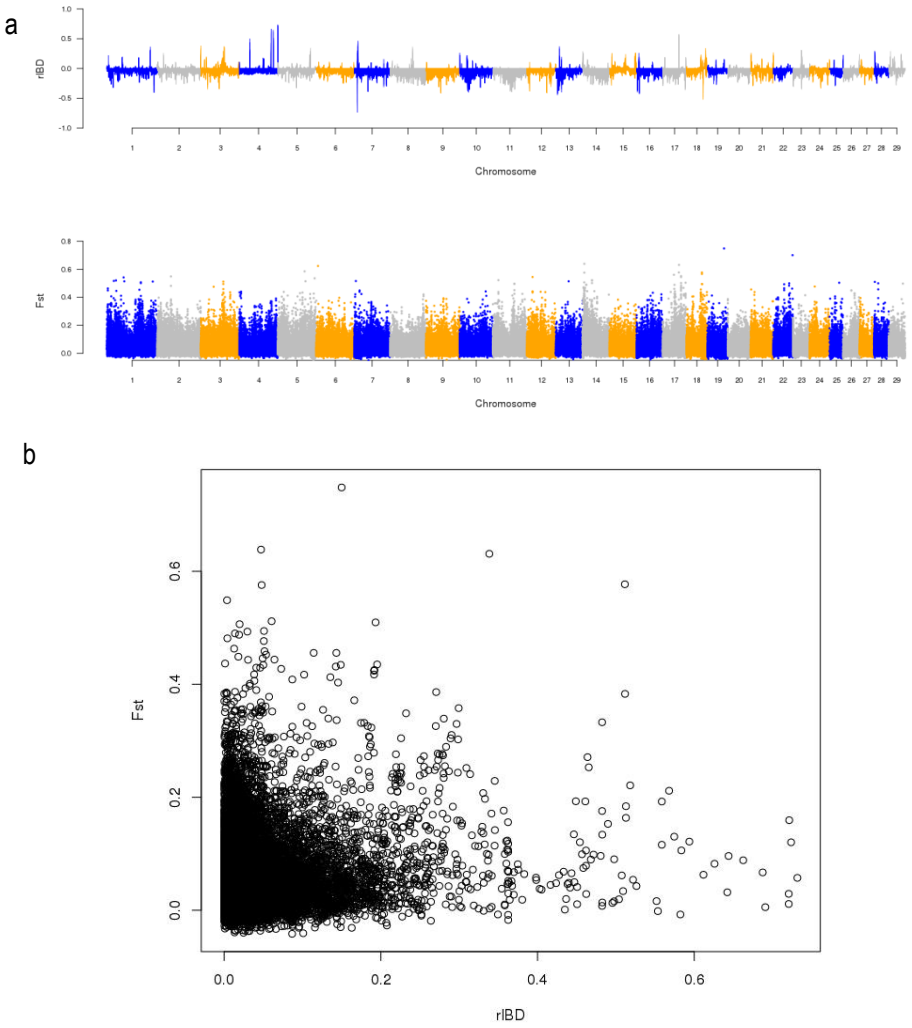


Figure 4.5 Comparison and correlation between Fst between modern and traditional Red Dairy cattle (mRDC and tRDC) and relative identity-by-descend (rIBD) score introgressed from Holstein (HOL) cattle in mRDC. a. comparison between Fst between mRDC and tRDC and rIBD score showing introgression from HOL cattle in mRDC. b. correlation between Fst between mRDC and tRDC and rIBD score showing introgression from BSW cattle in mRDC ($p<0.001$ and correlation =0.09).

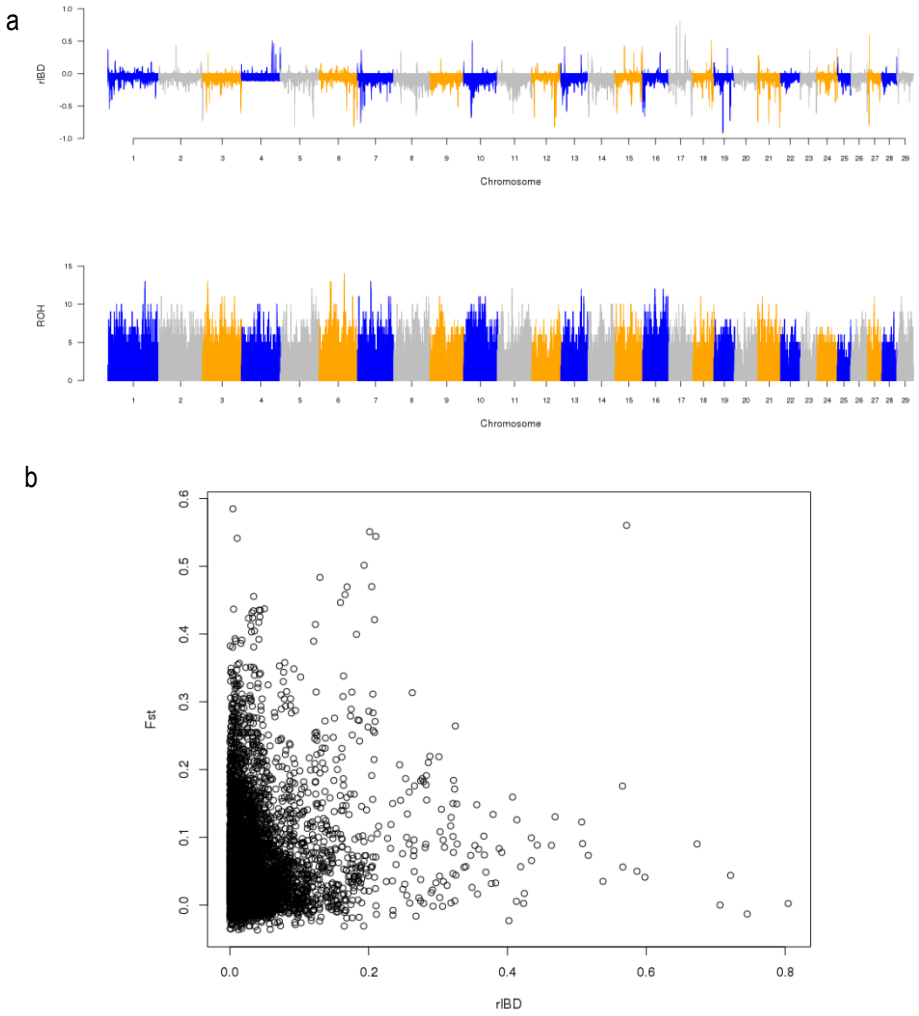


Figure 4.6 Comparison and correlation between Fst between modern and traditional Red Dairy cattle (mRDC and tRDC) and relative identity-by-descend (rIBD) score introgressed from Brown Swiss (BSW) cattle in mRDC. a. comparison between Fst between mRDC and tRDC and rIBD score showing introgression from BSW cattle in mRDC. b. correlation between Fst between mRDC and tRDC and rIBD score showing introgression from BSW cattle in mRDC ($p<0.001$ and correlation =0.09).

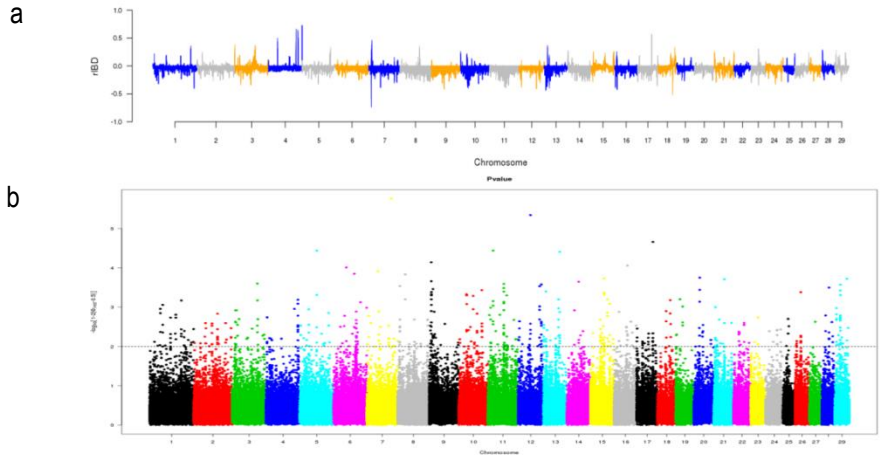


Figure 4.7 Comparison and correlation between iHS score detected within modern Danish Red dairy cattle (mRDC) and relative identity-by-descend (rIBD) score showing introgression from Holstein (HOL) cattle in mRDC (p<0.001 and correlation=-0.04). a. genome-wide pattern of rIBD scores showing introgression from HOL in mRDC. b. iHS sore detected within mRDC (p=-log[1-2|Φ(iHS)-0.5|]).

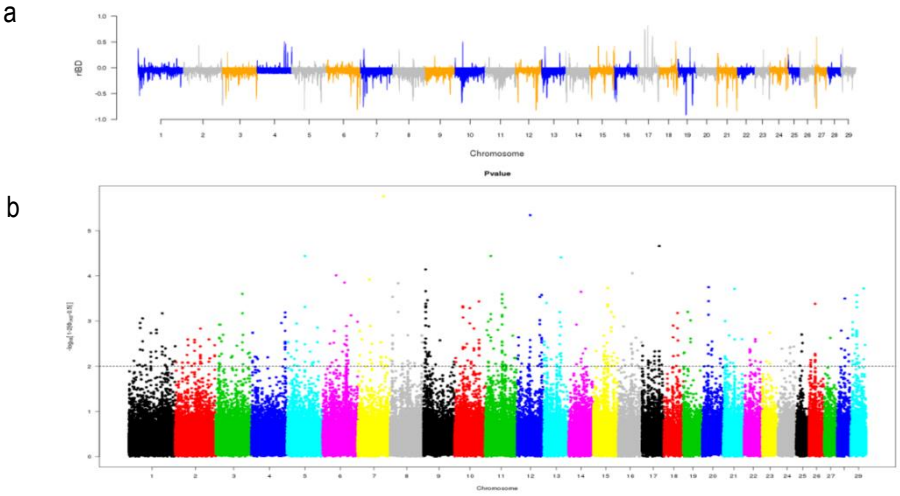


Figure 4.8 Comparison and correlation between iHS score detected within modern Danish Red dairy cattle (mRDC) and relative identity-by-descend (rIBD) score showing introgression from Brown Swiss cattle (BSW) in mRDC (p<0.05 and correlation=0.02). a. genome-wide pattern of rIBS scores showing introgression from BSW in mRDC. b. iHS sore detected within mRDC (p=-log[1-2|Φ(iHS)-0.5|]).

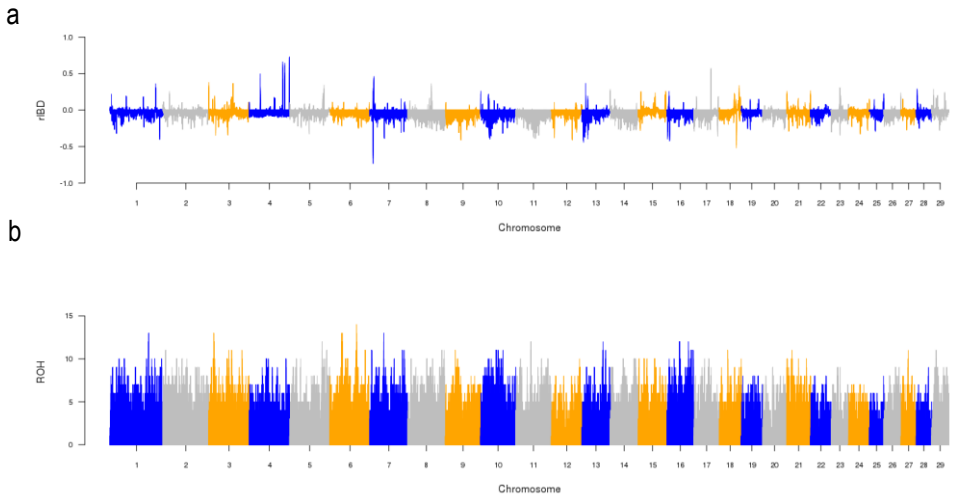


Figure 4.9 Comparison between Overlapped runs-of-homozygosity (ROH) among modern Red dairy cattle (mRDC) and relative identity-by-descend (rIBD) score showing introgression from Holstein cattle (HOL) in mRDC (p<0.001 and correlation =0.13). a. genome-wide pattern of rIBS scores showing introgression from HOL in mRDC. b. sharing of ROH among mRDC.

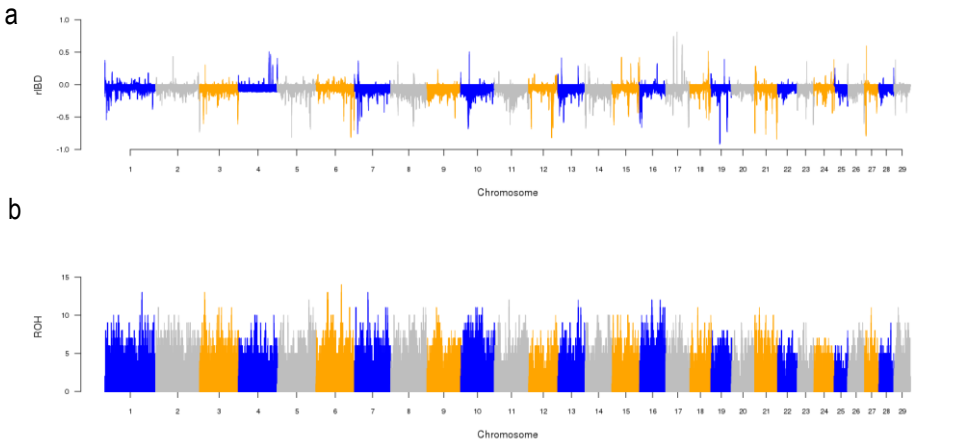


Figure 4.10 Comparison between Overlapped runs-of-homozygosity (ROH) among modern Red dairy cattle (mRDC) and relative identity-by-descend (rIBD) score showing introgression from Brown Swiss cattle (BSW) in mRDC (p>0.05 and correlation=0.004). a. genome-wide pattern of rIBS scores showing introgression from BSW in mRDC. b. sharing of ROH among mRDC.

4.4 Conclusion

Our study is the first to demonstrate genome-wide pattern of introgressed haplotypes in the modern dairy breed (mRDC) from high-yields breeds (HOL and BSW). We identified numerous regions that were likely introgressed from HOL or BSW, overlapping with important genes and QTLs affecting milk production, fat and protein content, calving traits, body confirmation, feed efficiency, carcass and fertility traits. These introgressed regions are correlated with the differentiation between mRDC and tRDC. Some of these regions containing genes and QTLs are of great importance for economic traits under selection in mRDC. The genomic footprints from introgression on the genome of hybrid mRDC are probably a result of interplay between artificial selection, recombination and drift. HOL or BSW introgressed haplotypes heavily shaped the genomic architecture of the hybrid mRDC as expected. Our study present the genomic regions which are introgressed and selected in the hybrid mRDC and contribute to the understanding of genomic consequence due to selective introgression in modern dairy cattle breed.

4.5 Appendix

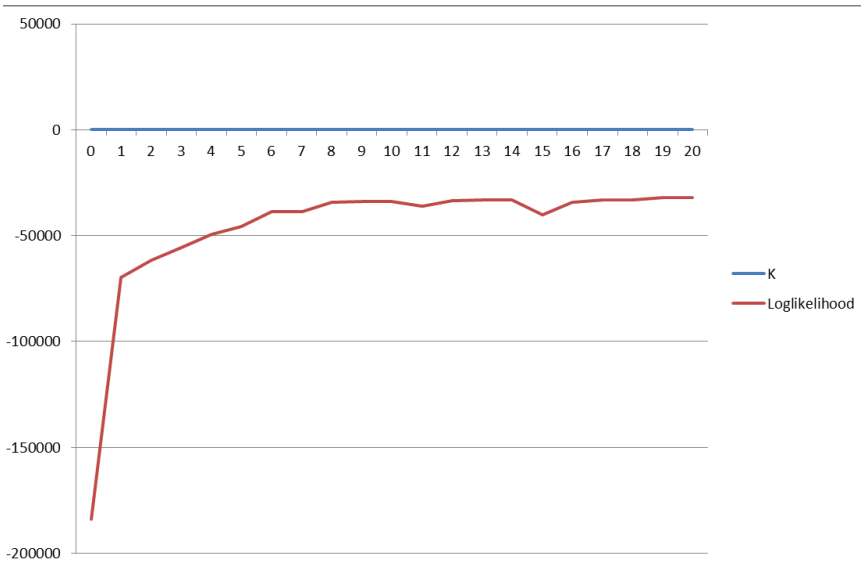


Figure S1 K values against Loglikelihood in Admixture analysis.

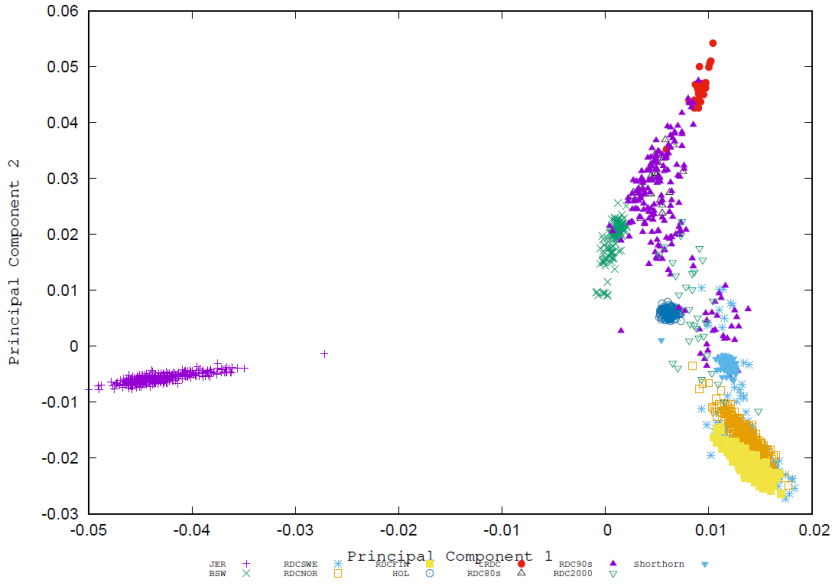


Figure S2 Principal component analysis (PCA) plots among different cattle breeds (Principal component 1 v.s. principal component 2). BSW – Brown Swiss, HOL – Holstein, tRDC – traditional Red Dairy Cattle, JER – Jersey cattle, RDCNOR – Norwegian Red Dairy Cattle, RDCSWE – Swedish Red Dairy Cattle, RDCFIN – Finish Red Dairy Cattle. Modern Red Dairy cattle include RDC90s and RDC2000 which are Red Dairy cattle born around 1990 and 2000.

4.6 Acknowledgement

We are grateful to Viking Genetics (Randers, Denmark) for providing samples for genotyping. Qianqian Zhang benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG". This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B). Mario Calus acknowledges financial support from the Dutch Ministry of Economic Affairs, Agriculture, and Innovation (Public-private partnership “Breed4Food” code BO-22.04-011-001-ASG-LR).

References

Abo-Ismail, M. K., M. J. Kelly, E. J. Squires, K. C. Swanson, S. Bauck, and S. P. Miller. 2013. Identification of single nucleotide polymorphisms in genes involved in digestive and metabolic processes associated with feed efficiency and performance traits in beef cattle. Journal of Animal Science 91(6):2512-2529.

Ai, H. S., X. D. Fang, B. Yang, Z. Y. Huang, H. Chen, L. K. Mao, F. Zhang, L. Zhang, L. L. Cui, W. M. He, J. Yang, X. M. Yao, L. S. Zhou, L. J. Han, J. Li, S. L. Sun, X. H. Xie, B. X. Lai, Y. Su, Y. Lu, H. Yang, T. Huang, W. J. Deng, R. Nielsen, J. Ren, and L. S. Huang. 2015. Adaptation and possible ancient interspecies introgression in pigs identified by whole-genome sequencing. Nature Genetics 47(3):217-255.

Benjamini, Y. and Y. Hochberg. 1995. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the royal statistical society. Series B (Methodological) 57(1):289-300.

Bennewitz, J., N. Reinsch, C. Grohs, H. Leveziel, A. Malafosse, H. Thomsen, N. Y. Xu, C. Looft, C. Kuhn, G. A. Brockmann, M. Schwerin, C. Weimann, S. Hiendleder, G. Erhardt, I. Medjugorac, I. Russ, M. Forster, B. Brenig, F. Reinhardt, R. Reents, G. Averdunk, J. Blumel, D. Boichard, and E. Kalm. 2003. Combined analysis of data from two granddaughter designs: A simple strategy for QTL confirmation and increasing experimental power in dairy cattle. Genetics Selection Evolution 35(3):319-338.

Bosse, M., M. S. Lopes, O. Madsen, H. J. Megens, R. P. M. A. Crooijmans, L. A. F. Frantz, B. Harlizius, J. W. M. Bastiaansen, and M. A. M. Groenen. 2015. Artificial selection on introduced Asian haplotypes shaped the genetic architecture in European commercial pigs. Proc. R. Soc. B. The Royal Society 282(1821): 20152019.

Bosse, M., H. J. Megens, L. A. F. Frantz, O. Madsen, G. Larson, Y. Paudel, N. Duijvesteijn, B. Harlizius, Y. Hagemeijer, R. P. M. A. Crooijmans, and M. A. M. Groenen. 2014a. Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. Nature Communications 5.

Bosse, M., H. J. Megens, O. Madsen, L. A. F. Frantz, Y. Paudel, R. P. M. A. Crooijmans, and M. A. M. Groenen. 2014b. Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent Sus scrofa populations. Molecular Ecology 23(16):4089-4102.

Browning, S. R. and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81(5):1084-1097.

Buzanskas, M. E., R. V. Ventura, T. C. S. Chud, P. A. Bernardes, D. J. D. Santos, L. C. D. Regitano, M. M. de Alencar, M. D. Mudadu, R. Zanella, M. V. G. B. da Silva, C. X. Li, F. S. Schenkel, and D. P. Munari. 2017. Study on the introgression of beef breeds in Canchim cattle using single nucleotide polymorphism markers. PLoS One 12(2): e0171660.

Capomaccio, S., M. Milanesi, L. Bomba, K. Cappelli, E. L. Nicolazzi, J. L. Williams, P. Ajmone-Marsan, and B. Stefanon. 2015. Searching new signals for production traits through gene-based association analysis in three Italian cattle breeds. Animal Genetics 46(4):361-370.

Cole, J. B., B. Waurich, M. Wensch-Dorendorf, D. M. Bickhart, and H. H. Swalve. 2014. A genome-wide association study of calf birth weight in Holstein cattle using single nucleotide polymorphisms and phenotypes predicted from auxiliary traits. Journal of Dairy Science 97(5):3156-3172.

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

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brondum, X. P. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46(8):858-865.

Davis, S. R., R. J. Spelman, and M. D. Littlejohn. 2017. BREEDING AND GENETICS SYMPOSIUM:Breeding heat tolerant dairy cattle: the case for introgression of the "slick" prolactin receptor variant into dairy breeds. Journal of Animal Science 95(4):1788-1800.

Fang, M., W. X. Fu, D. Jiang, Q. Zhang, D. X. Sun, X. D. Ding, and J. F. Liu. 2014. A Multiple-SNP Approach for Genome-Wide Association Study of Milk Production Traits in Chinese Holstein Cattle. PLoS One 9(8):e99544.

Fecteau, R. E., J. P. Kong, A. Kresak, W. Brock, Y. Song, H. Fujioka, R. Elston, J. E. Willis, J. P. Lynch, S. D. Markowitz, K. Guda, and A. Chak. 2016. Association Between Germline Mutation in VSIG1OL and Familial Barrett Neoplasia. Jama Oncol 2(10):1333-1339.

Fontanesi, L., D. G. Calo, G. Galimberti, R. Negrini, R. Marino, A. Nardone, P. Ajmone-Marsan, and V. Russo. 2014. A candidate gene association study for nine economically important traits in Italian Holstein cattle. Animal Genetics 45(4):576-580.

Galov, A., E. Fabbri, R. Caniglia, H. Arbanasic, S. Lapalombella, T. Florijancic, I. Boskovic, M. Galaverni, and E. Randi. 2015. First evidence of hybridization

between golden jackal (Canis aureus) and domestic dog (Canis familiaris) as revealed by genetic markers. Royal Society open science 2(12): 150450.

Gautier, M. and R. Vitalis. 2012. rehh: an R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28(8):1176-1177.

Hartwig, S., R. Wellmann, R. Emmerling, H. Hamann, and J. Bennewitz. 2015. Short communication: Importance of introgression for milk traits in the German Vorderwald and Hinterwald cattle. Journal of Dairy Science 98(3):2033-2038.

Hering, D. M., K. Olenski, and S. Kaminski. 2014. Genome-wide association study for poor sperm motility in Holstein-Friesian bulls. Animal Reproduction Science 146(3-4):89-97.

Hoglund, J. K., B. Buitenhuis, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015. Genome-wide association study for female fertility in Nordic Red cattle. BMC Genetics 16(1):10.

Jiao, X. L., B. T. Sherman, D. W. Huang, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. 2012. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics 28(13):1805-1806.

Kantanen, J., I. Olsaker, L. E. Holm, S. Lien, J. Vilkki, K. Brusgaard, E. Eythorsdottir, B. Danell, and S. Adalsteinsson. 2000. Genetic diversity and population structure of 20 North European cattle breeds. Journal of Heredity 91(6):446-457.

Kolbehdari, D., Z. Wang, J. R. Grant, B. Murdoch, A. Prasad, Z. Xiu, E. Marques, P. Stothard, and S. S. Moore. 2008. A whole-genome scan to map quantitative trait loci for conformation and functional traits in Canadian Holstein Bulls. Journal of Dairy Science 91(7):2844-2856.

Komohara, Y., K. Ohnishi, and M. Takeya. 2017. Possible functions of CD169-positive sinus macrophages in lymph nodes in anti-tumor immune responses. Cancer Science 108(3):290-295.

Kurland, M. D., M. B. Newcomer, Z. Peterlin, K. Ryan, S. Firestein, and V. S. Batista. 2010. Discrimination of Saturated Aldehydes by the Rat I7 Olfactory Receptor. Biochemistry-Us 49(30):6302-6304.

Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754-1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and G. P. D. Proc. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-2079.

Lindholm-Perry, A. K., A. K. Sexten, L. A. Kuehn, T. P. L. Smith, D. A. King, S. D. Shackelford, T. L. Wheeler, C. L. Ferrell, T. G. Jenkins, W. M. Snelling, and H. C. Freetly. 2011. Association, effects and validation of polymorphisms within the

NCAPG-LCORL locus located on BTA6 with feed intake, gain, meat and carcass traits in beef cattle. BMC Genetics 12(1):103.

Magee, D. A., K. M. Sikora, E. W. Berkowicz, D. P. Berry, D. J. Howard, M. P. Mullen, R. D. Evans, C. Spillane, and D. E. MacHugh. 2010. DNA sequence polymorphisms in a panel of eight candidate bovine imprinted genes and their association with performance traits in Irish Holstein-Friesian cattle. BMC Genetics 11(1):93.

Mao, X., N. K. Kadri, J. R. Thomasen, D. J. De Koning, G. Sahana, and B. Guldbrandtsen. 2015. Fine mapping of a calving QTL on Bos taurus autosome 18 in Holstein cattle. Journal of Animal Breeding and Genetics 133(3): 207-218.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297–1303.

Merks, J. W. M., P. K. Mathur, and E. F. Knol. 2012. New phenotypes for new breeding goals in pigs. Animal 6(4):535-543.

Moore, S. G., J. E. Pryce, B. J. Hayes, A. J. Chamberlain, K. E. Kemper, D. P. Berry, M. McCabe, P. Cormican, P. Lonergan, T. Fair, and S. T. Butler. 2016. Differentially Expressed Genes in Endometrium and Corpus Luteum of Holstein Cows Selected for High and Low Fertility Are Enriched for Sequence Variants Associated with Fertility. Biology of reproduction, 94(1), 19-1.

Ortega, M. S., A. C. Denicol, J. B. Cole, D. J. Null, and P. J. Hansen. 2016. Use of single nucleotide polymorphisms in candidate genes associated with daughter pregnancy rate for prediction of genetic merit for reproduction in Holstein cows. Animal Genetics 47(3):288-297.

Pimentel, E. C. G., S. Bauersachs, M. Tietze, H. Simianer, J. Tetens, G. Thaller, F. Reinhardt, E. Wolf, and S. Konig. 2011. Exploration of relationships between production and fertility traits in dairy cattle via association studies of SNPs within candidate genes derived by expression profiling. Animal Genetics 42(3):251-262.

Prakash, V., T. K. Bhattacharya, B. Jyotsana, and O. P. Pandey. 2011. Molecular Cloning, Characterization, Polymorphism, and Association Study of the Interleukin-2 Gene in Indian Crossbred Cattle. Biochemical genetics, 49(9-10), 638-644.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38(8):904-909.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155(2):945-959.

Sabater, L., C. Gaig, E. Gelpi, L. Bataller, J. Lewerenz, E. Torres-Vega, A. Contreras, B. Giometto, Y. Cornpta, C. Embid, I. Vilaseca, A. Iranzo, J. Santamaria, J. Dalmau, and F. Grays. 2014. A novel non-rapid-eye movement and rapid-eye-movement parasomnia with sleep breathing disorder associated with antibodies to lgLON5: a case series, characterisation of the antigen, and post-mortem study. Lancet Neurol 13(6):575-586.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419(6909):832-837.

Sahana, G., J. K. Hoglund, B. Guldbrandtsen, and M. S. Lund. 2015. Loci associated with adult stature also affect calf birth survival in cattle. BMC Genetics 16(1):1.

Sasago, N., T. Abe, H. Sakuma, T. Kojima, and Y. Uemoto. 2017. Genome-wide association study for carcass traits, fatty acid composition, chemical composition, sugar, and the effects of related candidate genes in Japanese Black cattle. Animal Science Journal 88(1):33-44.

Serao, N. V. L., D. Gonzalez-Pena, J. E. Beever, G. A. Bollero, B. R. Southey, D. B. Faulkner, and S. L. Rodriguez-Zas. 2013. Bivariate Genome-Wide Association Analysis of the Growth and Intake Components of Feed Efficiency. PLoS One 8(10): e78530.

vonHoldt, B. M., R. Kays, J. P. Pollinger, and R. K. Wayne. 2016. Admixture mapping identifies introgressed genomic regions in North American canids. Molecular Ecology 25(11):2443-2453.

Weir, B. S. and C. C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population-Structure. Evolution 38(6):1358-1370.

Weng, Z. Q., M. Saatchi, R. D. Schnabel, J. F. Taylor, and D. J. Garrick. 2014. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genetics Selection Evolution 46(1):34.

White, S. 2011. From globalized pig breeds to capitalist pigs: a study in animal cultures and evolutionary history. Environmental History, 16(1), 94-120.

Yates, A., W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G.

Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino, and P. Flicek. 2016. Ensembl 2016. Nucleic Acids Research 44(D1):D710-D716.

Zhang, Q., B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. 2015. Runs of homozygosity and distribution of functional variants in the cattle genome. BMC Genomics 16(1):542.

Zhang, Q. Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics 16(1):1.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10(4):R42.

5

Genome-wide association study for longevity with whole-genome sequencing in 3 cattle breeds

Qianqian Zhang[1,2], Bernt Guldbrandtsen[1], Jørn Rind Thomasen[3], Mogens S Lund[1] and Goutam Sahana[1]

[1] Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, Tjele DK-8830, Denmark; [2] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen 6700 AH, The Netherlands; 3 VikingGenetics, Assentoft, Randers DK-8960, Denmark

Journal of Dairy Science (2016) 99:7289–7298

Abstract

Longevity is an important economic trait in dairy production. Improvements in longevity could increase the average number of lactations per cow, thereby affecting the profitability of the dairy cattle industry. Improved longevity for cows reduces the replacement cost of stock and enables animals to achieve the highest production period. Moreover, longevity is an indirect indicator of animal welfare. Using whole-genome sequencing variants in 3 dairy cattle breeds, we carried out an association study and identified 7 genomic regions in Holstein and 5 regions in Red Dairy Cattle that were associated with longevity. Meta-analyses of 3 breeds revealed 2 significant genomic regions, located on chromosomes 6 (META-CHR6–88MB) and 18 (METACHR18–58MB). META-CHR6–88MB overlaps with 2 known genes: neuropeptide G-protein coupled receptor (*NPFFR2*; 89,052,210–89,059,348 bp) and vitamin D binding protein precursor (*GC*; 88,695,940–88,739,180 bp). The *NPFFR2* gene was previously identified as a candidate gene for mastitis resistance. META-CHR18–58MB overlaps with zinc finger protein 717 (*ZNF717*; 58,130,465–58,141,877 bp) and zinc finger protein 613 (*ZNF613*; 58,115,782–58,117,110 bp), which have been associated with calving difficulties. Information on longevity-associated genomic regions could be used to find causal genes/variants influencing longevity and exploited to improve the reliability of genomic prediction.

Key words: genome-wide association study, longevity, cattle, whole-genome sequencing

5.1 Background

Cows that remain healthy and reproduce regularly are of economic importance to breeders. The average lifespan of cows is far below their biological potential, and disposals due to old age are rare (Essl, 1998). In cows, longevity refers to the period from first parity until disposal, rather than the whole lifespan, although the length of productive life and herd life are more appropriate ways to describe the productive period (Essl, 1998). Longevity is of major importance for the economics of dairy production because involuntary culling of cows is often related to increased costs due to veterinary treatments, poor fertility, or low milk production. Selection for improved longevity is challenging because this trait is expressed late in life (the actual measurement is only available after the cow is culled or dead) and has a relatively low heritability (0.029–0.072; http://www.nordicebv.info/wp-content/uploads/2015/04/General-description_from-oldhomepage_06052015.pdf; Pritchard et al., 2013). Identification of genomic regions associated with longevity may help in identifying genes and causal mutations that influence this trait. This information could be included in the genomic selection model for better prediction of its breeding values (Brøndum et al., 2015, Wientjes et al., 2015).

Since 2008, longevity has been included as a trait in joint cattle breeding goals in Denmark, Finland, and Sweden. The Nordic Cattle Genetic Evaluation (NAV; www.nordicebv.info) defines the longevity index as the productive longevity of a bull's daughter, without correcting for other traits (e.g., yield or fertility). The economic value of longevity in the Nordic total merit index is 0.38 to 0.51 euros per day, depending on the breed. When this Nordic total merit index is used as a criterion in the selection strategy, 49 to 68% of the maximum progress for longevity can be obtained compared with a strategy that only includes longevity. The genetic trend for longevity has been positive for Holstein (HOL) and Nordic Red Dairy Cattle (RDC), but unchanged for Danish Jersey cattle (JER; https://www.landbrugsinfo.dk/Kvaeg/Avl/Avlsstatistik/Sider/aarsstat_2014.pdf?download = true).

Farmers and breeding companies in Nordic countries are taking increasing interest in breeding for longevity. Selection for longevity in cows could increase the number of productive cycles, with an indirect economic effect on the dairy cattle industry. Improved longevity for cows will reduce the replacement cost of stock, enable animals to achieve their highest production period, and improve animal welfare.

When the breeding goal in Nordic countries was changed to include both production and functional traits (e.g., fertility, health, and longevity), longevity was improved by about 21 and 18 breeding value units for RDC and HOL, respectively, between the years 1990 and 2010. However, this positive genetic progress for longevity has not been fully reflected in the phenotypic trend of Nordic cows. This discrepancy is because farm-related factors, such as feeding, housing, and management, have a large influence on cattle health (http://www.nordicebv.info/wp-content/uploads/2015/04/Longevity-of-Nordicdairy-cows-can-be-improved.pdf).

With the availability of genome-wide markers, genomic regions associated with longevity in cattle can be identified and studied. Therefore, the objective of this study was to identify QTL associated with longevity in 3 dairy cattle breeds. Within-breed association studies were carried out to detect longevity-associated QTL in 3 cattle breeds, followed by meta-analyses to determine whether additional QTL could be found by combining the breeds.

5.2 Material and Methods

Imputation of 50k genotypes to whole-genome sequencing (WGS) variants for bulls from 3 Nordic cattle breeds was done in 2 steps (Iso-Touru et al., 2016). High-density (HD) genotypes were imputed from 50k genotypes by using HD multibreed reference genotype data, and then the imputed HD genotypes were imputed to WGS variants by using a multibreed WGS reference. Genome-wide association analysis of the imputed WGS variants was also done in 2 steps. First, a genome scan was made separately for each of the 3 cattle breeds, using a modified mixed model approach (Kang et al., 2010). Then, a meta-analysis of the 3 breeds was performed.

5.2.1 Animals and Phenotypes

Because no animal experiments were performed, approval from the ethics committee was not required. The study included 10,575 genotyped and progeny-tested bulls from 3 Nordic cattle breeds from Denmark, Sweden, and Finland: 5,314 HOL, 4,200 RDC, and 1,061 Danish JER bulls. Population structure of these 3 breeds was reported by Kadri et al. (2015). Deregressed estimated breeding values (DRP) for the longevity trait from routine genetic evaluations were used as phenotypes in the association study. Average reliability of the DRP was 0.748.

In its evaluation of longevity, the NAV considers the number of days from calving to the end of lactation for each lactation up to the fifth, with a maximum of 365 d per lactation. To estimate breeding values of longevity, the NAV uses a multi-trait animal model, which includes age at first calving, calving month, and herd-year as fixed effects, heterosis effects as regression effects, and genetic groups, herd × year of first calving, and animal effects as random effects. Genetic groups are modeled as phantom parent grouping. Separate genetic evaluations are made for each breed (http://www.nordicebv.info/wp-content/uploads/2015/04/General-description_from-old-homepage_06052015.pdf). For details regarding recorded phenotypes and models used in routine breeding value estimates, see https://www.landbrugsinfo.dk/kvaeg/avl/sider/principles.pdf?download=true.

5.2.2 Genotypes of Animals

Single nucleotide polymorphism genotyping with a SNP chip and quality control (QC) analyses were performed as described by Sahana et al. (2015). In brief, DNA was extracted from whole blood or semen. All bulls with breeding values for longevity were genotyped with the BovineSNP50 BeadChip (Illumina, San Diego, CA) version 1 or 2. The QC analyses of SNP genotypes were carried out simultaneously for all 3 breeds. The SNP selection parameters were an 85% minimum call rate for individuals and 95% for loci. Marker loci without a known map position, with minor allele frequency <0.05%, or with deviation from Hardy-Weinberg proportions (χ^2-test, 1 df, $P < 0.00001$) were excluded. After QC, 43,415 SNP remained in the 50k data set.

A multi-breed reference of 3,383 animals (1,222 HOL, 1,326 RDC, and 835 Danish JER) obtained with the Illumina BovineHD Genotyping BeadChip was available in-house and from the EuroGenomics consortium (Lund et al., 2011). The QC parameter set and protocol for HD data were similar to those of the 50k chip described above. Imputation to HD genotypes was done with IMPUTE2 v2.3.1 (Howie et al., 2011). The HD genotyped reference was used to impute the 50k genotypes to the HD array as a bridge to impute to the WGS variants.

5.2.3 Imputation to WGS Level

Imputed HD genotypes were further imputed to the WGS level by using a multi-breed reference of 1,228 animals from run4 of the 1,000 bull genomes project (1,148 cattle, including 288 individuals from the global Holstein-Friesian population, 56 Nordic RDC individuals, 61 JER individuals, and 743 individuals from other breeds; Daetwyler et al., 2014) and private data from Aarhus University (80 cattle,

including 23 HOL, 30 Nordic RDC, and 27 Danish JER). Imputation to the WGS level was performed by using Minimac2 (Fuchsberger et al., 2015). A total of 22,751,039 bi-allelic variants were present in the imputed sequence data.

5.2.4 Statistical Methods for Association Analysis

Association analysis was carried out by using a modified linear mixed model (LMM) approach for within-breed QTL identification, followed by a meta-analysis across 3 breeds. The efficient mixed-model association expedited (EMMAX) approach was developed previously to correct for population structure and genetic relatedness, and to increase computational speed in association mapping (Kang et al., 2008). Association analysis for each imputed sequence variant was carried out in the EMMAX software tool by using a 2-step variance component-based approach to account for population stratification (Kang et al., 2010). Details about the model are given by Kang et al. (2008, 2010).

In the first step, polygenic and error variances were estimated with the variance component model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{a} + \boldsymbol{e}$$

where $\mathbf{y}$ is a vector of phenotypes (DRP for longevity); μ is the intercept; $\mathbf{a}$ is a vector of random polygenic effects, which are normally distributed as $\boldsymbol{a}$ ~ N $(0, \mathbf{G}\sigma_a^2)$; **G** is the genomic relationship matrix (GRM), which is built based on imputed HD SNP genotypes; σ_a^2 is the additive genetic variance; $\mathbf{Z}$ is an incidence matrix, relating phenotypes to the corresponding random polygenic effects; $\boldsymbol{e}$ is the vector of random individual error terms, assumed to follow N ~ $(0, \mathbf{I}\sigma_e^2)$; $\mathbf{I}$ is an identity matrix; and σ_e^2 is the error variance. The GRM is 2 times the kinship coefficient matrix between each pair of individuals. The kinship matrix is inferred by a simple identical-by-state allele-sharing matrix using “emmax-kin” in EMMAX. Inclusion of a candidate marker in the GRM can lead to loss of power due to double fitting of the candidate marker as both a fixed effect for testing association and a random effect as part of the GRM (Listgarten et al., 2013). Therefore, we followed the leave-one-chromosome-out approach (Yang et al., 2014) to build a kinship matrix specific to each chromosome.

In the second step, each SNP effect was obtained by using a linear regression model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{x}g + \boldsymbol{\eta}$$

where μ is the overall mean; $\mathbf{x}$ is a vector of imputed allele dosages (expected number of copies of a specified allele, ranging from 0 to 2); g is the SNP effect; and

$\boldsymbol{\eta}$ is a vector of random residual deviates, with variance $\mathbf{G}\sigma_a^2 + \mathbf{I}\sigma_e^2$. This approach was carried out separately for the WGS variants in each of the 3 dairy cattle breeds. A Bonferroni correction was applied to control for false positive associations.

The variance in the deregressed proofs of longevity explained by a SNP was calculated by $2p \times (1 - p) \times \alpha^2$/(variance of deregressed proofs of longevity) in which p is the frequency of the allele coded as 1 and α is the allele substitution effect.

5.2.5 Meta-Analysis Combining 3 Breeds

We carried out a sample size weighted meta-analysis based on P-values obtained from the EMMAX analysis for the 3 breeds. The weighted Z-score method was used, as follows:

$$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$$

Where $Z_i = \emptyset^{-1}(p_i/2) \times sign(\Delta_i)$; $w_i = \sqrt{N_i}$; N_i is the sample size; p_i is the P-value; Δ_i is the direction of effect for study *i*; and $\emptyset$ is the cumulative distribution function of the standard normal distribution. Significant SNP were detected by using the overall *P*-value for the meta-analysis, calculated as $P = 2\emptyset(-|Z|)$, and the Bonferroni-corrected threshold. Analyses were performed in the METAL program (Willer et al., 2010). This approach was carried out for WGS variants across the 3 Nordic dairy cattle breeds.

5.2.6 Search for Multiple QTL in a Genomic Region

We examined if multiple QTL are segregating in a genomic region on chromosome 6 and 18. First we adjusted the phenotype for the top associated SNP effect (Chr6:88840407 and Chr18:58015050 on chromosomes 6 and 18, respectively) using a **LMM** followed by a linear regression model to search for additional signal for association. The statistical model is described by the formula:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{x}g + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of deregressed proof breeding values for longevity, $\mathbf{1}$ is a vector of 1 s, μ is the general mean, g is the fixed additive genetic effect of the analyzed SNP, $\mathbf{x}$ is a vector of allele dosages for the top associated SNP of the chromosome, and $\mathbf{u}$ is a vector of random polygenic effects, which are normally distributed $\mathbf{u} \sim N(0, \mathbf{A}\sigma_u^2)$, where $\mathbf{A}$ is the pedigree-based additive relationship matrix, σ_u^2 is the polygenic variance, $\mathbf{Z}$ is an incidence matrix relating phenotypes to the corresponding random polygenic effects, and e is a vector of residual effects,

which are normally distributed $\mathbf{e} \sim N(0, \mathbf{D}\sigma_e^2)$, where $\mathbf{D}$ is a diagonal matrix with elements $d_{ii} = (1 - r_{DRP}^2)/r_{DRP}^2$ to account for heterogeneous residual variances due to different reliabilities of DRP (r_{DRP}^2), and σ_e^2 is the residual variance. Analyses were performed using the DMU package (Madsen et al., 2014). The residuals from LMM were used as response variable for the following linear regression model for single variant analysis.

$$\mathbf{y_r} = \mathbf{1}\mu + \mathbf{x}g + \boldsymbol{\varepsilon}$$

where $\mathbf{y_r}$ is the vector of residuals from the LMM, $\boldsymbol{\varepsilon}$ is random error assumed to be distributed $N \sim (0, \mathbf{I}\sigma_\varepsilon^2)$ and the rest of the terms in the model are as described above.

5.3 Results and discussion

5.3.1 Within-Breed Association Analyses

Within-breed genome scans for longevity-associated SNP revealed 8 genomic regions on 6 chromosomes in HOL and 5 regions on 4 chromosomes in Nordic RDC (Figures 5.1–5.3, Tables S1–S2; http://dx.doi.org/10.3168/jds.2015-10697). The 5 most significantly associated genomic regions and representative SNP in each region are presented in Tables S1 and S2 (http://dx.doi.org/10.3168/jds.2015-10697) for HOL and RDC. There were no significant SNP associated with longevity in JER (Figure 5.2).

HOL. A strong association signal in HOL was found on chromosome 6 around 88 Mbp (HOL-CHR6–88MB). The SNP with the strongest association (P = 5.36E-14) was located at 88,937,771 bp (rs383561794) (Figure 5.1 and Supplemental Table S1), between genes neuropeptide FF receptor 2 (*NPFFR2*; 89,052,210–89,059,348 bp) and groupspecific component (vitamin D-binding protein; *GC*; 88,695,940–88,739,180 bp). These genes were reported to be candidate genes associated with mastitis resistance (Sahana et al., 2014). In addition, these longevity QTL (HOL-CHR6–88MB and HOL-CHR6–89MB) also overlaps with the QTL not only associated with quality of udder and quality of feet and legs (Hiendleder et al., 2003, Pausch et al., 2016), but also some productive traits including fat and protein yield and percentage (Prinzenberg et al., 2003, Viitala et al., 2003, Raven et al., 2014). The HOL-CHR5–31MB overlaps with QTL associated with dressing percentage (Stone et al., 1999). The peak associated region around 88 Mbp on chromosome 9 (HOL-CHR9–88MB) was located on gene regulatory (inhibitor) subunit 14C (*PPP1R14C*;

88,384,683–88,500,749 bp) and overlaps with QTL associated with milk, protein, and fat yield (Zhang et al., 1998; Wiener et al., 2000).

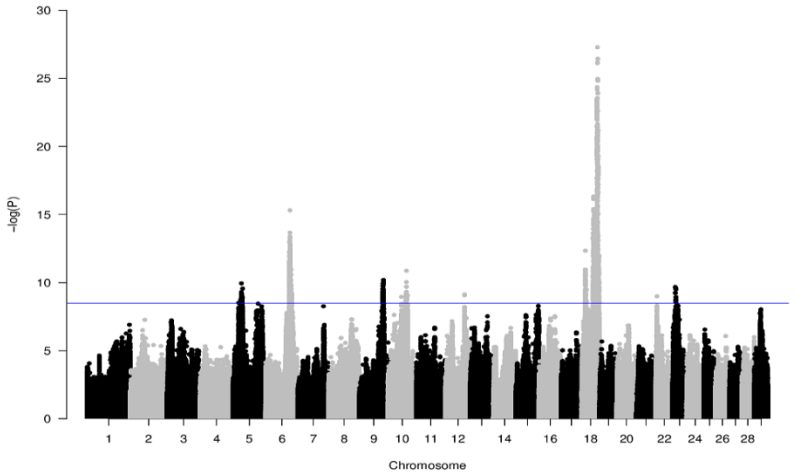


Figure 5.1 Manhattan plot for genome-wide association scan for longevity using whole-genome sequence variants in Holstein cattle.

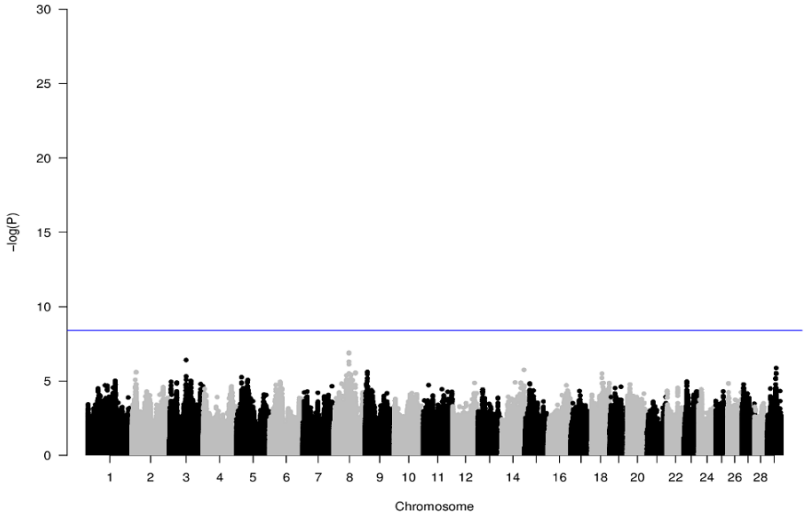


Figure 5.2 Manhattan plot for genome-wide association scan for longevity using whole-genome sequence variants in Jersey cattle.

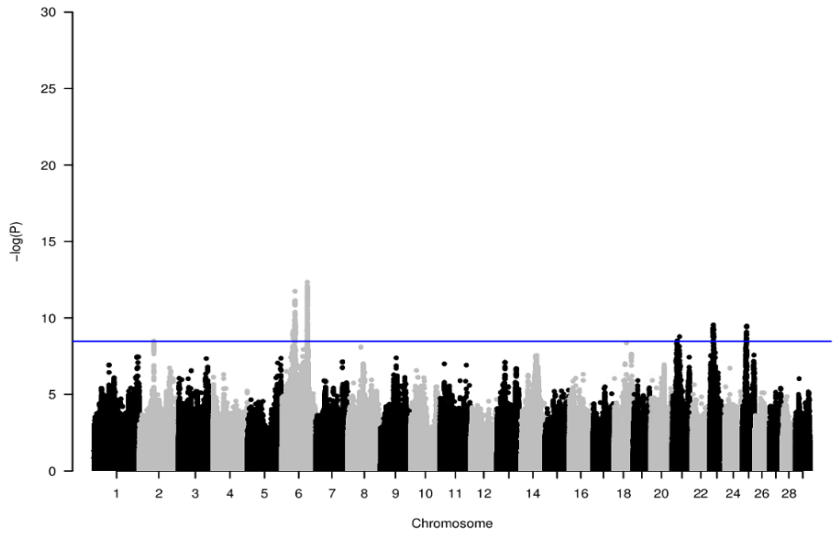


Figure 5.3 Manhattan plot for genome-wide association scan for longevity using whole-genome sequence variants in Nordic Red Dairy cattle.

The top associated SNP (P = 1.35E-11) in the genomic region around 67 Mbp on chromosome 10 (HOL-CHR10–67MB) was located at 67,632,847 bp (rs381727199), in close proximity to the GTP cyclohydrolase 1 gene (*GCH1*; 67,576,390–67,631,089 bp) and was overlapping with QTL associated with dressing percentage and ovulation rate (MacNeil and Grosz, 2002, Arias and Kirkpatrick, 2004). The HOL-CHR23-MB overlap with QTL associated with milk yield and preweaning ADG (Viitala et al., 2003, Kneeland et al., 2004). The strongest association signal in HOL was found on chromosome 18 around 58 Mbp (HOLCHR18–58MB). The most significant SNP on this region was located at 58,118,935 bp (rs521076153), very close to the zinc finger protein 613 gene (*ZNF613*; 58,115,782–58,117,110 bp) and zinc finger protein 717 gene (*ZNF717*; 58,130,465–58,141,877 bp), which have been associated with calving difficulties and which have been reported to be associated with calving traits (Mao et al., 2015). It is also very close to the QTL (BTA18: *SIGLEC12*) reported to be associated with productive life (Ma et al., 2012). This QTL on chromosome 18 also confirms the QTL reported previously for service-sire stillbirth and daughter calving ease (Cole et al., 2011). Meanwhile, this QTL (HOL-CHR18–58MB) overlaps with QTL associated with kidney, pelvic, and heart fat; live weight; and marbling score (MacNeil and Grosz, 2002). This SNP explained 3.45% of the variance in the de-regressed proofs of longevity (Supplemental Table S1).

RDC. In Nordic RDC, 2 significantly associated regions were present on chromosome 6. The most significant SNP on RDC-CHR6–88MB had the same locations as SNP on HOL-CHR6–88MB in HOL on chromosome 6 (Figure 5.3 and Supplemental Table S2). However, the QTL on chromosome 6 was not identified in the JER population, which is in line with previous findings for clinic mastitis traits (Sahana et al., 2014). Another region, RDC-CHR6–46MB at 46,099,648 bp (rs133661446) on chromosome 6, was located between 2 genes: coiled-coil domain containing 149 (*CCDC149*; 45,940,147–46,057,441 bp) and leucine-rich repeat LGI family, member 2 (*LGI2*; 46,127,372–46,155,777 bp). RDC-CHR6–46MB also overlaps with QTL associated with longissimus muscle area, hot carcass weight, birth and yearling weight (Casas et al., 2000). RDC-CHR21–19MB overlaps with QTL associated with birth weight (Kneeland et al., 2004). The most significant SNP on chromosome 23 in RDC-CHR23–13MB was located at 13,185,828 bp. This genomic region closely overlaps with the gene potassium channel, 2 pore domain subfamily K, member 16 (*KCNK16*; 13,175,069–13,181,327 bp) and QTL associated with milk yield and preweaning average daily gain (Viitala et al., 2003, Kneeland et al., 2004). The SNP on RDC-CHR6–88MB with highest significance explained 1.4% of the total variance in the deregressed proofs of longevity (Table 5.1). Mao et al. (2015) reported a major QTL affecting birth index in this HOL population, with the most associated SNP at 57,548,213 bp that explained 4.16% of the total genetic variance for birth index. It is possible that this QTL increases calving difficulties by increasing calf size at birth. Calving difficulties and related complications may lead to culling.

The genomic inflation (lambda) values for HOL, JER, RDC, and meta-analysis were 1.69, 1.30, 1.46, and 1.01, respectively (Supplemental Figures S1–S4). In the meta-analysis, the test-statistics were pre-adjusted for the inflation and therefore lambda was close to 1. The highest lambda value was observed for HOL in the within breed analyses. The trait analyzed here is polygenic in nature and phenotypes (DRP) had high reliability; besides, extensive linkage disequilibrium (LD) is present in these breeds due to small effective population sizes. All the markers in LD with the causal factor will show an effect on the trait proportional to their LD (r^2) with the causal variant. This is also evident from the fact that we see higher inflation in HOL where we have more power (larger sample sizes) and high LD. The extent of LD is high in HOL compared with the Red Dairy cattle, which are a combination of 3 sub-populations. Jerseys had the smallest sample size, and we observed the lowest inflation. Therefore, we can conclude that the inflation we see here is primarily due to the genetic architecture of the trait, sample sizes and

extent of LD. According to Yang et al. (2011), "In the absence of population structure and other technical artefacts, but in the presence of polygenic inheritance, substantial genomic inflation is expected. Its magnitude depends on sample size, heritability, LD structure and the number of causal variants." However, we used the leave-one-chromosome approach for building the relationship matrix. Therefore, we cannot exclude the possibility that some local structure remained unaccounted for, resulting in some inflation in test statistics.

5.3.2 Meta-Analysis

Meta-analyses of association results from the 3 cattle breeds did not reveal additional QTL for longevity. The 2 genomic regions on chromosomes 6 and 18 that were identified in within-breed analyses were significant in the meta-analyses (Figure 5.4). The top associated regions are reported in Table 5.1. The most significantly associated region from the meta-analysis was located on chromosome 6 (META-CHR6–88MB), and the SNP with the highest significance (Chr6:88840407: rs381295092) had an allelic substitution effect of 2.41 for the HOL population (Supplemental Tables S1 and S2). META-CHR6–88MB overlaps with 2 genes: neuropeptide G-protein coupled receptor gene (*NPFFR2*; 89,052,210–89,059,348 bp) and vitamin D-binding protein precursor (GC; 88,695,940–88,739,180 bp). This genomic region was reported to harbor mastitis QTL in HOL and RDC (Wu et al., 2015), suggesting that multiple causal variants for mastitis may be located in this region. The second most significantly associated region was META-CHR18–58MB, located very near the zinc finger protein 613 gene (*ZNF613*; 58,115,782–58,117,110 bp) and zinc finger protein 717 gene (*ZNF717*; 58,130,465–58,141,877 bp), which have been associated with calving difficulties.

We studied the LD pattern between the top associated SNP (Chr6:88840407, P = 4.23E-16) from the meta-analysis and other significant SNP at METACHR6–88MB. The most significant SNP was located upstream of *GC* and *NPFFR2*. Association results for significant SNP and the level of LD with the top SNP are presented in Supplemental Figure S5. Results of the search for additional QTL after adjusting phenotypes for the effect of Chr6:88840407 are presented in Supplemental Figure S6. Many of the highly significant SNP within this region were in high LD with Chr6:88840407. Several SNP were located around 88.88 Mbp with low LD with Chr6: 88840407 ($0.20 < r^2 \leq 0.40$), as well as around 89.05 Mbp. This result could indicate the presence of additional causal variant(s) in the latter region. However, no significantly associated SNP were observed when an association analysis was

Table 5.1 The significantly associated genomic regions with longevity in meta-analysis across three cattle breeds.

Region	Highest associated SNPs position (BP)	No. of significantly associated SNPs	Start and end positions of the region (BP)	Gene annotation	SNP Reference ID	Sample size	Z-Score	P-value
META-CHR6-88MB	88840407	2565	88025522 - 88995254	Intergenic variant	rs381295092	9571	-8.13	4.23E-16
	88899845			Intergenic variant	rs468282389	9571	8.08	6.73E-16
	88889152			Intergenic variant	rs109723567	9571	7.97	1.55E-15
	88892105			Intergenic variant	rs379544747	9571	-7.93	2.16E-15
	88889420			Intergenic variant	rs207890656	9571	7.93	2.18E-15
META-CHR18-58MB	58015050	596	58014529 - 58980943	Upstream gene variant	rs445560689	9571	-8.12	4.55E-16
	58141989			Downstream gene variant	rs381577268	9571	-8.04	9.35E-16
	58118935			Upstream gene variant	rs521076153	9571	7.97	1.53E-15
	58405555			Intergenic variant	rs479824187	5334	7.41	1.31E-13
	58410629			Intergenic variant	rs476912077	5334	-7.40	1.36E-13

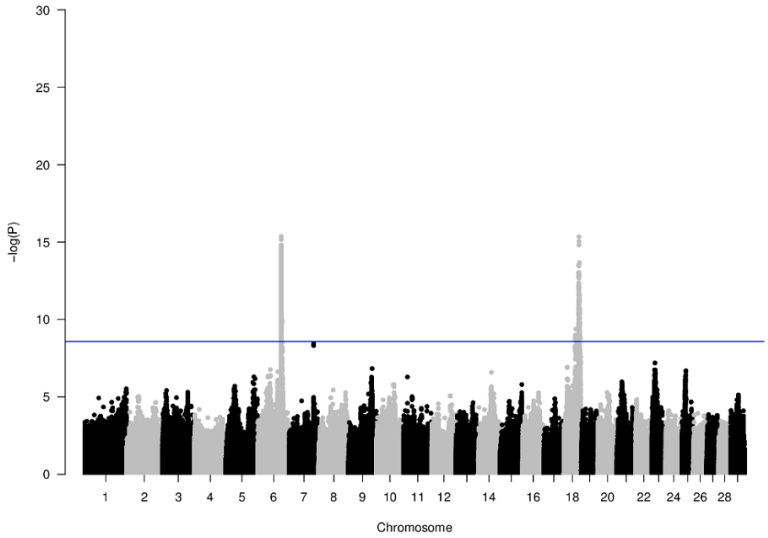


Figure 5.4 Manhattan plot for genome-wide association scan for longevity using whole-genome sequence variants in Nordic Red Dairy cattle.

carried out for residuals from a model that included the top associated marker (Chr6:88840407) and the polygenic component. The P-values of other SNP were under the threshold for significance (Supplemental Figure S6). Therefore, it is most likely that only one causal variant affecting longevity is segregating on chromosome 6 between 88 and 90 Mbp.

Analysis of LD between the second most significant region on chromosome 18 and the top associated SNP (Chr18:58015050: rs445560689, P = 4.55E-16) revealed SNP with high and low LD (Supplemental Figure S7). There might be multiple causal variants in this region. When the top associated marker (Chr18:58015050) and the polygenic component were included in an association analysis for residuals, the P-values of other SNP were below the threshold for significance (Supplemental Figure S8). Therefore, it is likely that only one causal variant affecting longevity is segregating on chromosome 18 between 57 and 59 Mbp.

5.4 General Discussion

Only the QTL on chromosome 6 overlapped in HOL and RDC in the within-breed analyses. Meta-analysis did not identify additional QTL. The RDC cattle consist of 3 RDC populations from Denmark, Sweden, and Finland. The population structure of breeds included in the present study has been reported (Kadri et al., 2015, Zhang et al., 2015a,b). Jersey is clearly separated and forms a tight cluster apart from the other populations. The admixture appears to be stronger among the 3 RDC subpopulations. Among the RDC breeds, the population from Denmark seems to be more closely related to HOL, as a result of the import of genes from Red Holstein in the 1970s to avoid inbreeding depression. It is possible that the major QTL segregating in these breeds are of recent origin and unique to the breeds. Some common QTL across these cattle breeds did not show significant association in within-breed analyses, perhaps due to a lack of power because of small effects on the trait or extremely low minor allele frequency. Genetic variants having major effects on longevity are segregating in these breeds, which could be the result of separation for many generations followed by strong artificial selection, as well as random drift due to small effective population sizes.

The QTL on chromosome 6 overlaps with mastitis resistance QTL in the same cattle populations (Sahana et al., 2014). The longevity trait could be influenced by mastitis, given the strong negative genetic correlation between mastitis resistance and risk of being culled (−0.40 to −0.53; Neerhof et al., 2000, Roxstrom and Strandberg, 2002). Multiple occurrences of mastitis may cause permanent damage to the udder, leading a cow to be culled from the herd. Alternatively, animals suffering from mastitis may have low immunity to other infections (Sordillo, 2005), which would increase their chances of being culled. The QTL on chromosome 18 overlaps with calf size and calving difficulty QTL in the same populations (Mao et al., 2015). A large size calf can cause damage of the birth canal of a cow during calving, which may lead to a higher risk of culling.

Overall, the most significantly associated SNP were not located within coding regions of genes or were between genes with unknown annotation. Thus, a link between cattle longevity and functions of the identified variants was not obvious. It was not possible to select one candidate gene, even for regions where the associated peaks were located within or near genes (e.g., META-CHR6–88MB and META-CHR18–58MB), due to LD among significant SNP from associated regions. Several other identified genomic regions associated with longevity harbor disease genes, including *KCNK16*, *PPP1R14C*, and *GCH1*. These disease genes and their

pathways could be studied further to short-list candidate genes for the longevity trait.

5.5 Appendix

Supplementary material can be found in the online version of the published paper or can be directly be accessed via
http://www.sciencedirect.com/science/article/pii/S0022030216303484

Zhang, Q., Guldbrandtsen, B., Thomasen, J. R., Lund, M. S., & Sahana, G. 2016. Genome-wide association study for longevity with whole-genome sequencing in 3 cattle breeds. Journal of Dairy Science, 99(9), 7289-7298.

3.7 Acknowledgement

We are grateful to the Danish Cattle Federation/NAV (Aarhus, Denmark) for providing the phenotypic data used in this study and Viking Genetics, Randers, Denmark, for proving samples for genotyping. Q. Zhang benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG." This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by The Danish Council for Strategic Research (Copenhagen, Denmark).

References

Arias, J., and B. Kirkpatrick. 2004. Mapping of bovine ovulation rate QTL; an analytical approach for three generation pedigrees. Animal Genetics 35:7–13.

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. Journal of Dairy Science 98:4107–4116.

Casas, E., S. D. Shackleford, J. W. Keele, R. T. Stone, S. M. Kappes, and M. Koohmaraie. 2000. Quantitative trait loci affecting growth and carcass composition of cattle segregating alternate forms of myostatin. Journal of Animal Science 78:560–569.

Cole, J. B., G. R. Wiggans, L. Ma, T. S. Sonstegard, T. J. Lawlor, B. A. Crooker, C. P. Van Tassell, J. Yang, S. W. Wang, L. K. Matukumalli, and Y. Da. 2011. Genome-

wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. BMC Genomics. 12:408.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brondum, X. P. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46:858–865.

Essl, A. 1998. Longevity in dairy cattle breeding: A review. Livestock Production Science 57:79–89.

Fuchsberger, C., G. R. Abecasis, and D. A. Hinds. 2015. minimac2: Faster genotype imputation. Bioinformatics 31:782–784.

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. Journal of Heredity 94:496–506.

Howie, B., J. Marchini, and M. Stephens. 2011. Genotype imputation with thousands of genomes. G3: Genes, Genomes. Genetics. 1:457–470.

Iso-Touru, T., G. Sahana, B. Guldbrandtsen, M. S. Lund, and J. Vilkki. 2016. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genetics 17:55.

Kadri, N. K., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015. Genetic dissection of milk yield traits and mastitis resistance quantitative trait loci on chromosome 20 in dairy cattle. Journal of Dairy Science 98:9015–9025.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42:348–354.

Kang, H. M., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723.

Kneeland, J., C. Li, J. Basarab, W. M. Snelling, B. Benkel, B. Murdoch, C. Hansen, and S. S. Moore. 2004. Identification and fine mapping of quantitative trait loci for growth traits on bovine chromosomes 2, 6, 14, 19, 21, and 23 within one commercial line of Bos taurus. Journal of Animal Science 82:3405–3414.

Listgarten, J., C. Lippert, E. Y. Kang, J. Xiang, C. M. Kadie, and D. Heckerman. 2013. A powerful and efficient set test for genetic markers that handles confounders. Bioinformatics 29:1526–1533.

Lund, M. S., A. P. W. de Roos, A. G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. T. Liu, R. Reents, C. Schrooten, F. Seefried, and G. S. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genetics Selection Evolution 43:43.

Ma, L., G. R. Wiggans, S. W. Wang, T. S. Sonstegard, J. Yang, B. A. Crooker, J. B. Cole, C. P. Van Tassell, T. J. Lawlor, and Y. Da. 2012. Effect of sample stratification on dairy GWAS results. BMC Genomics 13:536.

MacNeil, M. D., and M. D. Grosz. 2002. Genome-wide scans for QTL affecting carcass traits in Hereford x composite double backcross populations. Journal of Animal Science 80:2316–2324.

Madsen, P., J. Jensen, R. Labouriau, O. F. Christensen, and G. Sahana. 2014. DMU-A Package for Analyzing Multivariate Mixed Models in quantitative Genetics and Genomics. In Proc. 10th World Congress of Genetics Applied to Livestock Production, Vancouver, BC, Canada.

Mao, X., N. K. Kadri, J. R. Thomasen, D. J. De Koning, G. Sahana, and B. Guldbrandtsen. 2015. Fine mapping of a calving QTL on Bos taurus autosome 18 in Holstein cattle. Journal of Animal Breeding and Genetics 133:207–218. http://dx.doi.org/10.1111/jbg.12187.

Neerhof, H. J., P. Madsen, V. P. Ducrocq, A. R. Vollema, J. Jensen, and I. R. Korsgaard. 2000. Relationships between mastitis and functional longevity in Danish black and white dairy cattle estimated using survival analysis. Journal of Dairy Science 83:1064–1071.

Pausch, H., R. Emmerling, H. Schwarzenbacher, and R. Fries. 2016. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. Genetics Selection Evolution 48:14.

Prinzenberg, E. M., C. Weimann, H. Brandt, J. Bennewitz, E. Kalm, M. Schwerin, and G. Erhardt. 2003. Polymorphism of the bovine CSN1S1 promoter: Linkage mapping, intragenic haplotypes, and effects on milk production traits. Journal of Dairy Science 86:2696–2705.

Pritchard, T., M. Coffey, R. Mrode, and E. Wall. 2013. Understanding the genetics of survival in dairy cows. Journal of Dairy Science 96:3296–3309.

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

Roxstrom, A., and E. Strandberg. 2002. Genetic analysis of functional, fertility-, mastitis-, and production-determined length of productive life in Swedish dairy cattle. Livestock Production Science 74:125–135.

Sahana, G., B. Guldbrandtsen, B. Thomsen, L. E. Holm, F. Panitz, R. F. Brondum, C. Bendixen, and M. S. Lund. 2014. Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. Journal of Dairy Science 97:7258–7275.

Sahana, G., J. K. Hoglund, B. Guldbrandtsen, and M. S. Lund. 2015. Loci associated with adult stature also affect calf birth survival in cattle. BMC Genet. 16:47.

Sordillo, L. M. 2005. Factors affecting mammary gland immunity and mastitis susceptibility. Livestock Production Science 98:89–99.

Stone, R. T., J. W. Keele, S. D. Shackelford, S. M. Kappes, and M. Koohmaraie. 1999. A primary screen of the bovine genome for quantitative trait loci affecting carcass and growth traits. Journal of Animal Science 77:1379–1384.

Viitala, S. M., N. F. Schulman, D. J. de Koning, K. Elo, R. Kinos, A. Virta, J. Virta, A. Maki-Tanila, and J. H. Vilkki. 2003. Quantitative trait loci affecting milk production traits in Finnish Ayrshire dairy cattle. Journal of Dairy Science 86:1828–1836.

Wiener, P., I. Maclean, J. L. Williams, and J. A. Woolliams. 2000. Testing for the presence of previously identified QTL for milk production traits in new populations. Animal Genetics 31:385–395.

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

Willer, C. J., Y. Li, and G. R. Abecasis. 2010. METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26:2190–2191.

Wu, X., M. S. Lund, G. Sahana, B. Guldbrandtsen, D. X. Sun, Q. Zhang, and G. S. Su. 2015. Association analysis for udder health based on SNP-panel and sequence data in Danish Holsteins. Genetics Selection Evolution 47:50.

Yang, J., M. N. Weedon, S. Purcell, G. Lettre, K. Estrada, C. J. Willer, A. V. Smith, E. Ingelsson, J. R. O'Connell, M. Mangino, R. Mägi, P. A. Madden, A. C. Heath, D. R. Nyholt, N. G. Martin, G. W. Montgomery, T. M. Frayling, J. N. Hirschhorn, M. I. McCarthy, M. E. Goddard, and P. M. Visscher. 2011. Genomic inflation factors under polygenic inheritance. Europen Journal of Human Genetics 19:807–812.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price. 2014. Advantages and pitfalls in the application of mixed model association methods. Nature Genetics 46:100–106.

Zhang, Q., D. Boichard, I. Hoeschele, C. Ernst, A. Eggen, B. Murkve, M. Pfister-Genskow, L. A. Witte, F. E. Grignola, P. Uimari, G. Thaller, and M. D. Bishop. 1998.

Mapping quantitative trait loci for milk production and health of dairy cattle in a large outbred pedigree. Genetics 149:1959–1973.

Zhang, Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015a. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genetics 16:88.

Zhang, Q., B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. 2015b. Runs of homozygosity and distribution of functional variants in the cattle genome. BMC Genomics 16:542.

6

Comparison of gene-based rare variant association mapping methods for quantitative traits in a bovine population with complex familial relationships

Qianqian Zhang[1,2], Bernt Guldbrandtsen[1], Mario Calus[2], Mogens S Lund[1] and Goutam Sahana[1]

[1] Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, Tjele DK-8830, Denmark; [2] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen 6700 AH, The Netherlands.

Genetics Selection Evolution (2016) 48:60

Abstract

Background: There is growing interest in the role of rare variants in the variation of complex traits due to increasing evidence that rare variants are associated with quantitative traits. However, association methods that are commonly used for mapping common variants are not effective to map rare variants. Besides, livestock populations have large half-sib families and the occurrence of rare variants may be confounded with family structure, which makes it difficult to disentangle their effects from family mean effects. We compared the power of methods that are commonly applied in human genetics to map rare variants in cattle using whole-genome sequence data and simulated phenotypes. We also studied the power of mapping rare variants using linear mixed models (LMM), which are the method of choice to account for both family relationships and population structure in cattle.

Results: We observed that the power of the LMM approach was low for mapping a rare variant (defined as those that have frequencies lower than 0.01) with a moderate effect (5 to 8 % of phenotypic variance explained by multiple rare variants that vary from 5 to 21 in number) contributing to a QTL with a sample size of 1000. In contrast, across the scenarios studied, statistical methods that are specialized for mapping rare variants increased power regardless of whether multiple rare variants or a single rare variant underlie a QTL. Different methods for combining rare variants in the test single nucleotide polymorphism set resulted in similar power irrespective of the proportion of total genetic variance explained by the QTL. However, when the QTL variance is very small (only 0.1 % of the total genetic variance), these specialized methods for mapping rare variants and LMM generally had no power to map the variants within a gene with sample sizes of 1000 or 5000.

Conclusions: We observed that the methods that combine multiple rare variants within a gene into a meta-variant generally had greater power to map rare variants compared to LMM. Therefore, it is recommended to use rare variant association mapping methods to map rare genetic variants that affect quantitative traits in livestock, such as bovine populations.

Key words: Runs of homozygosity, Polymorphisms, Inbreeding, Cattle, Genome sequencing

6.1 Background

Genome-wide association studies (GWAS) have been successful in identifying common variants that are associated with complex diseases and quantitative traits. However, the common variants that have been identified thus far account for only a small fraction of the estimated heritabilities (Maher, 2008, Manolio et al., 2009, Gibson, 2012). Theoretical and empirical studies suggest that rare variants (defined as those that have frequencies lower than 0.01), could play a significant role in quantitative trait variation (Gibson, 2012, Kemper et al., 2012). In addition, studies on several Mendelian diseases indicate that common variants may often have a key role as modifiers of the effects of rarer, more highly penetrant contributors to disease risk in humans (Steinberg and Adewoye, 2006, Thein and Menzel, 2009). Therefore, the detection and investigation of rare variants should help researchers to further understand the genetic architecture of quantitative traits and may provide new ways to use such rare variants for mapping genes and improving accuracies of genomic prediction.

Rare variants are poorly captured by the commonly used single nucleotide polymorphism (SNP) chips, because SNPs on these chips typically have a much higher minor allele frequency (MAF) than rare variants and, thus, are generally in low linkage disequilibrium (LD) with these. Recent technological advances allow us to study individual genomes at the base-pair resolution (Elsik et al., 2016), including the detection of rare variants. Based on a large number of sequenced individuals (e.g. 1000), the optimal sequencing depth required for variants with a frequency lower than 0.01 is ~27 (Cao et al., 2013). Therefore, low coverage sequencing yields low calling accuracy at rare variant sites, and in addition, deep sequencing a large number of individuals remains economically prohibitive. The alternative is to impute high-density SNPs to whole-genome sequence. However, compared to common variants, rare variants are more often private to a sub-population or to families within a population (Tennessen et al., 2012), and thus, imputation accuracy for rare variants is considerably lower than for common variants. It has been shown in cattle that imputation accuracy from lower density SNP panels to whole-genome sequence data drops very quickly when allele frequency is lower than 0.1 (Bouwman and Veerkamp, 2014, Brondum et al., 2014, van Binsbergen et al., 2014). Although some imputation algorithms, such as that implemented in the IMPUTE2 software, tend to achieve higher imputation accuracies for rare variants than other algorithms, imputation accuracy remains rather low (Brondum et al.,

2014). Thus, imputation of rare variants remains a challenge, and is currently not sufficiently accurate to study the power of gene-based rare variant mapping. Instead deep re-sequencing of a large number of individuals (i.e. at least 1000) is necessary to identify rare variants, but currently this is economically prohibitive although the cost of whole-genome sequencing is continuously decreasing. An alternative approach to study the power to detect rare variants is to carry out a simulation study. Besides, simulated data has the advantage that the causal variants and their simulated effects are known with certainty and therefore, it is possible to compare methods for their accuracies of estimated effects. It is important that the genetic variation of the simulated dataset represents the complete spectrum of allele frequencies and retains the same haplotype structure as the empirical data (Moutsianas et al., 2015). Therefore, we used imputed sequence variants for a large number of SNP-array genotyped individuals to compare gene-based rare variant mapping approaches in cattle. Since the phenotypes were simulated based on imputed sequence variants, imputation errors did not distort the individuals' phenotypes.

Methods for GWAS based on common SNP variants are well established (Gibson, 2012). However, mapping rare variants remains a challenge and rare-variant association studies are generally "gene-based", in the sense that rare variants that are located within the same gene are grouped and then statistical methods are applied to assess the significance of the association between the phenotype and the combined rare variants. Cirulli (2016) emphasized the increasing importance of gene-based analyses in a review of 150 exome sequencing studies that claim that a disease can be caused by different rare variants in the same gene. Recently, guidelines on how to combine rare variants in gene-based analyses were formulated by MacArthur et al. (2014).

Several classes of statistical methods have been developed for the analysis of rare variants for 'case–control' designs and quantitative traits in humans for both randomly sampled and related individuals (Madsen and Browning, 2009, Price et al., 2010, Neale et al., 2011, Wu et al., 2011). A short overview of the approaches is given below.

One broad class of such methods is known as the "burden test" (Morgenthaler and Thilly, 2007, Li and Leal, 2008, Madsen and Browning, 2009, Price et al., 2010). A burden test collapses multiple rare variants in a region of the genome into a single meta-allele to represent a genetic burden score. These meta-alleles are then used

in association analyses. The power of these burden tests depends on the effect of the pooled variants and assumes that the effects of the rare alleles at different variant sites in a region of interest are in the same direction. Recent developments around these burden tests have enabled the analysis of data on related individuals (Chen et al., 2013, Schaid et al., 2013).

The second broad class of methods comprises variance component tests, such as that implemented in the C-alpha (Neale et al., 2011) and sequence kernel association test (SKAT) (Wu et al., 2011). Variance component tests aggregate individual variant statistics that measure the similarity of the variants within a region and incorporate flexible weights to boost the power of the analysis. Compared to the burden test, variance component tests are more robust for the identification of a gene even when multiple rare variants within the targeted gene have effects in different directions (positive and negative). There are also extensions for this kind of method for related individuals such as that implemented in famSKAT (Chen et al., 2013) and other similar approaches (Schifano et al., 2012, Ionita-Laza et al., 2013, Schaid et al., 2013).

The third category of methods combines burden tests and variance component tests to exploit the strengths of both approaches. This is implemented in the software SKAT-O for unrelated individuals (Lee et al., 2012) and in MONSTER (minimum *p* value optimized nuisance parameter score test extended to relatives) for related individuals (Jiang and McPeek, 2014). These methods introduce a nuisance parameter that defines the trade-off between burden tests and variance component tests, and is adaptively determined from the data to optimize power. Therefore, the combination of these two tests will be optimally balanced by the data itself and can detect both the common effect across rare variants (as in the burden tests) and the individual deviations from the average effect (as in the variance component tests).

Several studies have mapped rare variants that contribute to complex diseases in humans by using deep exome sequencing (Tennessen et al., 2012, Casals et al., 2013, Lee et al., 2014, MacArthur et al., 2014). However, to date, association studies for rare variants in cattle and other livestock species have not been reported. Increasing access to a large number of whole-genome sequences (Daetwyler et al., 2014) and availability of exome sequence data in the near future could be used to map rare variants in cattle. This will open new opportunities to

capture rare variants that affect economic traits in cattle especially those that are related to disease susceptibility, which, so far, was not possible by using SNP chip data. This should substantially improve success both in finding causative mutations and using the information for genomic selection to improve accuracy of prediction. Once the causative mutations are identified for one population, they can be directly tested in other populations and thus, results may be transposable from one breed to another.

The above-mentioned methods for rare variant association mapping were developed for human studies for which samples are obtained at random from a population or data that originate from small families, e.g. trio and sib-pair analyses. In contrast, bovine datasets usually include large half-sib families, and intensive artificial selection in cattle may pose special issues that are related to data analysis. For example, rare variants may be confounded with family structure, making it more difficult to disentangle their effects from family mean effects. In addition, the availability of large half-sib family sizes in cattle has the advantage that rare variants may be observed at a higher frequency within extended families compared to the population as a whole. The suitability of the above-described statistical methods that were developed to map rare variants for quantitative traits in humans still remains unexplored for data structures such as those of cattle and other livestock species. Thus, the objective of our study was to investigate power and type I errors of several approaches used to map rare variants in bovine data. Our hypothesis is that the power of the specialized methods that were developed to detect rare variants in the human genome will be higher than that of a linear mixed model approach, which is currently the method of choice to map common variants in the bovine genome. Thus, we propose method(s) for rare variant mapping in livestock populations, which should contribute to the development of models that are geared towards exploiting rare variants in genome-assisted breeding.

6.2 Methods

6.2.1 Statistical methods

6.2.1.1 Statistical methods for rare variant mapping

The statistical methods that we tested for rare variant mapping were famBT (Chen et al., 2013), famSKAT (Chen et al., 2013) and MONSTER (Jiang and McPeek, 2014).

The famBT method is a burden test that accounts for family relationships and assumes that the effects of all the rare variants are in the same direction (Chen et al., 2013) while the family-based SKAT (famSKAT) method makes no assumption on the direction of the effects of rare variants (Chen et al., 2013). The MONSTER method adaptively determines a nuisance parameter to adjust to the unknown composition of the effects at rare variant sites by applying a mixed effects model that accounts for covariates and additive polygenic effects (Jiang and McPeek, 2014).

When written in more conventional animal breeding notation, the MONSTER model becomes:

$$\mathbf{y} = \mathbf{X\gamma} + \mathbf{M\beta} + \mathbf{Zu} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of phenotypes, X is a design matrix for fixed covariates including the intercept, $\boldsymbol{\gamma}$ is a vector of unknown covariate effects, $\mathbf{Z}$ is an incidence matrix relating phenotypes to the corresponding random polygenic effect, $\mathbf{u}$ is a vector of random polygenic effects that follows a multivariate normal distribution N ~ (0, $\mathbf{A}\sigma_a^2$), where $\mathbf{A}$ is the additive genetic relationship matrix and σ_a^2 is the polygenic variance, e is a vector of random residuals, $\mathbf{e} \sim$ N(0, $\mathbf{I}\sigma_e^2$), $\mathbf{M}$ is a n × m matrix that encodes the genotype at the m tested variant loci and n is the number of individuals with m_{ij} representing the allele dosage (0, 1 or 2) of the minor allele at the j-th variant of individual i, and $\boldsymbol{\beta}$ is a vector of (possibly correlated) random effects of the m variants, $\boldsymbol{\beta} \sim N\left(0, \mathbf{R}_\rho \sigma_q^2\right)$, $\mathbf{R}_\rho = (1-\rho)\mathbf{I} + \rho\mathbf{I}$ with $0 \leq \rho \leq 1$. The limiting cases ρ = 0 and ρ = 1 correspond to models famSKAT and famBT, respectively. This method for detecting rare variants is referred to as MONSTER (Jiang and McPeek, 2014). A grid of 11 equally-spaced points: values of ρ i.e. ρ_1 = 0, 2 = 0.1, . . . , ρ_{10} = 0.9, ρ_{11} = 1 were tested in MONSTER. When ρ = 0, MONSTER is equivalent to famSKAT and when ρ = 1, MONSTER is equivalent to famBT.

To detect associations between a trait and a genomic region of interest, we tested the null hypothesis H_0 that σ_q^2 = 0 against H_1 that σ_q^2 > 0. This analysis was done using the software MONSTER (Jiang and McPeek, 2014). To access the type I error rate, the null model was tested for 1000 replicates for which the effects for all rare variants were assumed to be equal to 0. The genomic control coefficient λ (Devlin and Roeder, 1999), which for test statistics measures the departure of the median p value from its expectation under the null hypothesis, was calculated for all statistical methods considered to detect rare variants.

6.2.1.2 Statistical methods for GWAS with common variants

We compared MONSTER, famBT and famSKAT to two methods that are used for association mapping of common variants: a linear mixed model (Yu et al., 2006) and a simplified linear mixed model as implemented in the EMMAX software (Kang et al., 2010). These methods were included to investigate their ability to map rare variants and are briefly described below.

The linear mixed model (LMM) carries out a SNP-by-SNP analysis. Complex familial relationships are the primary confounding factor in GWAS of livestock populations. In cattle, LMM, which model the effects of relationships among individuals through polygenic effects, can control the false positive rate caused by family structure and population stratification (Sahana et al., 2010, Kadri et al., 2014). Here for the LMM, association between a SNP and a phenotype was assessed by a single-locus regression analysis using the following equation:

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

where $\mathbf{y}$ is the vector of phenotypes, $\mathbf{1}$ is a vector of ones, μ is the general mean, m is a vector of allele dosages (ranging from 0 to 2) that associate records to the marker effect, g is the scalar additive effect of the SNP, $\mathbf{Z}$ is an incidence matrix relating phenotypes to the corresponding random polygenic effect, $\mathbf{u}$ is a vector of random polygenic effects that follows a multivariate normal distribution N ~ (0, $\mathbf{A}\sigma_a^2$), where $\mathbf{A}$ is an additive relationship matrix and σ_a^2 is the polygenic variance, and e is a vector of random environmental deviates that follows a normal distribution N ~ (0, $\mathbf{I}\sigma_e^2$), where σ_e^2 is the error variance and $\mathbf{I}$ is an identity matrix. The model was fitted by restricted maximum likelihood (REML) using the software DMU (Madsen and Jensen, 2006), and the null hypothesis H_0 that g = 0 was assessed using a t-test. The null hypothesis was tested with 1000 replicates and the results are presented as the null model. The genomic control coefficient (Devlin and Roeder, 1999) was used to correct for stratification by adjusting association statistics at each SNP by the overall inflation factor (λ). A SNP was considered to be significantly associated with a trait if the p-value was below a significance threshold after correction for multiple-testing. We used two different multiple-testing correction approaches that are described in section “Comparison of different methods used to map rare variants in the simulation”.

Single variant association analysis using a LMM for full sequence variants is computationally demanding, i.e. it requires a computation time of $O(MN^3)$, where M is the number of SNPs and N is the number of samples, since variance component estimation is repeated for each candidate SNP (Yang et al., 2014).

Therefore, association analysis for each imputed sequence variant was also carried out using the efficient mixed-model association (EMMA) approach where the variance components are estimated once instead of for each variant using the EMMAX software (Kang et al., 2010). Briefly, the polygenic and error variances are estimated using the following variance component model: $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e}$, where $Var(y) = \mathbf{G}\sigma_a^2 + \mathbf{I}\sigma_e^2$, μ is the intercept, $\mathbf{y}$ is the vector of phenotypes, $\mathbf{G}$ is the genomic relationship matrix that is built based on high-density (HD) SNP genotypes, $\mathbf{I}$ is an identity matrix, σ_a^2 is the additive genetic variance and σ_e^2 is the error variance. In a second step, the SNP effect is obtained using a generalized linear regression model model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{m}g + \boldsymbol{\eta},$$

where $\mathbf{m}$ is a vector of the imputed allele dosages (ranging from 0 to 2), and $\boldsymbol{\eta}$ is a vector of random residual deviates with variance $\mathbf{G}\sigma_a^2 + \mathbf{I}\sigma_e^2$.

6.2.2 Individual genotypes and simulation of phenotypes

In total, the genotypes of 27,119 Holsteins animals were available for this study from the Illumina 54 k SNP array version 1 or 2 (Illumina Inc., San Diego). The number of SNPs remaining after quality control was equal to 43,415, for more details see (Iso-Touru et al., 2016). However, due to the computational constraints, we limited the analysis by including only the genes on chromosome 10 (arbitrarily picked) and for 5000 randomly selected individuals. The positions of the SNPs on the bovine genome were taken from the UMD3.1 Bovine genome assembly (Zimin et al., 2009). The 54 k genotypes for chromosome 10 of the 5000 randomly sampled animals together with 22,119 other animals were imputed to whole-genome sequence data using a two-step approach with the IMPUTE2 software (Howie et al., 2009). Average kinship between the sampled bulls was equal to 0.0017 and the 5000 sampled bulls were sired by 632 bulls that had 1–136 sons in the dataset, with a mean value of 7.9. The heat map of the relationships between the 5000 sampled individuals is in Additional file 1: Figure S1.

Approaches for rare variant mapping are gene-based, thus the results cannot be easily averaged across multiple genes. Therefore, based on the number of rare variants and level of LD between variant sites, two genes on bovine chromosome 10 were selected for this study: (1) the ENSEMBL Gene ID: *ENSBTAG00000018852* located between 1,116,669 and 1,212,429 bp that comprised 635 annotated SNPs in its transcribed region; the 222 rare variants (with a MAF < 0.01) within this gene were grouped into different SNP sets according to their MAF; the average pairwise

LD (r^2) for these variants was equal to 0.149 with an average distance of 83 bp between variants; and (2) the ENSEMBL Gene ID: *ENSBTAG00000035858* located between 610,854 and 933,224 bp that included 3015 annotated SNPs and 309 rare variants (MAF < 0.01); the average pairwise LD (r^2) for these variants was equal to 0.74 with an average distance of 106 bp between variants.

Phenotypes were simulated as the sum of three components, i.e. a polygenic effect, a QTL effect computed as the sum of the simulated effects of the underlying rare variants, and a random error. The polygenic effects were simulated based on pedigree records. The effects of rare variants were simulated as random effects. Four scenarios with respect to the MAF of the causal variants were considered. Rare variants were grouped into four classes based on MAF for the sampled individuals i.e.: 0.01 ≤ MAF < 0.02; 0.005 ≤ MAF < 0.01; 0.001 ≤ MAF < 0.005; and MAF < 0.001. Two different approaches for assigning effects to rare variants were followed: within a MAF class either multiple rare variants contributed to the total QTL effect or only one rare variant contributed to the whole QTL variance. In the scenarios with multiple causal variants, half of the rare variants within each MAF class were assigned an effect.

Three levels of heritability for the trait were considered i.e. 0.3, 0.5 and 0.8. In addition, three levels of QTL variances were considered. The variance explained by the QTL (i.e. the collective effect of the causal rare variants within the gene) was equal to 0.1, 0.5 or 1 % of the total genetic variance when the heritability was equal to 0.5. In the scenarios with multiple causal variants, the sum of the variance explained by individual causal variants was set equal to the predefined total QTL variance. The QTL effect (α) was then calculated by the following equation and each causal rare variant was assigned an effect with a certain weight (as defined next):

$$\alpha^2 = \frac{V_{qtl}}{Var(\mathbf{M})},$$

where V_{qtl} is the proportion of genetic variance explained by the QTL multiplied by the total genetic variance (i.e. 0.1, 0.5 and 1 %), **M** is the genotype dosage matrix including the loci which have a QTL effect. The weights were assigned in order to add QTL effects on the simulated phenotypes. Note that this formula considers the genotype variance at the QTL, as well as the co-variance between the QTL, and that it yields one value for α that is used for all QTL. Therefore, the total QTL effect for each animal was calculated as: α**M**.

In addition to the QTL effects, an additive polygenic effect was simulated with a variance component proportional to the kinship matrix. The polygenic effects were sampled from the following normal distribution, proceeding from the oldest to the youngest animal:

$$\text{Founder: } a^{F} \sim N(0,1),$$

$$\text{Offspring with one parent known: } a^{O1} \sim N\left(\frac{a}{2}, \left(\frac{3}{4} - \frac{F}{4}\right)\sigma_a^2\right),$$

$$\text{Offspring with two known parents: } a^{O2} \sim N\left(\frac{a_s + a_d}{2}, \left(\frac{1}{4}(1 - F_s) + \frac{1}{4}(1 - F_d)\right)\sigma_a^2\right),$$

where a^F is the polygenic effect for a founder, i.e. an animal with both parents unknown, and a^{O1} or a^{O2} are the polygenic effects for animals with one or two known parents, respectively, a, a_s and a_d are the polygenic effects for the known parent, the sire and the dam, respectively, F, F_s and F_d are the inbreeding coefficients for the known parent, the sire and dam, respectively, and σ_a^2 is the additive genetic variance.

Finally, an independent error variance component was also simulated to account for measurement error and individual-specific variability e ~ N(0, $\mathbf{I}\sigma_e^2$), where σ_e^2 is the error variance, which is equal to 20, 50 or 70 % of the phenotypic variance.

6.2.3 Simulated scenarios

A scenario with a sample size of 1000 individuals, a heritability of 0.5 and a QTL that explained 1 % of the total additive genetic variance was considered as the base scenario and used for comparison with the other scenarios (Table 6.1). Four MAF classes of rare variants (0.01 ≤ MAF < 0.02; 0.005 ≤ MAF < 0.01; 0.001 ≤ MAF < 0.005; and MAF < 0.001) based on MAF calculated from the whole population (27,119 Holsteins animals) were considered as causal variants for each heritability and QTL variance scenario. Two additional heritability levels (0.3 and 0.8) were simulated to compare with the heritability of the base scenario (h^2 = 0.5). Different proportions (0.1, 0.5 and 1 %) of additive genetic variance explained by the QTL were compared for the scenario with MAF class 0.001 ≤ MAF < 0.005. For low QTL variance scenarios (0.1 and 0.5 %), two sample sizes of 1000 and 5000 randomly selected individuals were compared. One hundred replicates were simulated for each scenario.

Table 6.1 Scenarios used in the simulation.

Heritability	MAF	Proportion of additive genetic variance explained by the QTL	Sample size in the test
0.3	0.01 ≤ MAF < 0.02 0.005 ≤ MAF < 0.01 0.001 ≤ MAF < 0.005 MAF < 0.001	0.01	1000
0.5	0.01 ≤ MAF < 0.02 0.005 ≤ MAF < 0.01 0.001 ≤ MAF < 0.005 MAF < 0.001	0.01	1000
0.8	0.01 ≤ MAF < 0.02 0.005 ≤ MAF < 0.01 0.001 ≤ MAF < 0.005 MAF < 0.001	0.01	1000
0.5	0.001 ≤ MAF < 0.005	0.001; 0.005 or 0.01	1000
0.5	0.001 ≤ MAF < 0.005	0.001	1000; 5000
0.5	0.001 ≤ MAF < 0.005	0.005	1000; 5000

6.2.4 Comparison of methods used to map rare variants in the simulation

To analyze samples of related individuals, three rare variant mapping methods famBT (Chen et al., 2013), famSKAT (Chen et al., 2013) and a combination of these two methods (MONSTER) (Jiang and McPeek, 2014) were compared. In addition, linear mixed model approaches as implemented by EMMAX (Kang et al., 2010) and DMU (Madsen and Jensen, 2006) were used. The kinship matrix used for LMM in DMU was based on the pedigree-based matrix (DMU_AMAT) while for EMMAX both a pedigree-based and a genomic relationship matrix using 50 k genotypes of the individuals computed by the "emmax-kin" option were used (EMMAX_AMAT; EMMAX_GMAT). No prior weights were assigned for any variants in the rare variant mapping of all tested methods.

The power of each method was estimated as the proportion of runs that significantly detected loci that were simulated to be causal. A significance level of 0.05 after Bonferroni correction was used for each scenario. The p values should be corrected by the total number of SNP sets tested for the MONSTER, famBT and famSKAT methods (there were five SNP tests: one for common variants (MAF ≥ 0.02) and four SNP sets based on the following MAF classes of rare variants: 0.01 ≤

MAF < 0.02; 0.005 ≤ MAF < 0.01; 0.001 ≤ MAF < 0.005; and MAF < 0.001). Thus, if the p value for a tested SNP set with simulated QTL was less than 0.05/5, it was considered to be significant. For EMMAX and DMU, the p value for simulated QTL was corrected by the total number of SNPs tested. For example, if the *p*-value for the simulated QTL was less than 0.05/635, it should be considered as significant for Gene ID: *ENSBTAG00000018852*. However, all the variants tested here are located within a gene and therefore are not independent because of the LD between them. Therefore, we used an alternative multiple-testing correction method based on calculating the effective number of independent SNPs for total number of SNPs according to (Gao et al., 2008). Based on this approach, the effective number of independent SNPs was equal to 17 for Gene ID: *ENSBTAG00000018852* and the corresponding eigenvalues explained 99.5 % of the SNP data variation. Based on these criteria, if the p value for the simulated QTL was less than 0.05/17 for the single variant analysis using EMMAX or DMU, it was considered as significant. The standard errors for each scenario were calculated from bootstrapping based on 100 re-samplings from the 100 simulation runs.

6.3 Results

6.3.1 Comparison of different methods with the null model

Figure 6.1 shows the quantile–quantile plots for the data simulated under the null model (no QTL present). The estimated λ (genomic control) values for MONSTER, famBT, famSKAT, EMMAX_AMAT, and EMMAX_GMAT were less than 1, indicating that the *p* values closely followed the expected distribution under the null hypothesis. Therefore, these methods showed no evidence of inflation of the p values under the null model. However, some of the observed χ^2 values for DMU_AMAT were far too large, which indicated very high false-positive values (Fig. 6.1). However, when rare variants with extremely low MAF (MAF < 0.001) were excluded, the estimated λ for DMU_AMAT followed the expected distribution under the null hypothesis very well (see Additional file 2: Figure S2). The type I error rate for DMU_AMAT was much higher than that for the other methods (MONSTER, famBT, famSKAT, EMMAX_AMAT, and EMMAX_GMAT) using either Bonferroni correction or multiple-testing correction based on the effective number of independent SNPs (see Additional file 3: Figure S3). However, using the effective number of SNPs to correct the significance level also increased type I error rate for linear mixed models (EMMAX_AMAT, EMMAX_GMAT and DMU_AMAT) (see Additional file 3: Figure S3).

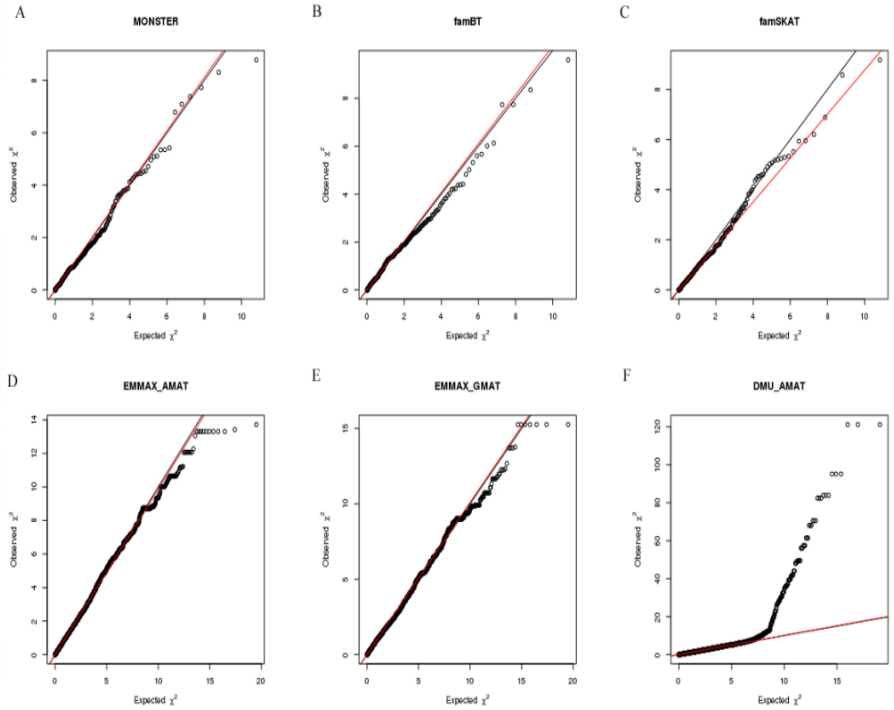


Figure 6.1 Quantile–quantile plots of the null models with different methods. a MONSTER. b famBT. c famSKAT. d EMMAX using the **A** matrix.

6.3.2 Comparison of the power of different methods with different scenarios

The power values of the methods used to detect rare simulated QTL averaged across 100 replicates are in Figs. 6.2, 6.3 and 6.4 (p values adjusted for the effective number of independent SNPs). First, the power values for all rare variant mapping methods across the four MAF classes (0.01 ≤ MAF < 0.02; 0.005 ≤ MAF < 0.01; 0.001 ≤ MAF < 0.005; and MAF < 0.001) were very similar under one of the scenarios. For the scenario with a moderate heritability (h^2 = 0.5), the powers of MONSTER, famBT and famSKAT ranged from 0.19 to 0.42 when multiple rare causal variants were assumed and from 0.09 to 0.30 when one causal rare variant was assumed. Increasing the heritability from 0.3 to 0.8, increased the power to detect QTL from ~0.17 to ~0.61 for MONSTER, famBT and famSKAT when multiple rare causal variants were assumed. No method was able to detect QTL (power ≤ 0.05) that only explained 0.1 % of the genetic variance (Fig. 6.4c, d). When a QTL explained 0.5 % of the genetic variance, the power increased from ~0.13 to ~0.86 as the number of individuals increased from 1000 to 5000 for MONSTER, famBT

and famSKAT (when multiple rare causal variants were assumed) (Fig. 6.4a, b). However, when the QTL explained only 0.1 % of the genetic variance, there was little increase in power (~0.04 to ~0.15) as the number of individuals increased from 1000 to 5000 (Fig. 6.4c, d).

When the *p* values of the total number of SNPs are adjusted by Bonferroni correction, DMU_AMAT, EMMAX_AMAT and EMMAX_GMAT had little power (<0.05) in all scenarios (see Additional file 4: Figure S4). However, when the p values were adjusted by multiple-testing correction based on the effective number of independent SNPs, DMU_AMAT, EMMAX_AMAT and EMMAX_GMAT had less power in all scenarios compared to the specialized methods for mapping multiple causal rare variants. When only one rare variant contributed to the total QTL variance, i.e. when there was only one variant with a relatively large effect, the powers of the LMM (DMU_AMAT, EMMAX_AMAT and EMMAX_GMAT) were similar compared to the specialized methods for rare variant mapping (MONSTER, famBT and famSKAT) (Figs. 6.2, 6.3, 6.4). With EMMAX, the powers were similar regardless of whether the A-matrix or G-matrix was used for the kinships (Figs. 6.2, 6.3, 6.4). When heritability increased from 0.3 to 0.8, the power of all methods increased (Figs. 6.2, 6.3). In general, the power was greater with multiple rare causal variants than with one causal rare variant across all scenarios for MONSTER, famBT and famSKAT (Figs. 6.2, 6.3, 6.4). With a heritability of 0.5, the power across scenarios with one rare causal variant simulated as a QTL remained similar compared to that across scenarios with multiple rare causal variants simulated as QTL for DMU_AMAT, EMMAX_AMAT and EMMAX_GMAT (Figs. 6.2, 6.3, 6.4) but if the total number of SNPs was adjusted by multiple-testing correction, power increased (see Additional file 4: Figure S4). The power of FamBT, compared to the other methods, was greatest across all scenarios for multiple rare causal variants, while that of famSKAT was highest across most scenarios with only one causal rare variant (Figs. 6.2, 6.3, 6.4).

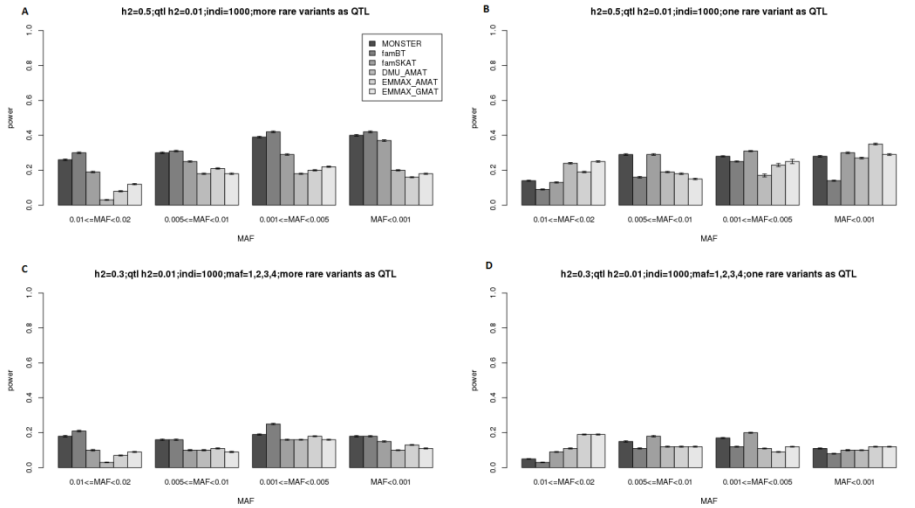


Figure 6.2 Comparison of the power of rare variant mapping methods in scenarios with different MAF for rare variants and heritabilities. a, b Heritability = 0.5; 0.01 ⩽ MAF < 0.02, 0.005 ⩽ MAF < 0.01, 0.001 ⩽ MAF < 0.005, MAF < 0.001; proportion of additive genetic variance explained by the QTL = 0.01; sample size in the test = 1000; with multiple rare variants simulated as QTL (a) and one rare variant simulated as a QTL (b). c, d Heritability = 0.3; 0.01 ⩽ MAF < 0.02, 0.005 ⩽ MAF < 0.01, 0.001 ⩽ MAF < 0.005, MAF < 0.001; proportion of additive genetic variance explained by the QTL = 0.01; sample size in the test = 1000; with multiple rare variant simulated as a QTL (c) and one rare variant simulated as a QTL (d)

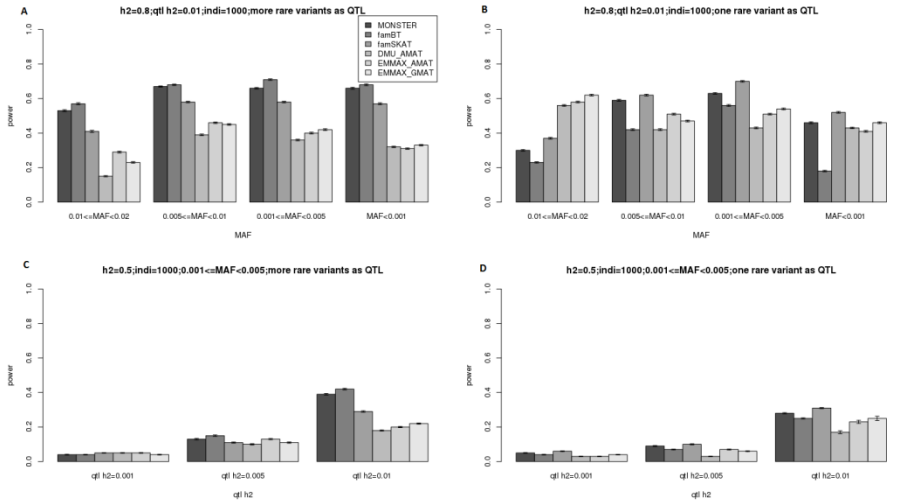


Figure 6.3 Comparison of the power of rare variant mapping methods in scenarios with different heritabilities and proportions of additive genetic variance explained by the QTL.

a, b Heritability = 0.8; 0.01 ⩽ MAF < 0.02, 0.005 ⩽ MAF < 0.01, 0.001 ⩽ MAF < 0.005, MAF < 0.001; proportion of additive genetic variance explained by the QTL = 0.01; sample size in the test = 1000; with multiple rare variants simulated as QTL (a) and one rare variant simulated as a QTL (b). c, d Heritability = 0.5; 0.001 ⩽ MAF < 0.005; proportion of additive genetic variance explained by the QTL = 0.01, 0.005, 0.001; sample size in the test = 1000; with multiple rare variant simulated as a QTL (c) and one rare variant simulated as a QTL (d)

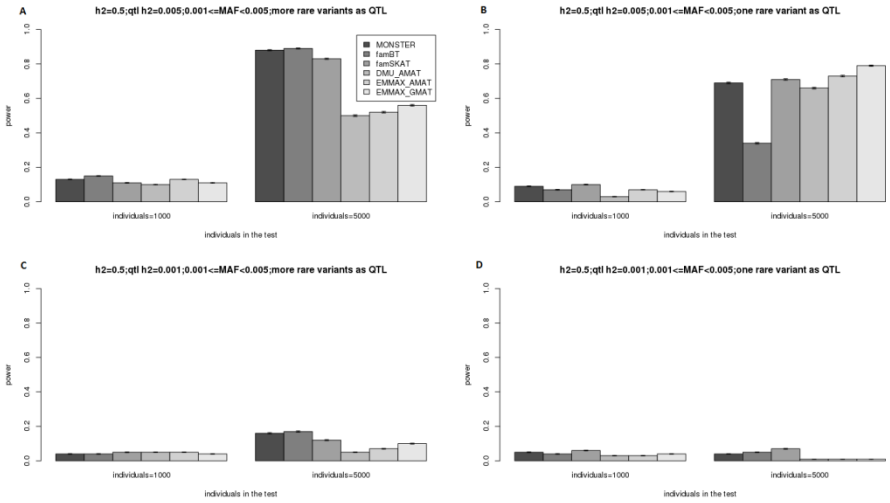


Figure 6.4 Comparison of the power of different methods in scenarios with different sample sizes. a, b Heritability = 0.5; 0.001 ⩽ MAF < 0.005; proportion of additive genetic variance explained by the QTL = 0.005; sample size in the test = 1000, 5000; with multiple rare variants simulated as QTL (a) and one rare variant simulated as QTL (b). c, d Heritability = 0.5; 0.001 ⩽ MAF < 0.005; proportion of additive genetic variance explained by the QTL = 0.001; sample size in the test = 1000, 5000; with multiple rare variants simulated as QTL (c) and one rare variant simulated as QTL (d)

6.4 Discussion

The objective of our study was to compare the power of several gene-based methods to detect rare variants using simulated phenotype data and imputed whole-genome sequence variants for a bovine population with a complex pedigree structure.

Methods that are specialized for the detection of rare variants in a population of individuals with family relationships (MONSTER, famBT and famSKAT) yielded more power than linear mixed models (DMU_AMAT, EMMAX_AMAT and EMMAX_GAMAT) for the detection of QTL with multiple rare causal variants. The

linear mixed model which is the method of choice for association mapping of common variants was less powerful for the detection of QTL with multiple rare causal variants (Figs. 6.2, 6.3, 6.4).

The observed association statistics (χ^2) for data simulated under the null model (no rare variant contributing to the phenotypic variance) followed closely the expected distribution under the null hypothesis for all methods except DMU_AMAT (see Additional file 2: Figure S2). A large number of loci showed a very high observed χ^2 (type I errors) under the null model for DMU_AMAT. This is probably because an extremely low frequency allelic variant will remain confined to a few families or individuals. If, by chance, these families or individuals have extreme phenotypes, that effect will be attributed to the allele resulting in a false positive association. The lower the MAF, the greater the chance that the minor allele is confined to a few families or individuals. Therefore, after filtering out the loci with a very low MAF (MAF < 0.001), the observed χ^2 followed closely the expected χ^2 for DMU_AMAT (see Additional file 2: Figure S2). This result suggests that it is necessary to filter out loci with extremely low MAF when using LMM in order to control false positives. However, this phenomenon was not observed with the EMMAX approach, which could be due to the adjustment of such effects in the first-step of EMMAX when the variance components are estimated.

MONSTER and linear mixed models implemented in DMU_AMAT and EMMAX_AMAT captured most of the total simulated heritability when considering both polygenic variance and the estimated QTL variance (see Additional file 5: Figure S5). DMU_AMAT and EMMAX_AMAT (see Additional file 6: Figure S6) yielded similar estimates of the genetic variance explained by QTL. The genomic heritability estimated by EMMAX_GMAT was considerably lower (0.3) than its simulated value (0.5) (see Additional file 5: Figure S5). The covariance structure among individuals was modeled based on pedigree records for phenotype simulation. The genomic relationships that were estimated from the 50 k SNP data differed considerably from the pedigree-based relationships and therefore explained only part of the additive genetic variance for the trait.

For the simulation with the *ENSBTAG00000035858* gene (see Additional file 7: Figure S7), a similar trend was observed as that found for the ENSBTAG00000018852 gene (see the "Result" section). The power of detecting QTL with a low MAF with the specialized methods for mapping rare variants was around ~30 % in the scenario with a heritability of 0.5 and where the QTL explained 1 % of

the additive genetic variance. Similar results were observed in the simulation with the *ENSBTAG00000035858* gene, i.e. the power of MONSTER, famBT and famSKAT when multiple rare variants explain all the QTL variance was greater (~40 %) than that of the linear mixed models (see Additional file 7: Figure S7). We observed relatively more power for low gene effects and small sample sizes, which is probably because all causal mutations were included in the association analyses. In analyses based on real data, it would be very unlikely that all the causal mutations were included in the SNP sets, for instance because variants may simply be removed during filtration of the data. In our simulation, we also considered the situation with only one rare variant explaining all the QTL variance, and we found that the power of MONSTER, famBT and famSKAT was also greater than that of the linear mixed models when the p-values were adjusted by multiple-testing correction for total number of SNPs (see Additional file 4: Figure S4). This was unexpected since rare variant mapping assumes an incorrect architecture for the locus when there is only one causal rare variant. Less power in the LMM analysis for scenarios with a single rare causal variant could result from the association signal being masked under stringent multiple-testing correction. When we used the effective number of independent SNPs to correct for multiple-testing, the powers for scenarios with single causal rare variants were similar to those of other specialized rare variant mapping methods (Figs. 6.2, 6.3, 6.4). However, in GWAS, *p* values are generally adjusted by Bonferroni correction i.e. by dividing the p values by the total number of SNPs. However, the false positive rate also increased when the p values were not divided by the total number of SNPs (Figs. 6.2, 6.3, 6.4).

Our findings across different scenarios probably reflect the overall power for the detection of rare variants based on QTL variance, genetic architecture and sample size for populations with family relationships as observed in cattle and other livestock species. However, when the QTL effect is small (0.1 % of the additive genetic variance), no method had more than 5 % power (i.e. type I error threshold) for the detection of rare variants with a sample size of 1000 individuals (Fig. 6.4). As expected, increasing the number of individuals increased the power to detect rare variants with small effects (Fig. 6.4).

The power of rare variant association mapping methods (MONSTER, famBT and famSKAT) depends on the genetic architecture of the trait because they differ in their assumption about the underlying variants, direction of their effects as well as the correlation structure between rare variants. This was also shown by the

simulation on the *ENSBTAG00000035858* gene in the main scenarios (see Additional file 7: Figure S7). Specifically, famBT had the greatest power when multiple rare variants in the test SNP set were simulated as QTL while famSKAT had the greatest power when only one rare variant in the test SNP set was simulated as a QTL. The correlation between rare variants in the test SNP set (ρ) was very low when only one rare variant was simulated as the QTL. Therefore, the power of famSKAT (ρ = 0) was greatest while that of famBT (ρ = 1) was greatest when the statistical method's assumptions matched the genetic architecture of the trait. However, the differences in power between MONSTER, famBT and famSKAT were very small across all scenarios. Therefore, when applying these methods on real data for mapping rare variants, it is reasonable to consider all three methods since the genetic architecture of the trait under study is usually unknown. In summary, in cattle, it is recommended to use rare variant association mapping methods to identify low frequency genetic variants especially when multiple rare variants are causal and contribute to the trait. Once identified, these rare variants could be exploited for whole-genome prediction of breeding values in the future.

Imputation accuracies of rare variants are lower than those of common variants and this could have a large impact in association analyses for rare variants on real data (Zheng et al., 2015). We used imputed rare sequence variants in this study instead of simulated genotypes. However, we used simulated phenotypes, assuming that the imputed variants were true. Therefore, imputation errors did not distort the individuals' phenotypes in our study. By using imputed genotypes, the LD structure and allele frequency spectrum are maintained as observed in our population. Therefore, we expect that using imputed genotypes did not affect the conclusions of our study. In real situations, high-coverage exome sequencing or low-coverage whole-genome sequencing of large number of samples may improve the accuracy of genotype call for the rare variants.

Mutations that change the protein structure or lead to a non-functional protein can have a strong phenotypic impact and may therefore be detectable. However, rare variants with subtle effects may be difficult to identify, even if the sample size is large. Therefore, the gene-based approaches used in our study should be considered for genome-wide mapping of rare variants. Besides, computational cost is an important factor to consider when performing genome-wide rare variant mapping. In our analyses, it took ~11 min to perform rare variant mapping for a sample size of 1000 and ~52 min for a sample size of 5000. Considering that there are ~22,000 annotated genes in the bovine genome, this still implies a huge

computational effort when considering all the genes. Therefore, it is important that the algorithms for gene-based mapping are further optimized, but it may also be useful to target rare variants in candidate genes only to save computational time.

6.5 Conclusions

Our findings showed that combining rare variants in a test SNP set with MONSTER, famBT and famSKAT yielded more power to map QTL than linear mixed models for bovine data. We also found that these methods could overcome the confounding of extreme phenotypes in the family mean when mapping rare variants compared to a one-step linear mixed model approach (Yu et al., 2006). In fact, linear mixed models were prone to yield large numbers of type I errors for loci with extremely low MAF (MAF < 0.001), while they were not able to correctly detect causal loci with extremely low MAF. However, EMMAX was robust to extremely low MAF. It is recommended to use methods such as the burden test or variance component tests for mapping rare variants in cattle and other livestock with a similar family structure.

6.6 Appendix

Supplementary material can be found in the online version of the published paper or can be directly be accessed via
https://gsejournal.biomedcentral.com/articles/10.1186/s12711-016-0238-5

Zhang, Q., Guldbrandtsen, B., Calus, M. P., Lund, M. S., & Sahana, G. 2016. Comparison of gene-based rare variant association mapping methods for quantitative traits in a bovine population with complex familial relationships. Genetics Selection Evolution, 48(1):60.

6.7 Acknowledgement

Qianqian Zhang benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”. This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (Grant 0603-00519B).

References

Bouwman, A. C. and R. F. Veerkamp. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genetics 15(1):105.

Brondum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. S. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15(1):728.

Cao, C. C., C. Li, Z. Huang, X. Ma, and X. Sun. 2013. Identifying Rare Variants With Optimal Depth of Coverage and Cost-Effective Overlapping Pool Sequencing. Genetic Epidemiology 37(8):820-830.

Casals, F., A. Hodgkinson, J. Hussin, Y. Idaghdour, V. Bruat, T. de Maillard, J. C. Grenier, E. Gbeha, F. F. Hamdan, S. Girard, J. F. Spinella, M. Lariviere, V. Saillour, J. Healy, I. Fernandez, D. Sinnett, J. L. Michaud, G. A. Rouleau, E. Haddad, F. Le Desit, and P. Awadalla. 2013. Whole-Exome Sequencing Reveals a Rapid Change in the Frequency of Rare Functional Variants in a Founding Population of Humans. PLoS Genetics 9(9).

Chen, H., J. B. Meigs, and J. Dupuis. 2013. Sequence Kernel Association Test for Quantitative Traits in Family Samples. Genetic Epidemiology 37(2):196-204.

Cirulli, E. T. 2016. The Increasing Importance of Gene-Based Analyses. PLoS Genetics 12(4):e1005852.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brondum, X. P. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46(8):858-865.

Devlin, B. and K. Roeder. 1999. Genomic control for association studies. Biometrics 55(4):997-1004.

Elsik, C. G., D. R. Unni, C. M. Diesh, A. Tayal, M. L. Emery, H. N. Nguyen, and D. E. Hagen. 2016. Bovine Genome Database: new tools for gleaning function from the Bos taurus genome. Nucleic Acids Research 44(D1):D834-839.

Gao, X., J. Starmer, and E. R. Martin. 2008. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. Genetic Epidemiology 32(4):361-369.

Gibson, G. 2012. Rare and common variants: twenty arguments. Nature Reviews Genetics 13(2):135-145.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genetics 5(6):e1000529.

Ionita-Laza, I., S. Lee, V. Makarov, J. D. Buxbaum, and X. H. Lin. 2013. Family-based association tests for sequence data, and comparisons with population-based association tests. European Journal of Human Genetics 21(10):1158-1162.

Iso-Touru, T., G. Sahana, B. Guldbrandtsen, M. S. Lund, and J. Vilkki. 2016. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genetics 17(1):1.

Jiang, D. and M. S. McPeek. 2014. Robust Rare Variant Association Testing for Quantitative Traits in Samples With Related Individuals. Genetic Epidemiology 38(1):10-20.

Kadri, N. K., B. Guldbrandtsen, P. Sorensen, and G. Sahana. 2014. Comparison of genome-wide association methods in analyses of admixed populations with complex familial relationships. PLoS One 9(3):e88926.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin. 2010. Variance component model to account for sample structure in genome-wide association studies. Nature Genetics 42(4):348-U110.

Kemper, K. E., P. M. Visscher, and M. E. Goddard. 2012. Genetic architecture of body size in mammals. Genome Biology 13(4):244.

Lee, S., G. R. Abecasis, M. Boehnke, and X. H. Lin. 2014. Rare-Variant Association Analysis: Study Designs and Statistical Tests. American Journal of Human Genetics 95(1):5-23.

Lee, S., M. C. Wu, and X. H. Lin. 2012. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762-775.

Li, B. S. and S. M. Leal. 2008. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. American Journal of Human Genetics 83(3):311-321.

MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, and C. Gunter. 2014. Guidelines for investigating causality of sequence variants in human disease. Nature 508(7497):469-476.

Madsen, B. E. and S. R. Browning. 2009. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genetics 5(2):e1000384.

Madsen, P. and J. Jensen. 2006. DMU A Package for Analysing Multivariate Mixed Models. 8th World Congress on Genetics Applied to Livestock Production. 247.

Maher, B. 2008. Personal genomes: The case of the missing heritability. Nature 456(7218):18-21.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-753.

Morgenthaler, S. and W. G. Thilly. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 615(1-2):28-56.

Moutsianas, L., V. Agarwala, C. Fuchsberger, J. Flannick, M. A. Rivas, K. J. Gaulton, P. K. Albers, G. McVean, M. Boehnke, D. Altshuler, M. I. McCarthy, and G. D. Consortium. 2015. The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. PLoS Genetics 11(4): e1005165.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly. 2011. Testing for an Unusual Distribution of Rare Variants. PLoS Genetics 7(3): e1001322.

Price, A. L., G. V. Kryukov, P. I. W. de Bakker, S. M. Purcell, J. Staples, L. J. Wei, and S. R. Sunyaev. 2010. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. American Journal of Human Genetics 86(6):832-838.

Sahana, G., B. Guldbrandtsen, L. Janss, and M. S. Lund. 2010. Comparison of association mapping methods in a complex pedigreed population. Genet Epidemiol 34(5):455-462.

Schaid, D. J., S. K. McDonnell, J. P. Sinnwell, and S. N. Thibodeau. 2013. Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods With Pedigree or Population Structured Data. Genetic Epidemiology 37(5):409-418.

Schifano, E. D., M. P. Epstein, L. F. Bielak, M. A. Jhun, S. L. R. Kardia, P. A. Peyser, and X. H. Lin. 2012. SNP Set Association Analysis for Familial Data. Genetic Epidemiology 36(8):797-810.

Steinberg, M. H. and A. H. Adewoye. 2006. Modifier genes and sickle cell anemia. Current opinion in hematology 13(3):131-136.

Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Q. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. M. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J.

Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, B. GO, S. GO, and N. E. S. Project. 2012. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science 337(6090):64-69.

Thein, S. L. and S. Menzel. 2009. Discovering the genetics underlying foetal haemoglobin production in adults. British Journal of Haematology 145(4):455-467.

van Binsbergen, R., M. C. A. M. Bink, M. P. L. Calus, F. A. van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 46(1):41.

Wu, M. C., S. Lee, T. X. Cai, Y. Li, M. Boehnke, and X. H. Lin. 2011. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. American Journal of Human Genetics 89(1):82-93.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price. 2014. Advantages and pitfalls in the application of mixed-model association methods. Nature Genetics 46(2):100-106.

Yu, J. M., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38(2):203-208.

Zheng, H. F., J. J. Rong, M. Liu, F. Han, X. W. Zhang, J. B. Richards, and L. Wang. 2015. Performance of Genotype Imputation for Low Frequency and Rare Variants from the 1000 Genomes. Plos One 10(1): e0116487.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10(4):R42.

7

Contribution of rare and low-frequency whole-genome sequence variants to complex traits variation in dairy cattle

Qianqian Zhang[1,2], Mario Calus[2], Bernt Guldbrandtsen[1], Mogens S Lund[1] and Goutam Sahana[1]

[1] Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, Tjele DK-8830, Denmark; [2] Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen 6700 AH, The Netherlands.

Genetics Selection Evolution (2017) 49:60

Abstract

Background: Whole-genome sequencing and imputation methodologies have enabled the study of the effects of genomic variants with low to very low minor allele frequency (MAF) on variation in complex traits. Our objective was to estimate the proportion of variance explained by imputed sequence variants classified according to their MAF compared with the variance explained by the pedigree-based additive genetic relationship matrix for 17 traits in Nordic Holstein dairy cattle.

Results: Imputed sequence variants were grouped into seven classes according to their MAF (0.001–0.01, 0.01–0.05, 0.05–0.1, 0.1–0.2, 0.2–0.3, 0.3–0.4 and 0.4–0.5). The total contribution of all imputed sequence variants to variance in de-regressed estimated breeding values or proofs (DRP) for different traits ranged from 0.41 [standard error (SE) = 0.026] for temperament to 0.87 (SE = 0.011) for milk yield. The contribution of rare variants (MAF < 0.01) to the total DRP variance explained by all imputed sequence variants was relatively small (a maximum of 12.5% for the health index). Rare and low-frequency variants (MAF < 0.05) contributed a larger proportion of the explained DRP variances (>13%) for health-related traits than for production traits (<11%). However, a substantial proportion of these variance estimates across different MAF classes had large SE, especially when the variance explained by a MAF class was small. The proportion of DRP variance that was explained by all imputed whole-genome sequence variants improved slightly compared with variance explained by the 50 k Illumina markers, which are routinely used in bovine genomic prediction. However, the proportion of DRP variance explained by imputed sequence variants was lower than that explained by pedigree relationships, ranging from 1.5% for milk yield to 37.9% for the health index.

Conclusions: Imputed sequence variants explained more of the variance in DRP than the 50 k markers for most traits, but explained less variance than that captured by pedigree-based relationships. Although in humans partitioning variants into groups based on MAF and linkage disequilibrium was used to estimate heritability without bias, many of our bovine estimates had a high SE. For a reliable estimate of the explained DRP variance for different MAF classes, larger sample sizes are needed.

Key words: Runs of homozygosity, Polymorphisms, Inbreeding, Cattle, Genome sequencing

7.1 Background

Associations of common genetic variants with complex diseases and quantitative traits have been successfully identified in humans and livestock (Purcell et al., 2009, Yang et al., 2010, Zhang et al., 2016b). However, these loci explain only a small fraction of the total genetic variance of a trait. In human genetics, the portion of the additive genetic variance that remains unexplained by the associated genetic variants is known as the "missing heritability" (Maher, 2008, Manolio et al., 2009, Gibson, 2012). One strategy to reduce the missing heritability is genomic prediction where all markers regardless of the magnitude and statistical significance of their effects are used to predict genetic values and estimate genetic variances (Eichler et al., 2010, Jensen et al., 2012, Roman-Ponce et al., 2014). Jensen et al. (2012) reported that on average 77.2% of the genetic variance for six dairy cattle traits was attributed to genomic relationships constructed based on the Illumina BovineSNP50 BeadChip (50 k) single nucleotide polymorphisms (SNP)s. Roman-Ponce et al. (2014) reported that a genomic relationship matrix based on the 50 k SNP chip could explain between 51 and 94% of the genetic variance, depending on the reliabilities of the phenotypes used for milk yield, fat yield, protein yield and somatic cell count (Yang et al., 2010). However, previous studies also showed that a wide gap remains between the proportion of variance explained using genomic relationships constructed from 50 k SNP chips and the genetic variance explained by pedigree-based relationships (Garrick et al., 2009, Jensen et al., 2012, Haile-Mariam et al., 2013, Roman-Ponce et al., 2014). This "missing" proportion of the genetic variance may affect the maximum accuracy that genomic prediction could achieve in livestock breeding (Dekkers, 2007).

Rare variants may play a significant role in quantitative trait variation (Gibson, 2012, Kemper et al., 2012, Zhang et al., 2016a) and contribute to the "missing heritability". With the development of whole-genome sequencing technologies, next-generation sequence data have been generated for a large number of individuals in various cattle populations (Daetwyler et al., 2014). These sequence data have predominantly been used as a reference to impute SNP array genotypes to whole-genome sequences for individuals with phenotypes (Brondum et al., 2014). By using imputed sequence data, rare and low-frequency variants can be identified and studied for much larger numbers of individuals.

When whole-genome sequence data are available, linkage disequilibrium (LD) between SNPs and causal variants increases and a large fraction of the causal variants themselves will be available for analysis. Therefore, an increase in the proportion of the variance that can be explained for quantitative traits is expected when whole-genome sequence variants are used compared with the use of SNP array data (Jensen et al., 2012, Roman-Ponce et al., 2014).

However, using whole-genome regressions which regress phenotypes on the whole-genome sequence variants using a linear model to infer the proportion of variance explained for a trait may result in biased estimates (de los Campos et al., 2015, Yang et al., 2015). First, if the causal variants are enriched in regions with higher or lower than average LD, heritability estimated based on genomic information is biased (Yang et al., 2010, Yang et al., 2015). Second, if causal variants have a different spectrum of minor allele frequencies (MAF) than the SNPs used, heritability estimated based on genomic information will also be biased (Yang et al., 2015). Due to strong artificial selection, causal variants in dairy cattle are expected to often have extreme allele frequencies, whereas the content of DNA chips is biased by design towards highly polymorphic SNPs. Therefore, the spectrum of the allele frequencies of causal variants is expected to be quite different from that of SNPs on the commonly used 50 k chip. The effect of differences in the spectrum of allele frequencies and in LD heterogeneity on heritability estimates based on genomic information has not yet been studied in dairy cattle. However, several studies have shown that LD in bovine populations is relatively high, with long haplotype blocks, compared to that in human populations (McKay et al., 2007, Qanbari et al., 2010). Thus, we expect that the effect of heterogeneity in LD on heritability estimates is relatively small in bovine populations.

Recently, Yang et al. (2015) proposed an LD- and MAF-stratified genomic-relatedness-based restricted maximum-likelihood (GREML-LDMS) method for human data that partitions the variance explained across classes of variants with different MAF. It also accounts for region-specific heterogeneity in LD (Yang et al., 2010). They showed that heritability estimates obtained with the GREML-LDMS method were unbiased for human height and body mass index and found negligible missing heritability for both traits when using imputed variants (Yang et al., 2015). Thus, we expect that, in cattle, the variance explained by imputed sequence data when estimated using the GREML-LDMS approach will capture larger proportions of the variance compared to estimates obtained from GREML using genomic

relationships based on SNP chip genotypes (VanRaden, 2008, Hayes et al., 2009, Yang et al., 2010).

The objectives of this study were to: (1) estimate the proportion of variance explained by whole-genome sequence variants for 17 traits in Nordic Holstein cattle; (2) estimate the proportion of variance explained by partitioning variants according to MAF, and with or without taking LD heterogeneity into consideration; and (3) compare estimates of the proportions of genetic variance explained by relationships based on pedigree, 50 k SNPs, and imputed whole-genome sequence variants.

7.2 Methods

7.2.1 Phenotypes and genotypes

In total, 5065 Holstein progeny-tested bulls with estimated breeding values were genotyped using the BovineSNP50 BeadChip (50 k) array version 1 or 2 (Illumina, San Diego, CA, USA). The phenotypes used in this study were de-regressed estimated breeding values or proofs (DRP) with a minimum reliability of 0.2 for 17 traits (Table 7.1). For details regarding the 17 traits, recording procedures and models to estimate breeding values for these three indices, see http://www.nordicebv.info/ntm-and-breeding-values. The number of bulls with both genotype data and DRP for different traits ranged from 4485 to 4949 (Table 7.1).

DNA was extracted using standard procedures from either semen or blood samples. Genotyping was performed by GenoSkan A/S, Tjele, Denmark or the Department of Molecular Biology and Genetics in Aarhus University. The data editing steps were the same as in (Iso-Touru et al., 2016). Quality parameters used to select SNPs were a minimum call rate of 85% for individuals and of 95% for loci. SNPs that were monomorphic or deviated from Hardy–Weinberg proportions ($P < 0.00001$) were excluded. The minimal acceptable GenCall score (GC) was 0.60 for SNPs and 0.65 for individuals. After quality control, 43,415 SNPs and 5065 individuals remained for analyses. The genomic positions of SNPs were taken from the UMD3.1 Bovine genome assembly (Zimin et al., 2009).

Table 7.1 Description of the traits.

Trait	Abbr.	Average DRP reliability	Standard deviation of DRP reliability	The range of DRP reliability	Number of bulls with DRP
Yield index	YIELD	0.936	0.027	0.634- 0.990	4649
Milk yield	MILK	0.934	0.031	0.634-0.990	4949
Protein yield	PROT	0.934	0.031	0.634-0.990	4876
Fat yield	FAT	0.933	0.031	0.634-0.990	4883
Udder index	MILKORG	0.773	0.080	0.444-0.990	4834
Milking speed	MILKSP	0.768	0.128	0.327-0.990	4753
Longevity	LONG	0.747	0.093	0.304-0.993	4551
Mastitis	MASTI	0.814	0.078	0.344-0.983	4858
Other-diseases (health)	HEALTH	0.577	0.132	0.207-0.990	4593
Feet and legs	LEG	0.570	0.121	0.204-0.990	4831
Daughter calving index	CALV	0.670	0.090	0.204-0.990	4788
Service sire calving index (birth index)	BIRTH	0.738	0.083	0.442-0.990	4795
Fertility	FERT	0.671	0.112	0.214-0.990	4806
Body conformation index	BODY	0.805	0.071	0.513-0.990	4832
Growth	GROWTH	0.912	0.048	0.513-0.990	4397
Temperament	TEMP	0.603	0.135	0.212-0.990	4526
Nordic total merit index	NTM	0.934	0.031	0.634-0.990	4834

In a previous study (Iso-Touru et al., 2016), the 50 k genotypes of 5065 animals were imputed to whole-genome sequence data using a two-step approach by first imputing 50 k genotypes to a high-density BovineHD BeadChip (HD, Illumina) using a multi-breed reference of 3383 animals, followed by imputing to the whole-genome sequence level using a multi-breed reference consisting of 1228 animals from run4 of the 1000 bull genomes project (Daetwyler et al., 2014) and additional whole-genome sequences from Aarhus University (Hoglund et al., 2014). The whole-genome sequence reference genotypes were pre-phased with BEAGLE4 r1274 (Browning and Browning, 2013). Imputation to HD genotypes was done by using IMPUTE2 v2.3.1 (Howie et al., 2009) and imputation to the whole-genome level by using Minimac2 (Fuchsberger et al., 2015). The imputed variants were filtered to remove those with a MAF lower than 0.001, which means that SNPs with less than ~10 copies of the minor allele in the data analysed were removed.

7.2.2 Contribution of different classes of genetic variants based on MAF to DRP variance

The GREML-MS and GREML-LDMS methods (Yang et al., 2015) were used to calculate the proportion of DRP variance explained by imputed sequence variants. For the GREMLMS method, the imputed sequence variants were grouped into seven classes based on their MAF (0.001–0.01, 0.01–0.05, 0.05–0.1, 0.1–0.2, 0.2–0.3, 0.3–0.4 and 0.4–0.5). The number of variants was very similar across MAF groups (Fig. 7.1). Rare variants were defined as those with a MAF ranging from 0.001 to 0.01; low-frequency variants as those with a MAF ranging from 0.01 to 0.05; and common variants had a MAF higher 0.05. Average imputation accuracies (IMPUTE-INFO score defined by (Marchini and Howie, 2010)) for rare and low-frequency variants were 0.850 and 0.873, respectively [see Additional file 1: Table S1]. We did not filter variants strictly based on imputation accuracy, i.e. all variants with IMPUTE-INFO score were included in the analyses, because a study using human data suggested that removing variants based on a more restrictive IMPUTEINFO threshold leads to a loss of variance explained (Yang et al., 2015).

Genomic relationship matrices (GRM) for each of the seven classes of variants were calculated following (Yang et al., 2010) and fitted jointly in a multicomponent REML analysis:

$$\mathbf{y} = \mathbf{1}\mu + \sum_{i=1}^{7} \mathbf{g}_i + \mathbf{e} \qquad (1)$$

where y is the vector of phenotypes (DRP), 1 is a vector of 1s, μ is the general mean, $\mathbf{g}_i$ is a vector of the genetic values for the ith variant class (i = 1,2,…,7), $\mathbf{g}_i \sim \mathrm{N}(0, \mathbf{G}_i\sigma_i^2)$, , where $\mathbf{G}_i$ is the GRM of the ith class, and e is a vector of residuals with $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_e^2)$. The variance components were estimated by using the REML approach implemented in the genome-wide complex trait analysis (GCTA) software (Jensen, 1997, Madsen and Jensen, 2006). The proportion of variance in DRP explained by class i of variants was calculated as:

$$\hat{\sigma}_i^2 \Big/ \sum_{i=1}^{7} \hat{\sigma}_i^2 + \hat{\sigma}_e^2.$$

To account for the region-specific heterogeneity in LD, we used the GREML-LDMS approach proposed by Yang et al. (2015). First, for each SNP, an LD score was computed as the sum of the LD measure r^2 between this SNP and other SNPs in a 20-Mb region centered on this SNP. Then, the mean LD score of the variants in each segment which contained twice the average number of variants per 100-kb window of a chromosome was calculated and these were used to partition the variants within each of the seven MAF classes into four equally-sized LD groups based on increasing mean LD scores, following Yang et al. (2015), resulting in 28 groups. Then, Model (1) was fitted using these 28 genetic components. In addition, to compare the estimates of variance components based on the GREML-MS and GREML-LDMS methods, the variants were also stratified into three different LD groups within each of the seven MAF classes, resulting in 21 genetic components. The proportion of DRP variance explained by rare, low-frequency and common variants, as defined previously, was divided by the sum of the DRP variances to compare their relative contribution to the total DRP variance explained.

The GRM used in GCTA, assumes that allelic effects of both common and rare variants follow the same distribution, similar to VanRaden's method 2 (VanRaden, 2008, Yang et al., 2011). This means that a common variant explains more variance than a rare variant. To verify whether this assumption is reasonable, expected contributions of different classes of MAF variants to the variance were compared to our empirical results. The expected variance explained by the variants from different MAF classes were computed under the assumptions of VanRaden's methods 1 and 2 (VanRaden, 2008). For VanRaden's method 1, the expected variance explained by a class of variants is:

$$\sum_{i=1}^{j_1} 2p_i(1-p_i) / \sum_{i=1}^{j_2} 2p_i(1-p_i),$$

where p_i is the MAF of the ith locus and the numerator is the sum for the variants in each class until the j1th locus and the denominator is the sum for all the variants

until the j_2^{th} locus. For VanRaden's method 2, the expected proportion of genetic variance explained by a class of variants is N_{class}/N, where N_{class} is the number of variants per class, and N is the total number of loci used in the calculation. Correspondingly, VanRaden's method 1 assigns a large amount of variance to common variants, while VanRaden's method 2 puts more emphasis on rare variants.

The phenotypes used in our analysis, as is often the case in animal breeding, were DRP derived from estimated breeding values with varying reliabilities. Weights derived from those reliabilities are commonly used in analyses that use DRP. However, the GCTA software does not support the use of weights, because it was developed in the context of human data analysis where the phenotypes used are typically directly measured on the genotyped individuals. However, the average reliability of the DRP used here were quite high (Table 7.1). For example, the average reliability of milk yield was 93.4%. Therefore, ignoring DRP reliabilities in our analyses is not expected to affect the results.

7.2.3 Proportion of DRP variance captured by pedigree and 50 k SNPs

The genetic variance estimated by using the pedigree relationship matrix was compared to the variance explained by the imputed sequence variants and the 50 k SNPs. The proportions of DRP variance explained by pedigree and genomic relationships were estimated by fitting pedigree and 50 k data separately or jointly in the model as described below:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}_{\boldsymbol{a}}\mathbf{a_1} + \mathbf{e} \qquad (2)$$

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}_{\boldsymbol{g}}\mathbf{g_1} + \mathbf{e} \qquad (3)$$

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}_{\boldsymbol{g}}\mathbf{g_2} + \mathbf{Z}_{\boldsymbol{a}}\mathbf{a_2} + \mathbf{e} \qquad (4)$$

where $\mathbf{y}$ is the vector of phenotypes, $\mathbf{1}$ is a vector of 1s, μ is the general mean. $\mathbf{Z}_{\boldsymbol{g}}$ and $\mathbf{Z}_{\boldsymbol{a}}$ are incidence matrices that relate DRP to breeding values in $\mathbf{g_1}$, $\mathbf{g_2}$, $\mathbf{a_1}$ and $\mathbf{a_2}$, respectively. Vectors $\mathbf{a_1}$ and $\mathbf{a_2}$ contain random effects with variance $\mathrm{var}(\mathbf{a}) = \mathbf{A}\sigma_a^2$, where $\mathbf{A}$ is the additive genetic relationship matrix computed from pedigree records. Finally, $\mathbf{e}$ is a vector of residuals with $\mathbf{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_e^2)$.

Models (2), (3) and (4) are labeled as "REML-PED", "REML-GRM" and "REML-PEDGRM", respectively. Analyses using pedigree relationships were implemented in the DMU software (Madsen and Jensen, 2006). The vectors $\mathbf{g_1}$ and $\mathbf{g_2}$ contain random effects with variance $\mathrm{var}(\mathbf{g}) = \mathbf{G}\sigma_g^2$, where $\mathbf{G}$ is the GRM calculated following VanRaden's method 1 (VanRaden, 2008):

$$\mathbf{G} = \frac{(\mathbf{X}-2\mathbf{p}\mathbf{1}')(\mathbf{X}-2\mathbf{p}\mathbf{1}')}{2\sum_{j=1}^{n} p_j(1-p_j)},$$

where $\mathbf{X}$ is the allele sharing matrix with the number of copies of the second allele, $\mathbf{p}$ is a vector with allele frequencies, and $\mathbf{1}$ is a vector of 1s. The factor $2\sum_{j=1}^{n} p_j(1-p_j)$ scales $\mathbf{G}$ to be comparable to the pedigree relationship matrix. Analyses using the 50 k data GRM were implemented using the REML-GRM model of the GCTA software (Yang et al., 2011). In addition, the REML-PEDGRM model was fitted with $\mathbf{a_2}$ and $\mathbf{g_2}$ simultaneously implemented in the DMU software. Reliabilities of DRP were not used in the models analyzed by DMU for consistency with the analyses using GCTA. The variance explained by pedigree relationship was re-scaled for REML-PED and REML-PEDGRM to use the same base genomic relationships, following Legarra (2016).

7.3 Results

7.3.1 Contribution of different classes of genetic variants based on MAF to DRP variance

Additional file 1: Table S1 shows the proportion of DRP variance explained and standard error (SE) for variants partitioned into seven MAF groups for 17 traits and Additional file 2: Table S2 presents the same for variants partitioned into seven MAF groups and four LD groups for 17 traits. A substantial proportion of the variance estimates had large SE for most traits when variants were partitioned into seven MAF groups and four LD groups [see Additional file 2: Table S2]. A similar pattern of large SE for the estimates was observed when variants were partitioned into seven MAF groups and three LD groups. However, relatively better estimates were obtained when variants were partitioned into seven MAF groups only [see Additional file 1: Table S1]. Therefore, only results for variants partitioned into seven MAF groups are presented here. However, partitioning variants into seven MAF groups also resulted in several variance estimates with large SE, especially when the estimates were small [see Additional file 1: Table S1].

Interestingly, we observed that the relative contribution of variants with a MAF between 0.2 and 0.3 was substantially higher than that of other classes for MILK, FAT and PROT, as well as for LEG and LONG, while the imputed sequence variants were more or less evenly distributed across each MAF class (Fig. 7.1). This might be due to the *DGAT1* gene (Grisart et al., 2004) (located on chromosome 14, position 1,802,265 bp with a MAF = 0.29), which is the largest milk-related QTL, explaining

11.2% of the DRP variance in MILK, 16.9% of the DRP variance in FAT and 2.9% of the DRP variance in PROT.

The proportion of DRP variance explained by rare (MAF < 0.01), low-frequency (MAF = 0.01–0.05) and common variants (i.e. MAF = 0.05–0.1, 0.1–0.2, 0.2–0.3, 0.3–0.4 and 0.4–0.5) in Additional file 1: Table S1 was divided by the total proportion of DRP variance explained and the results are summarized in Table 2 with three classes of variants (i.e. rare, low-frequency and common variants). The proportion of the DRP variance explained by the imputed sequence variants ranged from 0.406 (SE = 0.026) for TEMP to 0.872 (SE = 0.011) for MILK. The highest relative contribution among different classes of MAF was observed for the group of common variants (MAF ≥ 0.05) and ranged from 0.755 for HEALTH to 0.980 for BIRTH. For rare variants (MAF < 0.01), the contribution to the DRP variance explained was relatively small (ranging from 0 (with high SE) for FERT and BIRTH to 0.125 for HEALTH) compared with that from common variants (Table 7.2). The rare and low frequency variants (MAF < 0.05) contributed higher proportions of the explained DRP variance (in total >0.13 based on Table 7.2) for the health-related traits [i.e. fertility, other-diseases (health), longevity, feet and legs] compared with the production traits (in total <0.11 based on Table 7.2, i.e. yield index, protein yield and milk yield) (Table 7.2; Fig. 7.1).

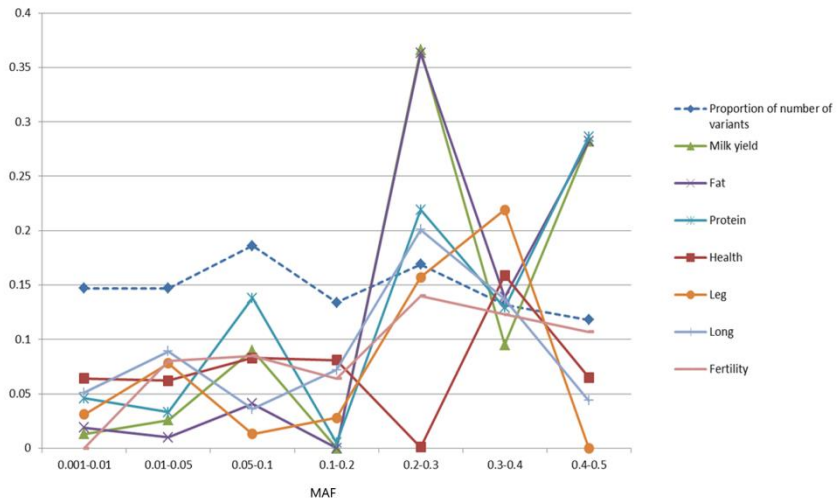


Figure 7.1 Proportion of variants in different MAF classes and their relative contribution to DRP variance for different traits. On the x axis are the variants with MAF classes: 0.001–0.01; 0.01–0.05; 0.05–0.1; 0.1–0.2; 0.2–0.3; 0.3–0.4 and 0.4–0.5 and on the y axis is the

proportion of number of variants over the total number of imputed sequence variants (dark blue), relative contribution of explained DRP variance for different MAF classes variants for different traits based on “Trait abbreviation” in different colors.

7.3.2 Proportion of DRP variance captured by pedigree and 50 k SNPs

The proportions of DRP variance explained for 17 traits by the different models (i.e. GREML-MS, REML-PED, REML-GRM and REML-PEDGRM) and using different information sources to construct relationship matrices (i.e. imputed sequence variants, 50 k SNPs or pedigree data) are in Table 7.3. Estimates of residual variance over the total variance of DRP are in Additional file 3: Table S3 and the Akaike information criterion (AIC) (Akaike, 1981) of the different models are in Additional file 4: Table S4. We observed that estimates of the residual variance and total DRP variance were similar across all models and information sources for a given trait. Therefore, the proportion of DRP variance explained was comparable across models and data sources for a trait [see Additional file 3: Table S3]. For most traits, REML-PEDGRM had the lowest AIC value, which means that this model fit the data best, whereas for some traits, GREML-MS fit the data best [see Additional file 4: Table S4].

Imputed sequence variants explained more DRP variance than 50 k SNPs for most traits (Table 7.3). However, the DRP variance explained by imputed sequence variants was still smaller than the genetic variance estimated by using the pedigree-based relationship matrix; the difference was smallest for MILK (0.015) and largest for HEALTH (0.379).

The variance explained by fitting both pedigree and genomic relationship matrices (GRM) using the 50 k data in the PED + 50 k-DMU model, relative to the variance explained by the pedigree-based relationship matrix alone (PED-DMU), ranged from 109.2% for MASTI to 90.3% for FERT (Table 7.3). Furthermore, the proportion of explained DRP variance by 50 k-based GRM in the total explained genetic variance from both 50 k-based GRM and pedigree-based relationship matrix using PED + 50 k-DMU model ranged from 79.8% for FAT to 26.1% for HEALTH. These results indicate that common variants were able to capture a large proportion of the genetic variance, especially for production traits.

Table 7.2 Relative contribution to the proportion of DRP variance explained by variants in different MAF classes for 17 Traits.

Traits and scenarios	Relative contribution of MAF classes to the explained DRP variance			Total proportion of DRP variance explained
	0.001-0.01	0.01-0.05	0.05-0.5	
YIELD	0.063	0.038	0.899	0.860
MILK	0.015	0.030	0.955	0.872
PROT	0.054	0.038	0.908	0.858
FAT	0.022	0.012	0.966	0.854
MILKORG	0.072	0.003	0.925	0.679
MILKSP	0.005	0.035	0.960	0.719
LONG	0.081	0.141	0.778	0.630
MASTI	0.019	0.000	0.981	0.669
HEALTH	0.125	0.121	0.755	0.514
LEG	0.059	0.149	0.796	0.525
CALV	0.037	0.000	0.963	0.507
BIRTH	0.000	0.020	0.980	0.602
FERT	0.000	0.133	0.867	0.600
BODY	0.088	0.026	0.886	0.568
GROWTH	0.010	0.087	0.903	0.814
TEMP	0.054	0.059	0.887	0.406
NTM	0.031	0.030	0.940	0.847

All the variants were partitioned into seven MAF classes. In this table, we report the proportion of DRP variance explained for three groups of MAF classes (rare: MAF < 0.01, low-frequency: 0.01 ≤ MAF < 0.05 and common: MAF ≥ 0.05). For the common variants group, the proportion of DRP variance explained was sum of proportion of DRP variance explained for classes of variants with MAF: 0.05-0.1; 0.1-0.2; 0.2-0.3; 0.3-0.4 and 0.4-0.5.

Table 7.3 Proportion of DRP variance explained using different Methods.

Traits	GREML-MS	REML-GRM	REML-PED	REML-PEDGRM
YIELD	0.860	0.845	0.923	0.941
MILK	0.872	0.844	0.887	0.927
PROT	0.858	0.847	0.943	0.963
FAT	0.854	0.840	0.898	0.914
MILKORG	0.679	0.703	0.811	0.816
MILKSP	0.719	0.715	0.748	0.840
LONG	0.630	0.606	0.884	0.881
MASTI	0.669	0.684	0.704	0.769
HEALTH	0.514	0.502	0.893	0.892
LEG	0.525	0.525	0.709	0.669
CALV	0.507	0.504	0.698	0.689
BIRTH	0.602	0.612	0.698	0.695
FERT	0.600	0.594	0.851	0.769
BODY	0.568	0.560	0.633	0.594
GROWTH	0.814	0.800	0.916	0.943
TEMP	0.406	0.403	0.645	0.645
NTM	0.847	0.839	-[a]	-

GREML-MS refers to estimation using the GREML-MS method with imputed sequence variants partitioned into MAF classes. REML-GRM refers to estimation using 50k markers with the REML-GRM model implemented in GCTA. REML-PED refers to using pedigree relationship in the REML-PED model implemented in DMU. REML-PEDGRM refers to fitting both 50k markers and pedigree relationship in the REML-PEDGRM model implemented in DMU.
a The model did not converge.

7.4 Discussion

7.4.1 Contribution of MAF classes to the variance of DRP

We estimated the relative contribution of genetic variants in different MAF classes to the explained DRP variance. However, many of these estimates had large SE when variants were partitioned into MAF and LD groups, or only into MAF groups. Although the method of partitioning variants in different MAF and LD groups was used to estimate heritability accurately in human data, many of our estimates for this bovine population had large SE. The number of individuals used in the human

study was 44,126 (Yang et al., 2015), which was much larger than the sample size used in this study in cattle (~5000). Therefore, to obtain reliable estimates of the explained DRP variance for different MAF classes, a larger sample size is needed in cattle population.

For all traits, the relative contribution of rare and low frequency variants to the proportion of DRP variance explained was small compared to the contribution of common variants. For health-related traits, the proportion of DRP variance explained by rare and low frequency variants was on average more than 13%, which was high compared to that for production traits. Gonzalez-Recio et al. (2015) also reported that rare variants explained 14% of the genetic variance for fertility in Holstein cattle. These results reflect that the genetic architecture of health-related traits probably differs from that of production traits in the sense that rare variants have a relatively larger impact on variation in health-related traits. This is expected since selection is purging the rare variants with a negative effect on fitness, for example, the rare deleterious variants will be purged by selection. However, the rare and low-frequency variants with a positive effect such as selective advantage could be very relevant for long-term selection response if they have a medium to large effect (MacLeod et al., 2014).

The variance explained by the class of variants with a MAF between 0.2 and 0.3 was low (0.001) for HEALTH (Fig. 7.1) and [see Additional file 1: Table S1] but is probably not biologically relevant given the large SE of this estimate. When we compared DRP variance among the traits analysed, we observed no specific pattern of rare frequency variants explaining more DRP variance than low-frequency variants. However, again the large SE for the estimates may mask any pattern that may be present. For YIELD, PROT, MILKORG, MASTI, CALV and BODY, rare variants explained more variance than low-frequency variants; for MILK, FAT, MILKSP, LEG, BIRTH, FERT and GROWTH, low-frequency variants explained more variance than rare variants; and for HEALTH, TEMP and NTM, rare variants explained a similar proportion of variance as that found for low-frequency variants. Rare or low-frequency variants with more explained DRP variance for different traits might reflect the genetic architecture (i.e. what kind of causal variants underlie the traits). Rare or low-frequency causal variants generally have larger effect sizes (Marouli et al., 2017) and might also have a larger contribution to phenotypic variation. For human height, rare variants explained 8.4% of the genetic variance and variants with a MAF ranging from 0.01 to 0.1 explained 13% of the genetic variance (Yang et

al., 2015). However, a previous study on bovine fertility reported that rare variants explained 14% of the genetic variance, while low-frequency variants (0.01 < MAF ≤ 0.05) explained 0% of the genetic variance (Akaike, 1981), but this may result from an imprecise estimate due to a small sample size, as in our study.

Computing correlations between the GRM that was constructed with rare variants and with the GRM constructed with other MAF class variants suggested that the GRM that were constructed with common variants captured at least some of the variance that was captured by the GRM built with rare variants (Table 7.4). Table 7.5 shows the comparison between expected and estimated variance explained by each MAF class for LEG. The differences between estimated and expected variances for the rare and low-frequency variants for LEG were large (0.137 and −0.125 for expected variances using VanRaden's methods 1 and 2, respectively) and the estimated variance was actually intermediate to the expected variances obtained with the two VanRaden methods (Hayes et al., 2009). The difference between expected variances with the two VanRaden methods was much larger for rare and low-frequency variants than for common variants. Thus, it might be necessary to correct the current model (two VanRaden's methods), as proposed by Speed et al. (2017); generally, the genomic relationship matrix ($\mathbf{X_{i,j}}$) is calculated as:

$$\mathbf{X}_{\boldsymbol{i,j}} = (\mathbf{S}_{\boldsymbol{i,j}} - 2f_j) \times (2f_j(1 - f_j))^{\alpha/2},$$

where $\mathbf{S_{i,j}}$ is the number of copies of the minor allele carried by individual i at SNP j, f_j is the allele frequency at the SNP j and is commonly set to −1 in human genetics and to 0 in animal and plant genetics (Speed et al., 2017). Speed et al. (2017) found that the optimal was −0.25 for their human data. Our results support the need of exploring the optimal to be used for constructing genomic relationship matrices.

It was previously shown that the contribution of rare variants to phenotypic variance of disease and stature in humans is large (Yang et al., 2015, Mancuso et al., 2016). In dairy cattle, we observed that rare variants play a bigger role for health-related traits than for production traits. Similar to the findings for human height, we also observed that rare variants contributed significantly (the contribution of rare variants for BODY was 0.088) to the body conformation index, for which stature is the main component trait.

In our study, the sequence data that was used to estimate the variance explained by different MAF classes of variants was imputed sequence data. Imputation errors can result in underestimation of the variance explained by rare variants since they

typically have a lower imputation accuracy (Brondum et al., 2014). The average imputation accuracy for rare variants in this study was 0.85, compared to 0.92 for other variants [see Additional file 1: Table S1], which indicates that imputation accuracy may be an important contributor in our study. The 17 traits studied in this analysis are all highly polygenic traits that are affected by a large number of loci. To better study rare variants, next-generation sequencing data from considerably more individuals in the reference population may be useful to improve imputation accuracy and reduce the cut-off threshold for MAF. In addition, the number of animals with phenotypes should be increased to obtain more reliable variance component estimates.

The models used in this study were originally developed to account for LD structure in human data. The LD structure observed from genome-wide loci in cattle differs greatly from that in humans, in that LD persists across much longer ranges and the LD scores are much higher in cattle than in humans, see (Yang et al., 2015) and Additional file 5: Figure S1; i.e. the LD score was in most cases higher than 1000 in cattle, while in humans it is lower than 200. Due to close family structures in cattle and the resulting LD structure, correlations between the GRM matrices based on different MAF classes may be higher in bovine than in human data. Figure 7.1a in Lee et al. (2012) shows that the estimated variances were very similar for each human chromosome, regardless of whether all chromosomes were fitted simultaneously or separately. Conversely, Daetwyler et al. (2012) showed that SNPs from a single chromosome can achieve up to 86% of the accuracy for genomic predictions using all (50 k) SNPs. Strong LD and resulting high correlations between effects is probably the main reason why the data did not contain enough information for the model to accurately partition variances by MAF class. Thus, when we partitioned the variants into LD groups, the SE for the estimates of DRP variance explained within each MAF class were large. We showed that the correlations between GRM that were built with common variants were high (more than 0.6), while correlations between GRM that were built with rare variants and common variants were low (ranging from 0.3 to 0.4) (Table 7.4). Therefore, for bovine data, due to the strong LD, the variance explained by a certain MAF class of common variants can also be explained by another class of common variants, but probably less by rare variants.

Table 7.4 Correlations of the off-diagonal elements of the genomic relationship matrix (GRM) built using variants in different classes of MAF.

MAF class of variants used to construct the GRM	0.001-0.01	0.01-0.05	0.05-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5
0.001-0.01	1.000	0.546	0.372	0.339	0.322	0.313	0.310
0.01-0.05		1.000	0.811	0.756	0.723	0.704	0.696
0.05-0.1			1.000	0.911	0.865	0.845	0.835
0.1-0.2				1.000	0.948	0.925	0.915
0.2-0.3					1.000	0.962	0.950
0.3-0.4						1.000	0.968
0.4-0.5							1.000

Table 7.5 Expectations and estimates of the proportion of variance explained by the variants in different MAF classes using imputed sequence data for the feet and legs trait.

	MAF class						
	0.001-0.01	0.01-0.05	0.05-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5
Expectation VR1[a]	0.006	0.065	0.079	0.252	0.210	0.194	0.193
Expectation VR2[a]	0.147	0.186	0.134	0.169	0.132	0.118	0.114
Estimate[b]	0.059	0.149	0.025	0.053	0.299	0.417	0.002

a Expectations of the proportion of variance explained based on the assumption of VanRaden method 1 (VR1) and 2 (VR2); see VanRaden (2008).
b The estimated proportion of DRP variance explained for feet and legs using the GREML-MS method with partitioning imputed sequence variants into seven MAF groups.

7.4.2 Proportion of DRP variance captured by pedigree and 50 k SNPs

We estimated the proportion of variance in DRP explained for 17 traits using different models and different data sources (Table 7.3). Imputed sequence variants explained a higher proportion of the DRP variance than the 50 k SNPs for most traits. However, the increase in variance explained was small (Table 7.3).

For all traits, estimation of DRP variance based on pedigree data explained the largest contribution of the total variance of DRP. This result is in line with other studies that used 50 k SNPs to construct the GRM (Jensen et al., 2012, Haile-Mariam et al., 2013, Roman-Ponce et al., 2014). The DRP were on progeny test bulls with adjustment for non-genetic effects with a pedigree-based model. Because the estimation and de-regression process was based on a pedigree- based model, it is not surprising that the pedigree-based model explained the largest proportion of variance in DRP. In fact, the REML-PED model is expected to yield EBV that are very similar to the EBV that were used to compute the DRP (Calus et al., 2016). For most health-related traits, the proportion of DRP variance estimated from pedigree relationships was small because the reliabilities of EBV for these traits were low.

7.5 Conclusions

Our results show that the 50 k SNP chip can explain most of the genetic variance estimated by using pedigree relationships and even that estimated by using whole-genome sequence. We observed that using high-density SNPs resulted in only a limited increase in the DRP variance explained. As a result, it is necessary to include pedigree information, i.e. polygenic effects, in genomic prediction in dairy cattle to capture variance that is not captured by genomic markers. Our study also showed the relative importance of rare and low-frequency genomic variants for 17 traits in dairy cattle. Although a human study showed that partitioning variants in different MAF and LD groups decreased the bias of heritability estimates, many of our estimates for the bovine population had high SE. To obtain a reliable estimate of the explained DRP variance for different MAF classes, a larger sample size is needed.

7.6 Appendix

Supplementary material can be found in the online version of the published paper or can be directly be accessed via
https://gsejournal.biomedcentral.com/articles/10.1186/s12711-017-0336-z

Zhang, Q., M. P. Calus, B. Guldbrandtsen, M. S. Lund, & G. Sahana. 2017. Contribution of rare and low-frequency whole-genome sequence variants to complex traits variation in dairy cattle. Genetics Selection Evolution 49(1):60.

7.7 Acknowledgement

Qianqian Zhang benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate “EGS-ABG”. This research was also supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (Grant 0603-00519B). Mario Calus acknowledges financial support from the Dutch Ministry of Economic Affairs, Agriculture, and Innovation (public–private partnership “Breed4Food” Code BO-22.04-011-001-ASG-LR).” Funding was provided by Strategiske Forskningsråd (Grant No. 12-132452), Breed4Food (Grant No. BO-22.04-011-001-ASG-LR), EGS-ABG.

References

Akaike, H. 1981. Citation Classic - a New Look at the Statistical-Model Identification. Springer Series in Statistics (Perspectives in Statistics) (51):22-22.

Brondum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. S. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15(1):728.

Browning, B. L. and S. R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194(2):459-471.

Calus, M. P. L., J. Vandenplas, J. ten Napel, and R. F. Veerkamp. 2016. Validation of simultaneous deregression of cow and bull breeding values and derivation of appropriate weights. Journal of Dairy Science 99(8):6403-6419.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brondum, X. P. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46(8):858-865.

Daetwyler, H. D., K. E. Kemper, J. H. J. van der Werf, and B. J. Hayes. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. Journal of Animal Science 90(10):3375-3384.

de los Campos, G., D. Sorensen, and D. Gianola. 2015. Genomic Heritability: What Is It? PLoS genetics 11(5):e1005048.

Dekkers, J. C. M. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. Journal of Animal Breeding and Genetics 124(6):331-341.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau. 2010. VIEWPOINT Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics 11(6):446-450.

Fuchsberger, C., G. R. Abecasis, and D. A. Hinds. 2015. minimac2: faster genotype imputation. Bioinformatics 31(5):782-784.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution 41(1):55.

Gibson, G. 2012. Rare and common variants: twenty arguments. Nature Reviews Genetics 13(2):135-145.

Gonzalez-Recio, O., H. D. Daetwyler, I. M. MacLeod, J. E. Pryce, P. J. Bowman, B. J. Hayes, and M. E. Goddard. 2015. Rare Variants in Transcript and Potential Regulatory Regions Explain a Small Percentage of the Missing Heritability of Complex Traits in Cattle. PLoS One 10(12):e0143945.

Grisart, B., F. Farnir, L. Karim, N. Cambisano, J. J. Kim, A. Kvasz, M. Mni, P. Simon, J. M. Frere, W. Coppieters, and M. Georges. 2004. Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. Proceedings of the National Academy of Sciences of the United States of America 101(8):2398-2403.

Haile-Mariam, M., G. J. Nieuwhof, K. T. Beard, K. V. Konstatinov, and B. J. Hayes. 2013. Comparison of heritabilities of dairy traits in Australian Holstein-Friesian cattle from genomic and pedigree data and implications for genomic evaluations. Journal of Animal Breeding and Genetics 130(1):20-31.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. (vol 91, pg 47, 2009). Genetics Research 91(2):143-143.

Hoglund, J. K., G. Sahana, R. F. Brondum, B. Guldbrandtsen, B. Buitenhuis, and M. S. Lund. 2014. Fine mapping QTL for female fertility on BTA04 and BTA13 in dairy cattle using HD SNP and sequence data. BMC Genomics 15(1):790.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genetics 5(6):e1000529.

Iso-Touru, T., G. Sahana, B. Guldbrandtsen, M. S. Lund, and J. Vilkki. 2016. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genetics 17(1):1.

Jensen J, M. E., Madsen P, Thompson R. 1997. Residual maximum likelihood estimation of (co) variance components in multivariate mixed linear models using average information. Journal of the Indian Society for Probability and Statistics 49:21-236.

Jensen, J., G. S. Su, and P. Madsen. 2012. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genetics 13(1):44.

Kemper, K. E., P. M. Visscher, and M. E. Goddard. 2012. Genetic architecture of body size in mammals. Genome Biology 13(4):244.

Lee, S. H., T. R. DeCandia, S. Ripke, J. Yang, P. F. Sullivan, M. E. Goddard, M. C. Keller, P. M. Visscher, N. R. Wray, S. P. Genome-Wide, I. S. C. ISC, and M. G. Schizophrenia. 2012. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. Nature Genetics 44(7):831-831.

Legarra, A. 2016. Comparing estimates of genetic variance across different relationship models. Theoretical population biology 107:26-30.

MacLeod, I. M., B. J. Hayes, and M. E. Goddard. 2014. The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. Genetics 198(4):1671.

Madsen, P. and J. Jensen. 2006. DMU A Package for Analysing Multivariate Mixed Models. 8th World Congress on Genetics Applied to Livestock Production. 247.

Maher, B. 2008. Personal genomes: The case of the missing heritability. Nature 456(7218):18-21.

Mancuso, N., N. Rohland, K. A. Rand, A. Tandon, A. Allen, D. Quinque, S. Mallick, H. Li, A. Stram, X. Sheng, Z. Kote-Jarai, D. F. Easton, R. A. Eeles, L. Le Marchand, A. Lubwama, D. Stram, S. Watya, D. V. Conti, B. Henderson, C. A. Haiman, B. Pasaniuc, D. Reich, and P. Consortium. 2016. The contribution of rare variation to prostate cancer heritability. Nature Genetics 48(1):30.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F.

C. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-753.

Marchini, J. and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nature Reviews Genetics 11(7):499-511.

Marouli, E. and M. Graff and C. Medina-Gomez and K. S. Lo and A. R. Wood and T. R. Kjaer and R. S. Fine and Y. C. Lu and C. Schurmann and H. M. Highland and S. Rueger and G. Thorleifsson and A. E. Justice and D. Lamparter and K. E. Stirrups and V. Turcot and K. L. Young and T. W. Winkler and T. Esko and T. Karaderi and A. E. Locke and N. G. D. Masca and M. C. Y. Ng, et al. 2017. Rare and low-frequency coding variants alter human adult height. Nature 542(7640):186-190.

McKay, S. D., R. D. Schnabel, B. M. Murdoch, L. K. Matukumalli, J. Aerts, W. Coppieters, D. Crews, E. Dias, C. A. Gill, C. Gao, H. Mannen, P. Stothard, Z. Q. Wang, C. P. Van Tassell, J. L. Williams, J. F. Taylor, and S. S. Moore. 2007. Whole genome linkage disequilibrium maps in cattle. BMC Genetics 8(1):74.

Purcell, S. M., N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, P. F. Sullivan, P. Sklar, D. M. Ruderfer, A. McQuillin, D. W. Morris, C. T. O'Dushlaine, A. Corvin, P. A. Holmans, S. Macgregor, H. Gurling, D. H. R. Blackwood, A. Corvin, N. J. Craddock, M. Gill, C. M. Hultman, G. K. Kirov, P. Lichtenstein, W. J. Muir, M. J. Owen, C. N. Pato, E. M. Scolnick, D. St Clair, N. J. Craddock, P. A. Holmans, N. M. Williams, L. Georgieva, I. Nikolov, N. Norton, H. Williams, D. Toncheva, V. Milanova, C. M. Hultman, P. Lichtenstein, et al. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460(7256):748-752.

Qanbari, S., E. C. G. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A. R. Sharifi, and H. Simianer. 2010. The pattern of linkage disequilibrium in German Holstein cattle. Animal Genetics 41(4):346-356.

Roman-Ponce, S. I., A. B. Samore, M. A. Dolezal, A. Bagnato, and T. H. E. Meuwissen. 2014. Estimates of missing heritability for complex traits in Brown Swiss cattle. Genetics Selection Evolution 46(1):36.

Speed, D., N. Cai, T. U. Consortium, M. Johnson, S. Nejentsev, and D. Balding. 2017. Re-evaluation of SNP heritability in complex human traits. BioRxiv doi: https://doi.org/10.1101/074310.

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

Yang, J., A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. B. Perry, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Magi, A. Metspalu, A. Hamsten, P. K. E. Magnusson, N. L. Pedersen, E.

Ingelsson, N. Soranzo, M. C. Keller, N. R. Wray, M. E. Goddard, P. M. Visscher, and L. C. Study. 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature Genetics 47(10).

Yang, J. A., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. E. Goddard, and P. M. Visscher. 2010. Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 42(7):565-U131.

Yang, J. A., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. American journal of human genetics 88(1):76-82.

Zhang, Q. Q., B. Guldbrandtsen, M. P. L. Calus, M. S. Lund, and G. Sahana. 2016a. Comparison of gene-based rare variant association mapping methods for quantitative traits in a bovine population with complex familial relationships. Genetics Selection Evolution 48(1):60.

Zhang, Q. Q., B. Guldbrandtsen, J. R. Thomasen, M. S. Lund, and G. Sahana. 2016b. Genome-wide association study for longevity with whole-genome sequencing in 3 cattle breeds. Journal of Dairy Science 99(9):7289-7298.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10(4).

8

Impact of rare and low-frequency sequence variants on reliability of genomic prediction in dairy cattle

Qianqian Zhang[1,2], Goutam Sahana[1] , Guosheng Su[1], Bernt Guldbrandtsen[1], Mogens S Lund[1] and Mario Calus[2]

[1] Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, Tjele DK-8830, Denmark; [2] Animal Breeding and Genomics, Wageningen University & Research, Wageningen 6700 AH, The Netherlands.

Abstract

Background: Common single nucleotide polymorphism (SNP) variants included in SNP arrays have been successfully utilized in genomic prediction in dairy cattle. Availability of whole genome sequence data for a large number of cattle and efficient imputation methodologies open a new opportunity to include rare and low-frequency variants (RLFVs) in genomic prediction. The objective of this study was to examine the impact of including RLFVs from whole genome sequence variants, in addition to the common variants on the routinely used 50k SNP array, on the reliability of genomic prediction in dairy cattle for fertility, health and longevity.

Results: RLFVs with a minor allele frequency less than 0.05 were extracted from imputed sequence data and subsequently grouped based on annotations or selected based on the significance of association with the targeted trait. Firstly, all RLFVs were included as an additional genetic component together with the 50k SNP based genomic relationship matrix in the prediction model. The reliability of prediction was improved by 0.011 for fertility, 0.007 for health and decreased by 0.008 for longevity. Secondly, significant RLFVs which are harbored in genes were identified using an association analysis method developed to map RLFVs, and used together with 50k SNPs for genomic prediction. There were marginal improvements in the reliabilities of genomic prediction, i.e. 0.002 for longevity, 0.002 for fertility, and no improvement for health index. Thirdly, the RLFVs with 'medium-to-high impact' (like protein altering variant, downstream gene variant and upstream gene variant) were added to the prediction model. This improved prediction reliability only up to 0.007 for the health index. However, when adding RLFVs with 'high impact' (like missense and protein altering variants), reliability of genomic prediction decreased. In addition to those empirical analyses, a simulation study was performed to evaluate the impact of adding RLFVs in the model on the reliability of prediction, depending on the amount of genetic variance explained by rare causal variants. Three sets of rare causal variants were generated. The first comprised 21,468 RLFVs randomly selected from 7-10 genes per chromosome, accounting for 10% of genetic variance. The second comprised 266 RLFVs randomly selected from 1 gene per chromosome, accounting for 10% of genetic variance. The third comprised 84 RLFVs randomly selected from 8 genes in the third scenario, accounting for 20% of genetic variance. The effects of these simulated causal RLFVs were added to the existing fertility De-regressed Proofs to form the new phenotypic values. The prediction reliabilities improved by 0.021 to 0.068 when the causal RLFVs were included in prediction model.

Conclusions: Using RLFVs from whole genome sequence data had a small impact on the empirical reliability of genomic prediction in dairy cattle. Our simulations revealed that to take advantage of the sequence data, the key is to identify the causal RLFVs.

Key words: Rare variants, Low-frequency variants, Imputed sequence data, Genomic prediction, Dairy cattle

8.1 Background

Due to the developments of sequencing technology, currently a large amount of whole genome sequence data is available in dairy cattle (Daetwyler et al., 2014, Druet et al., 2014). Using this resource, individual animals with SNP chip genotype data can be imputed to whole genome sequence variants (Brondum et al., 2014). It is expected that the causal variants can be better identified with whole genome sequence data, and therefore, can be used to improve the reliability of genomic prediction in dairy cattle (Hayes et al., 2009). Brondum et al. (2015) found that the reliability of genomic prediction was increased by up to 4% for milk yield traits using quantitative trait loci (QTL) derived from whole genome sequence data and the improvement was 0.5% for fertility. When all variants from the whole genome sequence were used in genomic prediction in dairy cattle, as opposed to using only the QTL derived from sequence data, the reliability of prediction did not increase (van Binsbergen et al., 2015).

SNP chips routinely used in genomic prediction in dairy cattle mostly include SNPs with relatively high minor allele frequency (MAF) and therefore are able to tag common variants very well. The ability of those chips to tag rare and low-frequency variants (RLFVs) is however limited, due to differences in allele frequency. There are several indications that RLFVs make an important contribution to genetic variance. For instance, nonsynonymous SNPs are expected to make a more important contribution to genetic variance than synonymous SNPs, and are significantly skewed towards low frequencies in human genome (Cargill et al., 1999). More generally, it has been observed that the rarer variants are significantly more likely to be functional than the more common variants (Zhu et al., 2011). Therefore, including RLFVs in genomic prediction, through (imputed) sequence data, might be a good alternative to increase reliability of genomic prediction using sequence data. This could be especially true for fitness traits, as the alleles deleterious to fitness may have been selected against due to purifying selection and therefore may have a low frequency. So far, it has not been studied if including selected RLFVs from whole genome sequence in addition to the chip data (e.g. 50k SNP chip data) can improve the reliability of genomic prediction in dairy cattle.

There are various ways to select a subset of RLFVs from whole genome sequence to be included in genomic prediction (Neale et al., 2011, Jiang and McPeek, 2014, Van den Berg, 2015, VanRaden, 2015). One approach is to perform genome-wide

association study (GWAS) for RLFVs to select those that are significantly associated with the trait of interest, and then including them in genomic prediction. Some specific methods can be used to map causal RLFVs, as those methods have been shown to be more powerful than commonly used mixed linear models in a simulation study in dairy cattle (Zhang et al., 2016). An alternative approach is to use annotations of variants that predict biological impact of variants in general (B Hayes, Perez-Enciso et al., 2015, MacLeod et al., 2016). It can be hypothesized that RLFVs with annotations of high impact, e.g. protein altering variants, probably have larger effect on phenotypes, and therefore should be included in genomic prediction. Gonzalez-Recio et al. (2015) and our data (Zhang et al., 2017) suggested that the relative contribution of RLFVs to the total genetic variance might be somewhat higher for health-related traits such as fertility, disease susceptibility and longevity, compared to milk production traits. Therefore, in this study, we hypothesize that the reliability of genomic prediction can be increased by including selected subsets of RLFVs for 3 indices related to fitness in dairy cattle namely fertility index, health index and longevity index.

The objective of this study was to test the above hypothesis and to examine the impact of using selected RLFVs from imputed whole genome sequence data on the reliability of genomic prediction in dairy cattle. We also undertook a simulation study to evaluate the (potential) impact of RLFVs on the reliability of genomic prediction, depending on the amount of genetic variance explained by rare causal variants.

8.2 Methods

8.2.1 Phenotypes and genotypes

In total, 6337 Holsteins sires with de-regressed proofs (DRPs) were genotyped using the Illumina BovineSNP50 BeadChip (50k) version 1 or 2 (Illumina Inc., San Diego, CA) (Iso-Touru et al., 2016). The quality parameters used for selection of SNPs were minimum call rates of 85% for individuals and 95% for loci. Marker loci that were monomorphic or deviated from Hardy-Weinberg proportions ($P < 0.00001$) were excluded. The minimal acceptable GC score was 0.60 for SNPs averagely, and individuals with average GC scores below 0.65 were excluded. The number of SNP remaining after quality control was 43,415 on autosomes. The genome positions of the SNPs were taken from the UMD3.1 Bovine genome assembly (Zimin et al., 2009). The 50k genotypes of the 6337 animals were imputed to full genome

sequence data using a two-step approach. Bulls' 50k genotypes were first imputed to a high-density SNP array (HD, 734,077 SNPs) using a multi-breed reference of 3383 animals which were genotyped with the Illumina BovineHD chip (Illumina Inc., San Diego, CA) using IMPUTE2 software (Howie et al., 2009). These imputed HD genotypes were subsequently imputed to the whole genome sequence level with a total number of 22,232,889 variants using a multi-breed reference of in total 1228 animals from run4 of the 1000 bull genomes project (Daetwyler et al., 2014) and Aarhus University (Hoglund et al., 2014). Both the 50k and the whole genome sequence genotypes were pre-phased with BEAGLE v3.3.2 (Browning and Browning, 2013). This imputation step was performed using Minimac2 software (Fuchsberger et al., 2015). The imputed variants were filtered with a MAF>0.001, which removed the SNP with less than 13 copies of the minor allele from the data (6337 animals). The average imputation accuracy ("INFO" from Minimac2) was 0.850 with a standard deviation of 0.233 for rare variants (MAF<0.01) and 0.873 with a standard deviation of 0.215 for low-frequency variants (0.01<MAF<0.05).

The called variants of each genomic site were annotated using ENSEMBL (v.67) databases with Variant Effect Predictor (VEP) (McLaren et al., 2010). Any sites with multiple transcripts resulting in multiple annotations were annotated only once using the by-gene option in VEP (McLaren et al., 2010) and the annotations of the non-reference alleles were classified according to SIFT scores of the variant (Velankar et al., 2013). VEP determines effects of variants (i.e. SNPs, insertions, deletions, CNVs, or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. SIFT predicts the potential effect a non-reference allele has on encoded proteins, and integrates effects of amino acid change, folded structure (predicted or known), and conservation score (Velankar et al., 2013).

Three fitness related traits i.e. fertility, health and longevity were studied. The numbers of bulls with DRPs, imputed sequence data, 50k data and pedigree information were 5043, 4926 and 4673 for fertility, health and longevity respectively. The fertility index includes breeding values for interval from first to last insemination, number of inseminations for heifers and cows and interval from calving to first insemination for cows. Health is the index for diseases other than clinical mastitis, describing genetic potential to resist reproductive, metabolic and feet-and-leg diseases, and health status is based on veterinarians' treatments in the first three lactations. The longevity index is defined as the productive longevity of a bull's daughters. For details regarding the component traits, recording

procedures and models to estimate breeding values for these three indices, see http://www.nordicebv.info/ntm-and-breeding-values.

8.2.2 Selection of RLFVs from imputed sequence data

The RLFVs were defined as imputed sequence variants with a MAF<0.05. We followed 4 strategies to select RLFVs in this study. The first strategy (based on all RLFVs) selected all RFLVs in 23,431 genes including non-coding genes. The second strategy (based on high impact annotations) selected the RLFVs in genes with annotations of 'high impact' were extracted (high impact: frameshift variant, inframe deletion, inframe insertion, missense variant, protein altering variant, start lost, stop gained, stop lost, splice acceptor variant, splice donor variant, splice region variant). The third strategy (medium-to-high impact of annotations) selected variants in the second strategy along with the RLFVs with annotations of 'medium impact' (3 prime UTR variant, 5 prime UTR variant, downstream gene variant, synonymous variant, upstream gene variant). The fourth strategy (based on association mapping) selected RLFVs in all genes that were found to be associated with three traits using the MONSTER software (Jiang and McPeek, 2014). The details for mapping RLFVs using the famSKAT approach (Chen et al., 2013) are described earlier by Zhang et al. (2016). The RLFVs within the genes with p<0.01 from famSKAT approach were selected. Briefly, the MONSTER model is:

$$\mathbf{y} = \mathbf{X\gamma} + \mathbf{Zu} + \mathbf{M\beta} + \mathbf{e},$$

where $\mathbf{y}$ is a vector of phenotypes, $\mathbf{X}$ is a design matrix for fixed covariates including the intercept, $\boldsymbol{\gamma}$ is a vector of unknown covariate effects, $\mathbf{Z}$ is an incidence matrix relating phenotypes to the corresponding random polygenic effect, $\mathbf{u}$ is a vector of random polygenic effects that follows a multivariate normal distribution $N(0, \mathbf{A}\sigma_a^2)$, where $\mathbf{A}$ is the pedigree-based additive genetic relationship matrix and σ_a^2 is the polygenic variance, $\mathbf{e}$ is a vector of random residuals, $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, $\mathbf{M}$ is a $n \times m$ matrix that encodes the genotype at the m tested variant loci and n is the number of individuals with m_{ij} representing the allele dosage (0, 1 or 2) of the minor allele at the *j*-th variant of individual *i*, and $\boldsymbol{\beta}$ is a vector of (possibly correlated) random effects of the m variants, $\boldsymbol{\beta} \sim N(0, \mathbf{R}_{\boldsymbol{\rho}}\sigma_q^2)$, $\mathbf{R}_{\boldsymbol{\rho}} = (1-\rho)\mathbf{I} + \rho\mathbf{11}'$ with $0 \leq \rho \leq 1$. The limiting case $\rho = 0$ correspond to model famSKAT was used in this study.

8.2.3 Simulation of causal RLFVs

Simulations were undertaken to evaluate the (potential) impact of including RLFVs in the prediction model on the reliability of genomic prediction, depending on the

amount of genetic variance explained by rare causal variants. Using simulated data, it is possible to examine the true impact of RLFVs on the reliability of genomic prediction when causal RLFVs are included in genomic prediction.

In these simulations, quantitative trait nucleotides (QTNs) were simulated by randomly drawing RLFVs from genes. Three scenarios were simulated. The first scenario simulated QTNs with small effect ("SQTN"): based on a whole-genome assembly, 7-10 genes were randomly selected per chromosome and all RLFVs from these genes were simulated as QTNs. The total variance explained by these simulated QTNs was 10% of the genetic variance explained by the markers on the 50k chip. The second scenario simulated QTN with medium effect ("MQTN"): one gene was randomly selected per chromosome and all RLFVs within these genes were simulated as QTNs. The total variance explained by these simulated QTNs was 10% of genetic variance explained by the markers on the 50k chip. The third scenario simulated QTNs with large effect ("LQTN"): 8 genes were randomly selected across the whole genome and all RLFVs in these genes were simulated as QTNs. The total variance explained by these simulated QTNs was 20% of genetic variance explained by the markers on the 50k chip. From scenario one to three, the number of simulated QTNs decreased, while the variance explained by each QTN increased. Therefore, we expected a gradual increase in power of detecting simulated rare QTNs from scenario one to three.

Three strategies of selection of RLFVs were applied in this simulation study for all three scenarios of QTN effects. In the first strategy, genotypes of the simulated QTNs were used to compute the second GRM used in prediction model, and thus we assumed that the QTNs were known without error. In the second strategy, the RLFVs were selected based on significance of association mapping (see 4th selection strategy in "Selection of RLFVs from imputed sequence data"). In the third strategy, RFLVs from randomly selected 10 genes per chromosome were added along with the simulated QTNs to construct the second GRM in the model, while none of these variants had a simulated effect. The third strategy mimicked real situations more closely where false positive associations can add noise in the prediction.

The simulated QTNs effects were added to the DRPs only for fertility, instead of all three traits, due to high computational demand in mapping RFLVs using MONSTER. Effects for all rare QTNs were sampled from a normal distribution following N ~ (0,

1). The true breeding value (TBV) for the simulated rare QTNs loci for each individual as a column vector was calculated as:

$$\mathbf{TBV_Q}=\mathbf{M_Q}*\boldsymbol{\alpha},$$

where $\mathbf{M_Q}$ is the genotype matrix including all the rare QTNs loci (one row per animal), and $\boldsymbol{\alpha}$ is the row vector of QTNs effects. Then the $\mathbf{TBV_Q}$ for each individual was scaled such that the variance jointly explained by all the rare QTNs loci was 10% or 20% of the genetic variance explained by the markers on the 50k chip. Finally, the scaled $\mathbf{TBV_Q}$ for each individual was added to the fertility DRP to obtain the simulated phenotypes.

Each simulation scenario was replicated 10 times, and the reported reliabilities and unbiasedness were the average across replicates. Standard errors of the average reliabilities and unbiasedness were calculated as the standard deviation of the results across the 10 replicates divided by $\sqrt{10}$. Due to computational limitations, we only randomly selected one replicate to do rare variants association mapping (the second strategy) using MONSTER and the result from this randomly selected replicate was presented in the results section.

8.2.4 Genomic prediction

The GBLUP model was used to predict genomic breeding values using the following model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z_g g} + \boldsymbol{e}, \qquad (1)$$

Where $\mathbf{y}$ is the vector of phenotypes, $\mathbf{1}$ is a vector of ones, μ is the general mean. is the design matrix which allocates $\mathbf{y}$ to $\mathbf{g}$. The vector $\mathbf{g}$ contains random effects with variance of var(g) = $\mathbf{G}\sigma_g^2$ where $\mathbf{G}$ is the genomic relationship matrix (GRM) calculated following VanRaden method 1 (VanRaden, 2008):

$$\mathbf{G} = \frac{(\mathbf{X} - 2\mathbf{p1}')(\mathbf{X} - 2\mathbf{p1}')}{2\sum_{j=1}^{n} p_j(1-p_j)}$$

where $\mathbf{X}$ is the allele sharing matrix with the number of copies of the second allele. $\mathbf{p}$ is a vector with allele frequencies and $\mathbf{1}$ is a vector of ones. The factor $2\sum_{j=1}^{n} p_j(1-p_j)$ scales $\mathbf{G}$ to be comparable to the pedigree relationship matrix.

An alternative GBLUP model is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z_{g_1} g_1} + \mathbf{Z_{g_2} g_2} + \boldsymbol{e}, \qquad (2)$$

used in the analysis when two GRMs were fitted simultaneously. The symbols in model 2 are the same as in model 1. The vectors $\mathbf{g_1}$ and $\mathbf{g_2}$ are the two random

effects corresponding to two GRMs. Calculations of the two GRMs are the same as in model 1.

Model 1 with the GRM built using 50k genotype data was used as the basic model for comparisons with other approaches. Genomic prediction with RLFVs were done using model 2 where one GRM was based on 50k data and the second GRM was built using the variants selected by one of the 4 strategies as described above. The models were solved using Restricted Maximum Likelihood (REML) implemented in DMU software (Madsen and Jensen, 2006).

For each trait, the 1000 youngest bulls with birth date ranging from 23/07/2005 to 14/01/2009 were used as validation bulls. The rest of the bulls (4043 for fertility, 3926 for health and 3673 for longevity) born before 23/07/2005 were used as training data. The reliability of genomic prediction was measured as the squared correlation between genomic estimated breeding values (GEBV) and DRP divided by the mean reliability of DRP for validation individuals i.e. $r^2_{GEBV} = (cor(GEBV, DRP))^2 / \bar{r}^2_{DRP}$. The unbiasedness of prediction was calculated as the regression coefficient (b) of DRP on GEBV: i.e. *b*=cov(DRP, GEBV)/ var (GEBV).

8.3 Results

8.3.1 Impact of including RLFVs on the reliability of genomic prediction

Number of selected RLFVs for four different strategies ranged from 25,780 to 1,605,553 (Table 8.1). Reliabilities of genomic prediction using models which include RLFVs are shown in Table 8.2. Across the three traits analyzed, the prediction reliability using 50k data was highest for fertility i.e. 0.392; similar to the results presented earlier by Su et al. and Brøndum et al. for the Nordic Holstein population (Su et al., 2014, Brondum et al., 2015). Reliability of prediction using 50k data was used as a base scenario to compare alternative scenarios for each trait. Adding an additional variance component with all the RLFVs across the whole genome along with 50k, prediction reliability increased 0.011 and 0.007 for fertility and health; however, it decreased 0.008 for longevity. When the additional variance component included only the significantly associated RLFVs, the prediction reliabilities improved by 0.002 for fertility and longevity, no improvement, however, has been shown for health. Adding RLFVs with medium-to-high impact annotations, the prediction reliability increased by 0.005 for fertility and 0.007 for

health index, but decreased by 0.014 for longevity. When RLFVs with high impact annotations were added as an additional genetic component, reliability of prediction decreased for all three traits. We observed that bias of prediction for health and longevity tended to reduce when adding RLFVs, regardless of how they were selected, although the differences with 50k results were very small (Table 8.3). The bias of prediction was increased for fertility for all scenarios except when significant RLFVs from association mapping were included (Table 8.3).

Table 8.1. The number of rare and low-frequency variants (RLFVs) selected to be included in the genomic prediction model.

Characteristics	Fertility	Health	Longevity
All RLFVs in the genome	1,585,116[1]	1,605,553[1]	1,598,760[1]
RLFVs in genes with significant association	25,780	45,652	174,499
RLFVs with high impact annotations	25,944	27,619	28,620
RLFVs with medium-to-high impact annotations	495,516	529,341	545,370

[1] These three numbers were different, because the RLFVs were extracted based on the individuals with phenotypes which differed across the three traits (5043, 4926 and 4673 for fertility, health and longevity respectively).

Table 8.2. Reliability of genomic prediction using different marker sets.

Marker sets	Fertility	Health	Longevity
50k SNP array	0.392	0.319	0.285
50k + All RLFVs from genes in the genome	0.403	0.326	0.277
50k + RLFVs in genes with significant association	0.394	0.319	0.287
50k + RLFVs with high impact annotations	0.384	0.315	0.263
50k + RLFVs with medium-to-high impact annotations	0.397	0.326	0.271

Table 8.3. Unbiasedness of genomic prediction in different methods of selection of rare and low-frequency variants (RLFVs).

Methods of selection of RVs	Fertility	Health	Longevity
50k SNP array	0.993	0.902	0.851
50k + All RLFVs from genes in the genome	1.046	0.950	0.939
50k + RLFVs in genes with significant association	1.002	0.909	0.920
50k + RLFVs with high impact annotations	1.038	0.935	0.902
50k + RLFVs with medium-to-high impact annotations	1.040	0.956	0.931

8.4.1 RLFVs simulated as QTNs

To validate the impact of rare causal variants on genomic prediction, we simulated RLFVs as QTNs and re-estimated the reliabilities of prediction using similar strategies to select RLFVs for prediction. One replicate was randomly selected from each scenario of simulation to present here (Table 8.4-8.6), because computational limitations prohibited mapping RLFVs for association before inclusion in the prediction model for all 10 replicates. Results across all 10 replicates for all scenarios, except the one based on association mapping of RLFVs are presented in the Supplementary Material (Table S1-S3).

The number of selected RLFVs in different scenarios from one replicate was presented in Table 8.4 (the average number of included variants across 10 replicates for each scenario is presented in Table S1). The scenario using only 50k marker was used as a base line to compare with using both 50k markers and simulated RLFVs as QTNs. When all simulated RLFVs as QTNs were included, the prediction reliabilities improved from 0.021 for simulation scenario MQTN to 0.068 for simulation scenario SQTN (Table 8.5). Adding randomly selected RLFVs (without simulated effect) along with the simulated RLFVs as QTNs, the reliability decreased from 0.012 in simulation scenario SQTN to 0.032 in simulation scenario LQTN compared with all simulated RLFVs as QTNs included in the prediction (Table 8.5). Across all the simulation scenarios, adding RLFVs from detected significant association mapping improved prediction reliability from 0.002 for scenario SQTN and 0.009 for simulation scenario MQTN (Table 8.5). We observed that bias of prediction reduced when adding RLFVs for scenario SQTN and LQTN (except when adding randomly selected RLFVs (without simulated effect) along with the simulated RLFVs as QTNs in LQTN) and increased for scenario MQTN (Tables 8.6).

Table 8.4. Characteristics for one random replicate of each simulation scenario (RLFVs refers to rare and low-frequency variants and QTNs refers to quantitative trait nucleotides).

Characteristics	SQTN	MQTN	LQTN
Number of genes	7-10 / chrom.	1 / chrom.	8 in total
Genetic variance explained by simulated QTNs	10%	10%	20%
Number of RLFVs simulated as QTNs	21,468	266	84
Number of RLFVs in the genes simulated as QTNs and RLFVs from 10 random selected genes from each chromosome	50,634	22,487	22,743
Number of RLFVs from genes mapped using MONSTER	25,984	81,068	24,087

SQTN corresponds to the scenario with RLFVs in 7-10 genes per chromosome simulated as

causal variants; MQTN corresponds to the scenario with RLFVs in 1 gene per chromosome simulated as causal variants; LQTN corresponds to the scenario with RLFVs in 8 randomly selected genes across the whole genome simulated as causal variants. The simulated total variances for the QTNs in SQTN, MQTN and LQTN were 10%, 10% and 20% of the variance explained by 50k markers for fertility index. Chrom. refers to chromosome.

Table 8.5. Reliabilities of genomic prediction for one random replicate in each simulation scenario and different strategies for selection of rare and low-frequency variants (RLFVs).

Scenarios	SQTN	MQTN	LQTN
50k	0.357	0.445	0.391
50k + All simulated RLFVs as QTNs	0.425	0.466	0.414
50k + RLFVs in the genes simulated as QTNs and RLFVs from 10 random selected genes from each chromosome	0.413	0.453	0.382
50k + RLFVs in genes mapped using MONSTER	0.359	0.454	0.394

SQTN corresponds to the scenario with RLFVs in 7-10 genes per chromosome simulated as causal variants; MQTN corresponds to the scenario with RLFVs in 1 gene per chromosome simulated as causal variants; LQTN corresponds to the scenario with RLFVs in 8 randomly selected genes across the whole genome simulated as causal variants. The simulated total variances for the QTNs in SQTN, MQTN and LQTN were 10%, 10% and 20% of the variance explained by 50k markers for fertility index.

Table 8.6. Unbiasedness of genomic prediction for one random replicate in each simulation scenario and different strategies for selection of rare and low-frequency variants (RLFVs).

Scenarios	SQTN	MQTN	LQTN
50k	0.929	1.012	0.986
50k + All simulated RLFVs as QTNs	1.009	1.034	1.019
50k + RLFVs in the genes simulated as QTNs and RLFVs from 10 random selected genes from each chromosome	0.995	1.036	0.981
50k + RLFVs in genes mapped using MONSTER	0.939	1.039	0.999

SQTN corresponds to the scenario with RLFVs in 7-10 genes per chromosome simulated as causal variants; MQTN corresponds to the scenario with RLFVs in 1 gene per chromosome simulated as causal variants; LQTN corresponds to the scenario with RLFVs in 8 randomly selected genes across the whole genome simulated as causal variants. The simulated total variances for the QTNs in SQTN, MQTN and LQTN were 10%, 10% and 20% of the variance explained by 50k markers for fertility index.

Similar to the observed increase of prediction reliability comparing between using only 50k markers, and both 50k markers and all simulated RLFVs as QTNs for one replicate, the average prediction reliabilities improved from 0.013 for scenario LQTN to 0.025 for scenario MQTN across 10 replicates of each simulation scenario (Table S2). A similar pattern was also observed when comparing between adding randomly selected RLFVs (without simulated effect) along with the simulated RLFVs

as QTNs and using all simulated RLFVs as QTNs included in the prediction for each simulation scenario i.e. reliabilities decreased from 0.009 for scenario SQTN to 0.020 for scenario MQTN (Table S2). Moreover, we did observe that bias of prediction reduced for the 10 replicates for all simulation scenarios compared between using 50k markers and adding RLFVs regardless how they were selected (Table S3).

8.4 Discussion

We expected improvement in the reliability of genomic prediction for fertility, health and longevity by including imputed RLFVs which explained ranging from 13.3% to 24.6% of the DRP variance for these traits in Nordic Holsteins (Zhang et al., 2017). In the current study, however, we observed that including RLFVs in genomic prediction only marginally improved the reliability of prediction for these three fitness related traits compared with the prediction reliability obtained by 50k genotype data, regardless of how those RLFVs were selected. We observed that none of the strategies to select RLFVs improved more than 1.5% of prediction reliability. In addition, improvements of prediction reliabilities were not consistent across all the strategies of selecting RLFVs and prediction reliability even decreased for two scenarios for longevity and one scenario for fertility and health. The imputation accuracy was 0.79 for the high impact annotation variants and 0.81 for medium-to-high impact annotation variants, both being lower than the average for all RLFVs (0.85). The decrease of reliability using the high impact annotated RLFVs, therefore, might be a result of lower imputation accuracy due to on average a lower MAF (the MAF for high impact annotation variants was 0.5% lower than medium-to-high impact variants). Lower MAF levels imply that these high impact annotated RLFVs are often private to a small number of individuals and may add more noise than signal in the model. In addition, RLFVs might have been selected that are not causal for fertility, health and longevity. Other studies also showed that utilizing sequence variants or preselected sequence variants, regardless of their MAF, yielded no or only small improvements in the accuracy of genomic prediction in dairy cattle (Calus et al., 2016, Heidaritabar et al., 2016, Veerkamp et al., 2016). However, we did observe that the bias of prediction slightly reduced when adding RLFVs no matter how they were selected for health and longevity (Table 8.3) and for the simulation (Table S3). This result indicated that selecting of RLFVs and adding them in genomic prediction could reduce bias of prediction,

which was also in line with the findings in Heidaritabar et al., 2016, Veerkamp et al., 2016.

There are several possible explanations for the observed results. Firstly, rare and low-frequency sequence variants have relatively lower call rates (Brondum et al., 2014) and also low imputation reliabilities (Brondum et al., 2014) compared with common variants. These could also reduce the power of detection when mapping RLFVs. Secondly, the contribution of the identified RLFVs to the genomic prediction reliability likely depends on the power of the mapping methods used. Power to identify rare variants is extremely low unless those variants have large effect and a large sample size is used. For example a rare QTL explaining 0.1% of the total genetic variance has no power to be identified with the existing sample sizes in dairy cattle (Zhang et al., 2016). This suggests that the power to detect rare QTLs for the real data of health-related traits was low, and consequently adding selected RLFVs in the prediction models did not only add extra information, but also noise. Thirdly, RLFVs may explain a limited proportion of the trait variance. However, for the traits analyzed here, we showed that RLFVs explained 13.3%, 24.6% and 22.2% of the DRP variance for fertility, health and longevity (Zhang et al., 2017). Similarly, Gonzalez-Recio et al. also reported that RLFVs (MAF<0.01) explained a larger proportion of the total genetic variance for fertility compared to production traits (14%) (Gonzalez-Recio et al., 2015). The common SNPs on the 50k array are, however, probably able to explain some part of the trait variance due to RLFVs, leaving little room for improvement when adding RLFVs to the model. The small effective population size in dairy cattle results in long range of LD across the whole genome and the common variants might be able to capture a part of effect from RLFVs due to co-segregation of QTL and marker alleles because of close family relationship (MacLeod et al., 2014).

We used a simulation to study some of these possible issues with RLFVs. In simulations, we observed that the prediction reliabilities increased by 0.021 for simulation scenario MQTN to 0.068 for simulation scenario SQTN when all simulated RLFVs as QTNs were included (Table 8.5). To better explain the observed reliability in the simulated scenarios, we compared the simulated and estimated variance explained by 50k and simulated RLFVs for all scenarios (Table S4). The higher increase of prediction reliability for MQTN than SQTN is because the 50k SNPs are best able to capture variance from RLFVs for the SQTN and MQTN relative to the LQTN (Table S4). These results suggest that substantial improvement in

prediction reliability could be achieved, provided that the causal RLFVs affecting complex traits can be identified.

8.5 Conclusions

We compared the prediction reliabilities between using 50k markers, and both 50k markers and selected RLFVs with different strategies to examine the impact of RLFVs on the reliability of genomic prediction. The reliability of genomic prediction was marginally improved using a sub-set of RLFVs (selected based on association or annotations) or using all the RLFVs in the genome. There are several possible reasons for this small improvement, including low imputation reliabilities for RLFVs, little or no power to map RLFVs with the existing sample size, and that common variants on the 50k chip are able to capture a large proportion of variance due to RLFVs. However, using simulations we did observe that prediction reliability was improved when known rare QTNs were added as a separate genetic component in the model. This indicates that the prediction reliability using both 50k data and selected RLFVs can be improved largely unless we could identify the causal RLFVs. This study improves our knowledge on the impact of RLFVs from imputed sequence data on genomic prediction in dairy cattle.

8.6 Appendix

Table S1. Characteristics for each simulation scenario across 10 replicates.

Scenarios	SQTN	MQTN	LQTN
Number of genes	7-10 / chrom.	1 / chrom.	8 in total
Genetic variance explained by simulated QTNs	10%	10%	20%
Number of SNPs on 50k array	54,323	54,323	54,323
Number of RLFVs simulated as QTNs	22,212 (4359)	212 (38)	85 (22)
Number of RLFVs in the genes simulated as QTNs and RLFVs from 10 random selected genes from each chromosome	46,036 (5177)	25,806 (3482)	23,369 (5158)

The results were presented as mean (standard deviation). RLFVs refers to rare and low-frequency variants and QTNs refers to quantitative trait nucleotides. Scenario SQTN corresponds to the scenario with RLFVs in 7-10 genes per chromosome simulated as causal variants that in total explained 10% of 50k explained variance; Scenario MQTN corresponds to the scenario with RLFVs in 1 gene per chromosome simulated as causal variants that in total explained 10% of 50k explained variance; Scenario LQTN corresponds to the scenario with RLFVs in 8 randomly selected genes across the whole genome simulated as causal

variants that in total explained 20% of 50k explained variance.

Table S2. Reliabilities of genomic prediction using different marker sets for the traits affected by different rare QTNs sets, averaged over 10 replicates.

Scenarios	SQTN	MQTN	LQTN
Number of genes	7-10 / chrom.	1 / chrom.	8 in total
Genetic variance explained by simulated QTNs	10%	10%	20%
Number of SNPs on 50k array	54,323	54,323	54,323
Number of RLFVs simulated as QTNs	22,212 (4359)	212 (38)	85 (22)
Number of RLFVs in the genes simulated as QTNs and RLFVs from 10 random selected genes from each chromosome	46,036 (5177)	25,806 (3482)	23,369 (5158)

The results were presented as mean (standard error). RLFVs refers to rare and low-frequency variants and QTNs refers to quantitative trait nucleotides. Scenario SQTN corresponds to the scenario with RLFVs in 7-10 genes per chromosome simulated as causal variants that in total explained 10% of 50k explained variance; Scenario MQTN corresponds to the scenario with RLFVs in 1 gene per chromosome simulated as causal variants that in total explained 10% of 50k explained variance; Scenario LQTN corresponds to the scenario with RLFVs in 8 randomly selected genes across the whole genome simulated as causal variants that in total explained 20% of 50k explained variance.

Table S3. Unbiasedness of genomic prediction using different marker sets for the traits affected by different rare QTNs sets, averaged over 10 replicates.

Scenarios	SQTN	MQTN	LQTN
50k	0.968 (0.008)	0.994 (0.017)	0.970 (0.006)
50k + All simulated RLFVs as QTNs	0.998 (0.008)	1.000 (0.008)	0.988 (0.009)
50k + RLFVs in the genes simulated as QTNs and RLFVs from 10 random selected genes from each chromosome	0.995 (0.009)	1.001 (0.010)	0.986 (0.006)

The results were presented as mean (standard error). RLFVs refers to rare and low-frequency variants and QTNs refers to quantitative trait nucleotides. Scenario SQTN corresponds to the scenario with RLFVs in 7-10 genes per chromosome simulated as causal variants that in total explained 10% of 50k explained variance; Scenario MQTN corresponds to the scenario with RLFVs in 1 gene per chromosome simulated as causal variants that in total explained 10% of 50k explained variance; Scenario LQTN corresponds to the scenario with RLFVs in 8 randomly selected genes across the whole genome simulated as causal variants that in total explained 20% of 50k explained variance.

Table S4. The variance explained in the models for one replicate (the same replicate selected for Table 4-6) in each simulation scenario and different strategies for selection of rare and low-frequency variants (RLFVs).

Scenarios	SQTN		MQTN		LQTN	
	50k	RLFVs	50k	RLFVs	50k	RLFVs
Variance simulated	143.5	14.3	143.5	14.3	143.5	14.3
50k	157.1		167.0		156.0	
50k + All simulated RLFVs as QTNs	141.4	14.2	143.9	25.4	145.9	10.8
50k + RLFVs in the genes simulated as QTNs and RLFVs from 10 random selected genes from each chromosome	138.6	15.9	151.2	15.7	143.1	9.5
50k + RLFVs in genes mapped using MONSTER	152.9	3.8	150.0	15.8	151.9	3.7

SQTN corresponds to the scenario with RLFVs in 7-10 genes per chromosome simulated as causal variants; MQTN corresponds to the scenario with RLFVs in 1 gene per chromosome simulated as causal variants; LQTN corresponds to the scenario with RLFVs in 8 randomly selected genes across the whole genome simulated as causal variants. The simulated total variances for the QTNs in SQTN, MQTN and LQTN were 10%, 10% and 20% of the variance explained by 50k markers for fertility index.

8.7 Acknowledgement

We are grateful to the Nordic Cattle Genetic Evaluation (NAV, Aarhus, Denmark) for providing the phenotypic data used in this study and Viking Genetics (Randers, Denmark) for providing samples for genotyping. Qianqian Zhang benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG". This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B). Mario Calus acknowledges financial support from the Dutch Ministry of Economic Affairs, Agriculture, and Innovation (Public-private partnership “Breed4Food” code BO-22.04-011-001-ASG-LR).

References

Hayes, B., A. C., H Daetwyler, CJ Vander Jagt, ME Goddard. 0415 Improving genomic selection across breeds and across generations with functional annotation. Journal of Animal Science 94(5):201-201.

Brondum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. S. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15(1):728.

Brondum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. Journal of Dairy Science 98(6):4107-4116.

Browning, B. L. and S. R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194(2):459-471.

Calus, M. P. L., A. C. Bouwman, C. Schrooten, and R. F. Veerkamp. 2016. Efficient genomic prediction based on whole-genome sequence data using split-and-merge Bayesian variable selection. Genetics Selection Evolution 48(1):49.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C. R. Lane, E. P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G. Q. Daley, and E. S. Lander. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nature Genetics 23(3):373-373.

Chen, H., J. B. Meigs, and J. Dupuis. 2013. Sequence Kernel Association Test for Quantitative Traits in Family Samples. Genetic Epidemiology 37(2):196-204.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brondum, X. P. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46(8):858-865.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112(1):39-47.

Fuchsberger, C., G. R. Abecasis, and D. A. Hinds. 2015. minimac2: faster genotype imputation. Bioinformatics 31(5):782-784.

Gonzalez-Recio, O., H. D. Daetwyler, I. M. MacLeod, J. E. Pryce, P. J. Bowman, B. J. Hayes, and M. E. Goddard. 2015. Rare Variants in Transcript and Potential Regulatory Regions Explain a Small Percentage of the Missing Heritability of Complex Traits in Cattle. PLoS One 10(12):e0143945.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. (vol 91, pg 47, 2009). Genetics Research 91(2):143-143.

Heidaritabar, M., M. P. L. Calus, H. J. Megens, A. Vereijken, M. A. M. Groenen, and J. W. M. Bastiaansen. 2016. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics 133(3):167-179.

Hoglund, J. K., G. Sahana, R. F. Brondum, B. Guldbrandtsen, B. Buitenhuis, and M. S. Lund. 2014. Fine mapping QTL for female fertility on BTA04 and BTA13 in dairy cattle using HD SNP and sequence data. BMC Genomics 15.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genetics 5(6):e1000529.

Iso-Touru, T., G. Sahana, B. Guldbrandtsen, M. S. Lund, and J. Vilkki. 2016. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genetics 17(1):1.

Jiang, D. and M. S. McPeek. 2014. Robust Rare Variant Association Testing for Quantitative Traits in Samples With Related Individuals. Genetic Epidemiology 38(1):10-20.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes, and M. E. Goddard. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17(1):144.

MacLeod, I. M., B. J. Hayes, and M. E. Goddard. 2014. The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. Genetics 198(4):1671.

Madsen, P. and J. Jensen. 2006. DMU A Package for Analysing Multivariate Mixed Models. 8th World Congress on Genetics Applied to Livestock Production. 247.

McLaren, W., B. Pritchard, D. Rios, Y. A. Chen, P. Flicek, and F. Cunningham. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26(16):2069-2070.

Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly. 2011. Testing for an Unusual Distribution of Rare Variants. PLoS genetics 7(3): e1001322.

Perez-Enciso, M., J. C. Rincon, and A. Legarra. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genetics Selection Evolution 47(1):43.

Su, G., O. F. Christensen, L. Janss, and M. S. Lund. 2014. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. Journal of Dairy Science 97(10):6547-6559.

van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genetics Selection Evolution 47(1):71.

Van den Berg, I. 2015. The use of whole sequence data for genomic selection in dairy cattle. . PhD thesis, AgroParisTech, Paris; Aarhus University, Aarhus, Paris.

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

VanRaden, P. M. a. J. R. O. C. V., P. M. and J. R. O'Connell. . 2015. Strategies to choose from millions of imputed sequence variants. Interbull Bulletin 10-13.

Veerkamp, R. F., A. C. Bouwman, C. Schrooten, and M. P. Calus. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. Genetics Selection Evolution 48(1):95.

Velankar, S., J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M. J. Martin, and G. J. Kleywegt. 2013. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Research 41(D1):D483-D489.

Zhang, Q., M. P. L. Calus, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2017. Contribution of rare and low-frequency whole-genome sequence variants to complex traits variation in dairy cattle. Genetics Selection Evolution 49(1):60.

Zhang, Q. Q., B. Guldbrandtsen, M. P. L. Calus, M. S. Lund, and G. Sahana. 2016. Comparison of gene-based rare variant association mapping methods for quantitative traits in a bovine population with complex familial relationships. Genetics Selection Evolution 48(1):60.

Zhu, Q. Q., D. L. Ge, J. M. Maia, M. F. Zhu, S. Petrovski, S. P. Dickson, E. L. Heinzen, K. V. Shianna, and D. B. Goldstein. 2011. A Genome-wide Comparison of the Functional Properties of Rare and Common Genetic Variants in Humans. American journal of human genetics 88(4):458-468.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biology 10(4):R42.

9

General discussion

9.1 Introduction

Although animal breeding has achieved spectacular success using quantitative genetics, selection decisions still mostly rely on black box approaches and underlying mechanisms need to be better understood to reach a new milestone in animal breeding. To achieve this, the first task is to generate relevant data, for example, whole genome sequencing data and extract useful information from the data combining with existing biological knowledge to understand the biology mechanism behind it. Currently, large numbers of individuals from multiple populations are sequenced for unraveling genetic mechanisms underlying complex traits in animal breeding (Daetwyler et al., 2014). More and more omics data will soon be available to better study the biological mechanisms underlying phenotypic variation among animals, such as DNA methylation signature, RNA expression profiles, metabolite profiles etc. Utilizing and integrating these omics data in genomic evaluation will be an important and promising area in animal breeding in the future.

This thesis exploits whole genome sequence variants to contribute on understanding the genetic mechanisms underlying cattle breeding. To achieve this, the first step is to understand how the landscape of genetic variation is shaped by different evolutionary forces including selective breeding. The second step is to dig into the extra information from whole genome sequence variants, not exploited by the routinely used SNP chip data, i.e. utilizing rare variants. This thesis presents and investigates the potential of different ways of utilize whole genome sequence variants, both common and rare, in cattle breeding to demonstrate the distribution of genetic variants shaped by evolutionary process with emphasis on selection and the role of rare variants in genomic evaluation. The knowledge provided in this thesis could be exploited to identify associated variants as well as improving genomic prediction with better utilization of whole genome sequence variants. The methods used in exploiting the use of whole genome sequence variants can be useful in studying other livestock species. In this chapter, I discuss the main findings in this thesis, explore the added value from this thesis compared to the existing literature, and describe the perspectives to further improve the existing knowledge and implementation of the obtained knowledge in practice.

9.2 Inbreeding and introgression inferred from whole genome sequence variants

With the advent of whole genome sequence data, it is possible to characterize the genetic variation quantifying genetic diversity at the resolution of single base pair in individuals. This thesis uses whole genome sequence variants to demonstrate the distribution of genetic variation which reflects inbreeding and introgression in cattle.

9.2.1 Estimation of inbreeding

Managing the level of inbreeding in the population has long been an important objective in breeding programs to maintain genetic diversity and to reach long-term breeding goals. Pedigree information has been used to estimate inbreeding levels before genomic information was available (Blackwell et al., 1995, Cassell et al., 2003, Mc Parland et al., 2007). However, precision of inbreeding estimated using pedigree information depends on the depth and accuracy of pedigree records (Cassell et al., 2003). In addition, pedigree information captures the expected proportion that genomes that is identity-by-descent (IBD), but cannot capture the Mendelian sampling variance and linkage during gamete formation on IBD relationships (Franklin, 1977, Hill, 1993, Guo, 1996, Hill and Weir, 2011).

The availability of genomic information, e.g. used for genetic evaluations, allows estimation of inbreeding levels and management of genetic diversity across the entire genome. In chapter 2, I estimated the region-specific genetic diversity across the whole genome for multiple individuals and studied the genome-wide scale of homozygosity across individual genomes in different cattle breeds. The ROH, stretches of homozygous genomic regions, constitutes genomic inbreeding of an individual. The lengths of ROH regions reflect the number of generations since the last common ancestor (Fisher, 1954), and correlate with the changing of effective population size over the population history. Long ROH are generated in the first generation since the last common ancestor. These long ROH are broken down by recombination in the following generations. Several generations later, the long ROH become short ROH accumulating in the genome. While long ROH are more related to recent inbreeding, short ROH are more due to selection. Therefore, the length of ROH reflects the number of generations tracing back to the common ancestor. Specifically, following the calculation of Fisher's theory and assumption, a ROH region with length of 2.5 Mb, 10 and 25 Mb corresponds with a common ancestor from 20, 5 and 2 meiosis or generations ago, respectively (Fisher, 1954, Howard et al., 2017). After identification of ROH regions, the realized proportion of

these ROH regions is summed across the genome as an estimate of the inbreeding coefficient.

Different methods and data sources to estimate inbreeding coefficient were compared in chapter 3. The estimators used were Genomic Relationship Matrix (GRM) based inbreeding coefficients, based on excess of homozygosity and correlation of uniting gametes. The ROH based estimator is more stable and directly reflects the genomic regions related to inbreeding. The ROH based estimators using different density of markers including whole genome sequence variants are also highly correlated due to the overlap of ROH detected. However, the estimators based on GRM, excess of homozygosity and correlation of uniting gametes, which are dependent on allele frequencies in the base population (VanRaden et al., 2013), had a poor correlation with ROH and pedigree based estimators. In chapter 3, I examined the impact of allele frequencies in different cattle breeds on these estimators based on GRM, excess of homozygosity and correlation of uniting gametes. The expectations of these estimators for a homozygote or heterozygote show that the estimates can be very different for different allele frequencies. These expectations are related to the actual allele frequencies distribution in these cattle breeds. The estimation of allele frequencies in the base population has a high fluctuation using a small population. Therefore, these estimators based on GRM, excess of homozygosity and correlation of uniting gametes are not recommended for estimation of inbreeding coefficients especially in small populations. Instead, it is recommended to compute inbreeding based on the ROH estimator using either SNP chip or sequence data.

The density of marker used to detect ROH affects the accuracy of estimating ROH and length of ROH which can be estimated. The ROH detected from SNP chip data tends to be longer since inter-marker distance put a limit on how small ROH can be identified. However, it is notable that the long stretch of ROH detected from SNP chip data might not be completely homozygous (Ramey et al., 2013). By whole genome sequence data, the short ROH which probably contains important information about selection other than inbreeding can also be detected (Chapter 2). The ROH estimated from whole genome sequence variants is the most accurate due to the nature of whole genome sequence data.

Knowledge of these specific inbred genomic regions are quiet useful in terms of avoiding the risks of obtaining undesirable inbreeding depressions (Howard et al., 2017). Especially for fitness traits such as health-related, reproduction and survival

traits, managing inbreeding is more important than for production traits because inbreeding depression seems to affect fitness traits more (Charlesworth and Charlesworth, 1987, González-Recio, 2007, Leroy, 2014, Howard et al., 2017). Functional inbreeding is that the genetic variants are likely to be functional and associated with inbreeding depression in different traits. Specific association tests have been developed to examine the inbreeding depression effects of these ROH regions on fitness traits (i.e. homozygosity mapping). Different unfavorable haplotypes with inbreeding effects within a ROH have been identified in different species (Power et al., 2014, Pryce et al., 2014, Howard et al., 2015, Howard et al., 2017). However, it is still very complex to identify the exact ROH haplotypes with unfavorable effects as the ROH regions associated with inbreeding are often containing different haplotypes with variable effects on different phenotypes (Howard et al., 2017). Another challenge of identifying the inbred regions associated with phenotypes is that functional inbreeding is more likely to happen in the loci with a low frequency (Xue et al., 2017). Chapter 6, 7 and 8 deal with identification and utilization of rare and low-frequency variants in cattle. Limited power of detection with current sample size in cattle adds to the complexity to identify rare and low-frequency variants associated with functional inbreeding. In conclusion, there is still a need of developing more efficient methods to identify the functional inbreeding genomic regions, so that animal breeders can put weight on these unfavorable haplotypes for controlling functional inbreeding.

9.2.2 Detection of introgression

Introgression due to gene-flow commonly happens within or between populations in a species, and even between species (Hedrick, 2013). Hybridization and introgression often results in high level of genetic variation in the admixed population (Bosse et al., 2014b). In the long term, introgression generates important sources of genetic variation that are important for adaptation (Grant et al., 2005). With the advent of selective breeding technology, breeds with high production are hybridized to local breeds to improve the production of local breeds (Hartwig et al., 2015, Davis et al., 2017). With hybridization, heterosis is created in the admixed population and is strongest in the first generation. Hybridization is routinely used in producing finishing pigs, broiler and layers (Bichard, 1977, Schneider et al., 1982). The distribution of genetic variation is strongly shaped by the long-term process of introgression (Bosse et al., 2014a, Bosse et al., 2014b). The introgressed haplotypes are subject to various evolutionary forces such as selection and drift, therefore, the introgressed haplotypes remaining in the admixed population are of potential importance for adaption or may yield selective

advantage otherwise. In animal breeding, these retained introgressed haplotypes are worthwhile to examine further because they are potential candidates with beneficial effect on economic traits that are under selection in the admixed population. I used the Nordic Red Dairy cattle (RDC), which has a history of gene flow from Holstein (HOL) and Brown Swiss (BSW) cattle, as an example to demonstrate the pattern of introgressed haplotypes and how some of the (favorable) haplotypes were retained in the admixed population.

In chapter 2, I showed the levels of nucleotide diversity in RDC are highest among the four cattle breeds. It reflects the admixed nature of RDC, therefore, RDC is a good population to study introgression from other cattle breeds. I demonstrated the genome-wide signals of introgressed haplotypes in RDC in chapter 4. The signature of introgression in the genome of RDC coincided with genes and QTLs associated with milk, protein and fat content in milk, calving traits, body confirmation, feed efficiency, carcass and fertility traits. However, we did observe that some introgression haplotypes contain no annotated gene and are located in a genomic region with low recombination rate resulting in long range introgressed haplotypes. This suggests that the retained introgressed haplotypes are a result of interplay between many factors i.e. artificial selection, gene-flow, local recombination rate and genetic drift.

9.3 Effect of selection on distribution of genetic variants

Numerous quantitative trait loci (QTL) have been identified that are associated with different phenotypic traits by genome-wide association studies. However, the estimated effects of these QTLs highly rely on the linkage disequilibrium between causal variants and markers. Estimated QTL effects may vary across populations, different time points, and different sets of markers. To better understand quantitative traits, it is necessary to understand the evolutionary significance of QTLs, i.e. the nature and change of allele frequency of genetic variants by evolutionary and artificial forces (Barton and Keightley, 2002). Detection of evolutionary and artificial signatures on genomes has been receiving continuous attention to improve our understanding of the genetic basis of different traits in human and agricultural species from an evolutionary perspective (Barton and Keightley, 2002).

Chapter 2 showed that artificial selection affects the distribution of functional variants in different length of ROH regions. The haplotypes located in ROH regions which are likely to be IBD. The frequency of a homozygote in IBD (p) is supposed to be higher than the expected frequency of the homozygote of deleterious alleles (p^2) (Szpiech et al., 2013). Consequently, it is observed that deleterious homozygotes are more accumulated in ROH regions than non-ROH regions in both cattle and human genomes (Szpiech et al., 2013). In this case, the variants are annotated as "deleterious" due to their large impact on protein. In cattle populations, some supposedly "deleterious" variants which actually have a beneficial effect on economic traits are selected, thus the frequency of these deleterious variants increase. An example is the myostatin gene mutation in Belgian Blue cattle that results in the double muscling phenotype (Kambadur et al., 1997).

The length of ROH regions is correlated to the number of generations since the last common ancestor. The functional variants in short (<100 Kbp) and medium ROH regions (0.1-3 Mbp) are more likely to be beneficial variants under selection and preserved during several generations while other variants originally in long ROH regions will disappear due to recombination (Bosse et al., 2012). Specifically, these short or medium ROH regions containing favorable variants are expected to rapidly spread over the population. This has been proven by the observation that short ROH regions significantly more often occurred in genomic regions putatively under selection (Chapter 2, Figure S10 and S11). In contrast, the variants in long ROH regions are behaving more or less neutral reflecting recent inbreeding. Interestingly, this pattern observed in cattle populations is opposite to the pattern observed in human where deleterious variants are more enriched in long ROH regions (Szpiech et al., 2013). This is mostly likely related to strong artificial selection in cattle in contrast to natural selection in human. Similar analysis in other livestock species under selection will be helpful to support this hypothesis. It demonstrates that extrapolating the results from human to livestock species should be done with caution.

Chapter 4 showed that the pattern of introgressed haplotypes was also highly shaped by artificial selection in admixed cattle population. When comparing between New Danish Red (New-RED) and Holstein (HOL) or Brown Swiss (BSW), the distribution of introgressed haplotypes under selection from HOL and BSW in New-RED was different with respect to the different stages of selection. The haplotypes under ongoing selection remain diverse whereas haplotypes containing alleles with highly selective advantage will soon be fixed in the population with intense

selection. The introgressed haplotypes from HOL or BSW were firstly highly correlated with the differentiated sites comparing between Old Danish Red (Old-RED) and New-RED. The introgressed regions from BWS were correlated with regions putatively under selection from iHS (Integrated Haplotype Score) test whereas the introgressed haplotypes from HOL was correlated with shared short ROH regions among individuals. This result reflects that the introgressed haplotypes from HOL or BSW are indeed under selection, however, most introgressed haplotypes from HOL are already fixed while introgressed haplotypes from BSW are still under ongoing selection. This is probably because introgressed haplotypes from HOL contain more haplotypes with higher selective advantage than BSW introgressed haplotypes if the introgression happened from HOL and BSW at the similar time with the same intensity of selection.

In the first half of this thesis (Chapter 2, 3 and 4), I used cattle to demonstrate the effect of selection on the distribution of genetic variants in ROH and introgressed genomic regions in cattle genomes. The shared ROH regions and introgressed regions are candidates for selection, likely to be missed in traditional methods to detect signatures of selection. These studies on ROH and introgressed regions can serve as an example strategy model to detect genomic regions under selection in other livestock species.

9.4 The role of rare variants in genomic evaluation in cattle

Genome-wide association studies have identified large numbers of common variants affecting phenotypes (Duerr et al., 2006, Pryce et al., 2010, Satake et al., 2010, Bolormaa et al., 2011). For example, common whole genome sequence variants have been identified associated with longevity in three cattle breeds in chapter 5. However, the genetic variance cannot be fully explained by these common variants (Manolio et al., 2009, Eichler et al., 2010, Zuk et al., 2012, Bloom et al., 2013). A relatively large proportion of whole genome sequence variants are rare, and rare variants are known to have an important role in the evolution of different species (Pritchard, 2001, Gorlov et al., 2011, Tennessen et al., 2012, Bhatia et al., 2013). Thus, rare variants might play an important role in this so-called missing heritability problem (Maher, 2008, Manolio et al., 2009).

How about the role of rare and low-frequency variants in the context of cattle breeding? Rare and low-frequency variants are typically missed when using SNP chip data, but can be studied using whole genome sequence data. The first question to ask is how much do rare and low-frequency variants contribute to the variation of complex traits in cattle? The relative contribution of variants in different MAF classes for 17 complex traits was estimated in chapter 7 and observed to be small (1.9%-24.6%) compared with common variants. However, the relative contribution of rare and low-frequency variants was larger for health-related traits (average more than 13%) than production traits (average less than 11%) (Chapter 7). Gonzalez-Recio et al. (2015) also observed that 14% of the genetic variance for fertility in cattle was explained by rare and low-frequency variants, suggesting that the genetic architecture of health-related traits in cattle is different from production traits. One hypothesis that can explain these results is that rare and low-frequency variants are usually deleterious variants and expectedly selected against for health-related traits such as diseases (Szpiech et al., 2013, MacLeod et al., 2014). In contrast, rare and low-frequency variants seldom have selective advantage with large effect size and will be selected for in the long term (MacLeod et al., 2014, Marouli et al., 2017). The relative contributions of variants of different MAF classes also reflect that causal variants or the variants with large effect might locate in the MAF class with highest explained variance. For example, the contribution of variants with MAF between 0.2 and 0.3 is the highest for milk, fat and protein yield, and the QTL with largest effect affecting milk, fat and protein yield locating in gene *DGAT1* has a MAF of 0.29 (Chapter 7). In summary, the general trend of contribution of variants in different MAF classes reflects the genetic architecture underlying these complex traits.

Another question to ask is whether rare and low-frequency variants can be utilized in improving selection for complex traits in cattle breeding? Chapter 8 showed that the contribution of rare and low-frequency variants was relatively higher for health-related traits than production traits. Different strategies to select sequence variants for genomic prediction were examined for rare and low-frequency variants in chapter 8. Marginal improvements have been observed in some strategies of selecting rare and low-frequency variants compared with only using 50k data in genomic prediction.

The difficulties in utilization of these rare and low-frequency variants in genomic evaluation are also obvious. Firstly, the models and methods commonly used in genomic prediction might need to be adjusted when including rare and low-

frequency variants in genomic prediction. The variance explained by rare variants is intermediate to the expected variance calculated following VanRaden method 1 (VanRaden, 2008) and 2 for building GRM (Table 7.5 in chapter 7). Moreover, rare variants are more private for individuals in the population; therefore, the off-diagonal elements of a GRM based on rare variants are relatively small compared to those of a GRM based on common variants. The results in this thesis support the need for developing new method or model for genomic prediction especially when including rare and low-frequency variants in genomic prediction. The weights for rare and low-frequency variants should be realized according to how much variance rare and low-frequency variants can explain. Alternative ways of building GRM have been proposed such as Speed et al. (2017), which allows to have different weights for rare and low-frequency variants intermediate between VanRaden method 1 and 2 when build up GRM. Secondly, accurate imputation of rare variants and larger sample size are needed for better characterization of the role of rare variants. The inaccurate imputation of rare variants likely results in underestimation of variance explained by rare variants and missing of true association between rare variants and phenotypes. Next generation sequencing of a large number of individuals might be a solution but with an extra cost. Alternatively, whole exome sequencing on large number of animals than smaller number of animals on whole genome sequencing or better tools for imputing rare variants might be needed. For example, to improve the accuracy of imputation for rare and low frequency variants, more information from pedigree might be helpful. If the next generation sequencing data or exome sequencing data from large number of individuals from different populations are made available, more reliable results such as QTLs including both rare and common variants from meta-GWAS studies can be obtained and utilized in genomic prediction. Thirdly, the LD structure is very different in cattle population than human. The LD in cattle usually persists in a long range (Yang et al., 2015) while the LD between loci is much lower in human population (see figure S1 in chapter 7 and Yang et al., 2015). The high LD in cattle positions the obstacle that it is harder to disentangle the effects between rare, low-frequency and common variants as they are highly correlated with each other due to close pedigree relationships. In human, the explained variance in different MAF classes can be differentiated across the same range of LD simply by classifying variants into different LD groups. In cattle, however, this is not feasible. The long range LD in cattle population also results in partly capturing variance from rare and low-frequency variants in common variants (Chapter 8). Therefore, in cattle a 50k SNP chip is able to capture a large amount of genetic variance and reach a high reliability in prediction, also shown by Daetwyler et al. (2012).

Results of chapter 7 and 8 jointly show that rare variants explain small amount of phenotypic variance and hardly improve reliability in genomic prediction for the short-term selection. The results in chapter 7 and 8 suggest that rare variants have low contribution on improving accuracy of genomic prediction for the breeding values of next generations. However, putting weights on rare and low-frequency variants in genomic selection model are very relevant for long-term genomic improvement in cattle breeding programs. More and more efforts have been put on the utilization of rare variants on long-term genomic improvement in breeding schemes (Eynard et al., 2015, Liu et al., 2015, De Beukelaer et al., 2017). To maximize short-term genomic improvement, lower weights are required to be given to rare variants with small effects in genomic selection model (Bijma, 2012). In evolutionary theory, rare deleterious variants are selected against to keep the frequency low (Gibson, 2012). But for rare variants with beneficial effect, the frequency of these rare variants is expected to increase in the population after years of selection. Therefore, rare variants with beneficial effect are useful for long-term genomic improvement, but might be lost due to short-term selection on the variants with relatively larger effect. Different strategies have been proposed for long-term genomic selection, such as optimal contributions selection (Meuwissen and Sonesson, 1998). To reach a maximum long-term genetic gain with balance between genetic gain and inbreeding, both allele frequency and kinship should be taken into consideration in selection scheme (Bijma, 2012). With increasing attention on the loss of rare variants with beneficial effects for long-term genomic improvement, weighted genomic selection has been proposed, which put more weights on rare variants (Goddard, 2009, Jannink, 2010). Therefore, whole genome sequence variants especially rare and low-frequency variants should be used for genomic prediction to maximize the long-term genetic improvement and monitoring allelic frequency for genetic diversity.

9.5 Concluding remarks

In the thesis, I exploited the utilization of whole genome sequence variants in cattle breeding. This thesis provides insight in how selective breeding shapes the distribution of genetic variants in cattle populations. Whole genome sequence variants in different MAF classes contribute different proportions in the phenotypic variance. Common variants have been identified associated with longevity in different cattle breeds. The role of rare variants in terms of their contribution to

phenotypic variance and improving reliability in genomic prediction in cattle was studied. These results provide valuable insight for understanding how selective breeding in cattle influence the distribution of variants on the genomes, and shed light on the importance and possible utilization of rare variants in genomic evaluation. A new era of animal breeding with integrating different omics data in animal breeding is in the horizon. To optimally use the new information, it is necessary to extract useful and meaningful information from massive data according to the breeding goal. The knowledge gained in this thesis demonstrate the different strategies of extracting useful information and better utilization of the whole genome sequence data in animal breeding.

References

Barton, N. H. and P. D. Keightley. 2002. Understanding quantitative genetic variation. Nature Reviews Genetics 3(1):11-21.

Bhatia, G., N. Patterson, S. Sankararaman, and A. L. Price. 2013. Estimating and interpreting FST: the impact of rare variants. Genome Research 23(9):1514-1521.

Bichard, M. 1977. Economic-Efficiency of Pig Breeding Schemes - Breeding Company View. Livestock Production Science 4(3):245-254.

Bijma, P. 2012. Long-term genomic improvement - new challenges for population genetics. Journal of Animal Breeding and Genetics 129(1):1-2.

Blackwell, B. F., P. D. Doerr, J. M. Reed, and J. R. Walter. 1995. Inbreeding Rate and Effective Population-Size - a Comparison of Estimates from Pedigree Analysis and a Demographic-Model. Biological Conservation 72(3):407-407.

Bloom, J. S., I. M. Ehrenreich, W. T. Loo, T. L. V. Lite, and L. Kruglyak. 2013. Finding the sources of missing heritability in a yeast cross. Nature 494(7436):234-237.

Bolormaa, S., B. J. Hayes, K. Savin, R. Hawken, W. Barendse, P. F. Arthur, R. M. Herd, and M. E. Goddard. 2011. Genome-wide association studies for feedlot and growth traits in cattle. Journal of Animal Science 89(6):1684-1697.

Bosse, M., H. J. Megens, L. A. F. Frantz, O. Madsen, G. Larson, Y. Paudel, N. Duijvesteijn, B. Harlizius, Y. Hagemeijer, R. P. M. A. Crooijmans, and M. A. M. Groenen. 2014a. Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. Nature Communication 5.

Bosse, M., H. J. Megens, O. Madsen, L. A. F. Frantz, Y. Paudel, R. P. M. A. Crooijmans, and M. A. M. Groenen. 2014b. Untangling the hybrid nature of modern pig genomes: a mosaic derived from biogeographically distinct and highly divergent Sus scrofa populations. Molecular Ecology 23(16):4089-4102.

Bosse, M., H. J. Megens, O. Madsen, Y. Paudel, L. A. Frantz, L. B. Schook, R. P. Crooijmans, and M. A. Groenen. 2012. Regions of homozygosity in the porcine

genome: consequence of demography and the recombination landscape. PLoS Genetics 8(11):e1003100.

Cassell, B. G., V. Adamec, and R. E. Pearson. 2003. Effect of incomplete pedigrees on estimates of inbreeding and inbreeding depression for days to first service and summit milk yield in Holsteins and Jerseys. Journal of Dairy Science 86(9):2967-2976.

Charlesworth, D. and B. Charlesworth. 1987. Inbreeding Depression and Its Evolutionary Consequences. Annual Review of Ecology and Systematics 18:237-268.

Daetwyler, H. D., K. E. Kemper, J. H. J. van der Werf, and B. J. Hayes. 2012. Components of the accuracy of genomic prediction in a multi-breed sheep population. Journal of Animal Science 90(10):3375-3384.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. Van Binsbergen, R. F. Brondum, X. P. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nature Genetics 46:858–865.

Davis, S. R., R. J. Spelman, and M. D. Littlejohn. 2017. BREEDING AND GENETICS SYMPOSIUM: Breeding heat tolerant dairy cattle: the case for introgression of the "slick" prolactin receptor variant into Bos taurus dairy breeds. Journal of Animal Science 95(4):1788-1800.

De Beukelaer, H., Y. Badke, V. Fack, and G. De Meyer. 2017. Moving Beyond Managing Realized Genomic Relationship in Long-Term Genomic Selection. Genetics 206(2):1127-1138.

Duerr, R. H., K. D. Taylor, S. R. Brant, J. D. Rioux, M. S. Silverberg, M. J. Daly, A. H. Steinhart, C. Abraham, M. Regueiro, A. Griffiths, T. Dassopoulos, A. Bitton, H. Y. Yang, S. Targan, L. W. Datta, E. O. Kistner, L. P. Schumm, A. T. Lee, P. K. Gregersen, M. M. Barmada, J. I. Rotter, D. L. Nicolae, and J. H. Cho. 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314(5804):1461-1463.

Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau. 2010. VIEWPOINT Missing heritability and strategies for finding the underlying causes of complex disease. Nature Reviews Genetics 11(6):446-450.

Eynard, S. E., J. J. Windig, G. Leroy, R. van Binsbergen, and M. P. L. Calus. 2015. The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genetics 16(1):24.

Fisher, R. A. 1954. A Fuller Theory of Junctions in Inbreeding. Heredity 8(2):187-197.

Franklin, I. R. 1977. The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. Theoretical population biology 11(1):60-80.

Gibson, G. 2012. Rare and common variants: twenty arguments. Nature Reviews Genetics 13(2):135-145.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136(2):245-257.

Gonzalez-Recio, O., H. D. Daetwyler, I. M. MacLeod, J. E. Pryce, P. J. Bowman, B. J. Hayes, and M. E. Goddard. 2015. Rare Variants in Transcript and Potential Regulatory Regions Explain a Small Percentage of the Missing Heritability of Complex Traits in Cattle. PLoS One 10(12):e0143945.

González-Recio, O., López de Maturana, E., & Gutiérrez, J. P. 2007. Inbreeding depression on female fertility and calving ease in Spanish dairy cattle. Journal of Dairy Science 90(12):5744-5752.

Gorlov, I. P., O. Y. Gorlova, M. L. Frazier, M. R. Spitz, and C. I. Amos. 2011. Evolutionary evidence of the effect of rare variants on disease etiology. Clinical genetics 79(3):199-206.

Grant, P. R., B. R. Grant, and K. Petren. 2005. Hybridization in the recent past. The American Naturalist 166(1):56-67.

Guo, S. W. 1996. Variation in genetic identity among relatives. Human heredity 46(2):61-70.

Hartwig, S., R. Wellmann, R. Emmerling, H. Hamann, and J. Bennewitz. 2015. Short communication: Importance of introgression for milk traits in the German Vorderwald and Hinterwald cattle. Journal of Dairy Science 98(3):2033-2038.

Hedrick, P. W. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular Ecology 22(18):4606-4618.

Hill, W. G. 1993. Variation in Genetic Identity within Kinships. Heredity 71:652-653.

Hill, W. G. and B. S. Weir. 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genetics Research 93(1):47-64.

Howard, J. T., M. Haile-Mariam, J. E. Pryce, and C. Maltecca. 2015. 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(1):813.

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. Journal of Dairy Science 100(8):6009-6024.

Jannink, J. L. 2010. Dynamics of long-term genomic selection. Genetics Selection Evolution 42(1):35.

Kambadur, R., M. Sharma, T. P. L. Smith, and J. J. Bass. 1997. Mutations in myostatin (GDF8) in double-muscled Belgian blue and Piedmontese cattle. Genome Research 7(9):910-916.

Leroy, G. 2014. Inbreeding depression in livestock species: review and meta-analysis. Animal genetics 45(5):618-628.

Liu, H. M., T. H. E. Meuwissen, A. C. Sorensen, and P. Berg. 2015. Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs. Genetics Selection Evolution 47(1):19.

MacLeod, I. M., B. J. Hayes, and M. E. Goddard. 2014. The Effects of Demography and Long-Term Selection on the Accuracy of Genomic Prediction with Sequence Data. Genetics 198(4):1671.

Maher, B. 2008. Personal genomes: The case of the missing heritability. Nature 456(7218):18-21.

Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature 461(7265):747-753.

Marouli, E. and M. Graff and C. Medina-Gomez and K. S. Lo and A. R. Wood and T. R. Kjaer and R. S. Fine and Y. C. Lu and C. Schurmann and H. M. Highland and S. Rueger and G. Thorleifsson and A. E. Justice and D. Lamparter and K. E. Stirrups and V. Turcot and K. L. Young and T. W. Winkler and T. Esko and T. Karaderi and A. E. Locke and N. G. D. Masca and M. C. Y. Ng and P. Mudgal and M. A. Rivas and S. Vedantam and A. Mahajan and X. Q. Guo, et al. 2017. Rare and low-frequency coding variants alter human adult height. Nature 542(7640):186-190.

Mc Parland, S., J. F. Kearney, M. Rath, and D. P. Berry. 2007. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. Journal of Animal Science 85(2):322-331.

Meuwissen, T. H. E. and A. K. Sonesson. 1998. Maximizing the response of selection with a predefined rate of inbreeding: Overlapping generations. Journal of Animal Science 76(10):2575-2583.

Power, R. A., C. Nagoshi, J. C. DeFries, R. Plomin, and W. T. C. Control. 2014. Genome-wide estimates of inbreeding in unrelated individuals and their association with cognitive ability. European Journal of Human Genetics 22(3):386-390.

Pritchard, J. K. 2001. Are rare variants responsible for susceptibility to complex diseases? American Journal of Human Genetics 69(1):124-137.

Pryce, J. E., S. Bolormaa, A. J. Chamberlain, P. J. Bowman, K. Savin, M. E. Goddard, and B. J. Hayes. 2010. A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. Journal of Dairy Science 93(7):3331-3345.

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

Ramey, H. R., J. E. Decker, S. D. McKay, M. M. Rolf, R. D. Schnabel, and J. F. Taylor. 2013. Detection of selective sweeps in cattle using genome-wide SNP data. BMC Genomics 14(1):382.

Satake, W., I. Mizuta, M. Kubo, T. Kawaguchi, T. Tsunoda, T. Yoshikawa, S. Sakoda, M. Yamamoto, N. Hattori, M. Murata, Y. Nakamura, T. Toda, and J. P. G. Consortium. 2010. Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. Nature Genetics 41(12): 1303-1307.

Schneider, J. F., L. L. Christian, and D. L. Kuhlers. 1982. Crossbreeding in Swine - Genetic-Effects on Pig Growth and Carcass Merit. Journal of Animal Science 54(4):747-756.

Speed, D., N. Cai, T. U. Consortium, M. Johnson, S. Nejentsev, and D. Balding. 2017. Re-evaluation of SNP heritability in complex human traits. BioRxiv doi: https://doi.org/10.1101/074310.

Szpiech, Z. A., J. Xu, T. J. Pemberton, W. Peng, S. Zollner, N. A. Rosenberg, J. Z. Li. 2013. Long runs of homozygosity are enriched for deleterious variation. American Journal of Human Genetics 93(1):90-102.

Tennessen, J. A., A. W. Bigham, T. D. O'Connor, W. Q. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. M. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, B. GO, S. GO, and N. E. S. Project. 2012. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science 337(6090):64-69.

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

VanRaden, P. M., D. J. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. C. H. M. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, and G. A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of Dairy Science 96(1):668-678.

Xue, Y. L., M. Mezzavilla, M. Haber, S. McCarthy, Y. Chen, V. Narasimhan, A. Gilly, Q. Ayub, V. Colonna, L. Southam, C. Finan, A. Massaia, H. Chheda, P. Palta, G. Ritchie, J. Asimit, G. Dedoussis, P. Gasparini, A. Palotie, S. Ripatti, N. Soranzo, D. Toniolo, J. F. Wilson, R. Durbin, C. Tyler-Smith, and E. Zeggini. 2017. Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. Nature Communication 8.

Yang, J., A. Bakshi, Z. Zhu, G. Hemani, A. A. E. Vinkhuyzen, S. H. Lee, M. R. Robinson, J. R. B. Perry, I. M. Nolte, J. V. van Vliet-Ostaptchouk, H. Snieder, T. Esko, L. Milani, R. Magi, A. Metspalu, A. Hamsten, P. K. E. Magnusson, N. L. Pedersen, E. Ingelsson, N. Soranzo, M. C. Keller, N. R. Wray, M. E. Goddard, P. M. Visscher, and L. C. Study. 2015. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. Nature Genetics 47(10):1114.

Zuk, O., E. Hechter, S. R. Sunyaev, and E. S. Lander. 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. Proceedings of the National Academy of Sciences of the United States of America 109(4):1193-1198.

Summary
&
Sammendrag

Summary

The advent of whole genome sequencing technology enables to better explore the genetic architecture of complex traits. Understanding how selective breeding has shaped the genomic architecture and the contribution of genetic variants across the allele frequency spectrum to trait variation potentially enables to better use of genomic tools to select the best animals. I used whole genome sequence variants to demonstrate the genome-wide signals of inbreeding and introgression in different cattle breeds, and how the distribution of genetic variants in these inbred and introgressed regions was shaped by selective breeding. Whole genome sequence variants in different minor allele frequencies (MAF) contribute differently to the trait variation. Common variants account for a large proportion of genetic variance while the contribution of rare variants is largely unknown in cattle. I examined the contribution of rare and low-frequency variants to the variation in different complex traits. Meanwhile, the impact of rare and low-frequency variants on reliability of genomic prediction was explored using different strategies of selecting rare and low-frequency variants. This thesis provides different perspectives on the distribution of genetic variants shaped by selective breeding, and strategies to exploit whole genome sequence variants including the role of rare variants in genomic evaluation in cattle breeding.

In chapter 2, the genetic variation and inbreeding was characterized in four cattle breeds. Runs of homozygosity (ROH) reflect recent and ancient inbreeding depending on the length of ROH. Long ROH reflect recent inbreeding while the sharing of short ROH across individuals is correlated with genomic regions putatively under selection. Artificial selection shapes the distribution of 'deleterious' variants in different length of ROH in cattle genomes. The interplay between long-term selective breeding and inbreeding resulted in that deleterious variants are more enriched in short or medium ROH regions. This pattern was strikingly different from human population where deleterious variants were more enriched in long ROH regions. The findings contribute to the understanding of effects of inbreeding and artificial selection on the distribution of functional genetic variants in ROH regions in cattle populations.

Chapter 3 compared ROH-based inbreeding coefficients with other estimators using different data sources. The other estimators are based on the genomic relationship matrix (GRM), excess of homozygosity and the correlation between uniting gametes, and are all sensitive to allele frequencies in the base population.

The ROH-based estimator directly reflects the genome-wide level of homozygosity and has the advantage of not being affected by the allele frequencies in the base population. The inbreeding coefficient estimated using pedigree was moderately correlated with the ROH-based estimator provided that the pedigree was averagely 8 generations deep. Marker density affected the estimation of ROH; however, ROH-based estimates using different density of markers were highly correlated. The inbreeding coefficient based on ROH using 50k data is shown to be generally close to using whole genome sequence data. However, to obtain the full picture of ROH regions in different length, whole genome sequence data is required.

Chapter 4 used data from Red Danish Cattle (DK-RED) to demonstrate the genome-wide pattern of introgression from the high yielding breeds Holstein (HOL) and Brown Swiss (BSW). Numerous identified introgressed haplotypes were overlapping with genes and QTLs affecting milk production, fat and protein content, calving traits, body confirmation, feed efficiency, carcass and fertility traits. The introgressed haplotypes from HOL and BSW were correlated with signatures of selection in the DK-RED population. The findings clearly show the genomic consequences of selective introgression in modern dairy cattle breed.

Chapter 5 identified whole genome sequence variants associated with longevity in three cattle breeds i.e. Holstein, Jersey and Nordic Red cattle. Most significantly associated SNP were located between genes within unknown annotation or outside coding regions of genes. The high linkage disequilibrium (LD) between causal variants and markers hinders selecting candidate genes within the associated peaks. Identified genomic regions associated with longevity harbor disease genes, including KCNK16, PPP1R14C, and GCH1, which can be studied further for functionality.

Chapter 6 compared different methods for genome-wide association mapping of rare variants using a simulation study. Combining rare variants within a gene into a meta-variant improved the power of detecting associations for rare variants compared with a standard single marker test. When the minor allele frequency is extremely low, high type I error rates for linear mixed models were found in mapping rare variants. The results suggest that methods such as the burden test or variance component tests should be used instead of linear mixed model for mapping rare variants in cattle and other livestock species with a similar family and population structure.

Chapter 7 explored the contribution of variants in different MAF classes to the genetic variation of complex traits. Rare and low-frequency variants contributed relatively more to health-related traits compared with production traits. However, the results suggest that larger samples with both genotype and phenotype are required to be able to draw more general conclusions. Compared with variance explained by 50k SNP markers and pedigree in the trait variation, the gain from whole genome sequence variance was limited in terms of the total amount of variance explained. To explain the variance which is not captured by the genomic markers, it is necessary to include a polygenic component based on pedigree information in genomic prediction in dairy cattle.

Chapter 8 investigated the impact of rare and low-frequency variants on reliability of genomic prediction. The reliabilities of genomic prediction were compared between using 50k markers, and both 50k markers and rare and low-frequency variants (RLFV) selected from sequence data using different strategies. The reliability of genomic prediction was only marginally improved using selected rare and low-frequency variants based on associations and annotations or using all rare and low-frequency variants. There are several reasons for this result including low reliabilities for RLFVs, little power to map RLFVs with the existing sample size, and that common variants on the 50k chip are able to capture a large proportion of variance due to RLFVs. However, the results in simulation indicate that the reliability of genomic prediction using 50k data and selected rare and low-frequency variants can be improved provided that the causal rare and low-frequency variants explain a reasonable proportion of genetic variance and can be identified accurately in advance.

Finally, in the general discussion I put the findings in this thesis in a broader perspective by discussing the main results and the contribution of these results to the existing literature. Different perspectives on the utilization of whole genome sequence variants in cattle breeding have been discussed. The topics include genomic inbreeding and introgression inferred from whole genome sequence, the effect of selection on the distribution of genetic variants in inbred and introgressed regions, the role of rare variants in genomic evaluation and the concluding remarks.

Sammendrag

Helgenomsekventeringsteknologiens indtog gør det muligt at udforske komplekse egenskabers genetiske arkitektur. Forståelsen af, hvordan selektiv avl har formet den genomiske arkitektur, og genetiske varianters bidrag på tværs af allelfrekvensspektret til variation i egenskaber, gør det potentielt muligt at bruge genomisk værktøjer til bedre udvalg af de bedste dyr. Jeg har brugt helgenomsekvensvarianter til at finde tegn på indavl og introgression i hele genomet i forskellige kvægracer, og hvordan fordelingen af genetiske varianter i disse indavlede og indkrydsede områder blev formet af selektiv avl. Helgenomsekvensvarianter med forskellige minor allele frequencies (MAF) bidrager forskelligt til variationen i egenskaberne. Almindelige varianter forklarer en stor andel af genetisk varians, mens bidraget fra sjældne varianter stort set er ukendt hos kvæg. Jeg undersøgte sjældne og lavfrekvente varianters bidrag til variationen i forskellige, komplekse egenskaber. Imens blev sjældne og lavfrekvente varianters indvirkning på sikkerheden af genomisk prædiktion undersøgt ved at bruge forskellige strategier til udvalg af sjældne og lavfrekvente varianter. Denne afhandling præsenterer forskellige perspektiver på, hvordan selektiv arv har bidraget til at formefordelingen af genetiske varianter, samt strategier til at udnytte helgenomsekvensvarianter, herunder sjældne varianters, rolle i genomisk evaluering i kvægavl.

I **kapitel 2** karakteriseres mængden af genetiske variation og indavl karakteriseret i fire kvægracer. Runs of homozygosity (ROH) afspejler nyere og ældre indavl afhængigt af længden af ROH. Lange ROH afspejler nyere indavl, mens deling af korte ROH på tværs af individer hænger sammen med genomiske områder, der formodentlig er under selektion. Kunstig selektion former fordelingen af skadelige varianter i af forskellige længder i ROH i kvæggenomer. Samspillet mellem langsigtet, selektiv avl og indavl har resulteret i, at skadelige varianter er ophobet i korte eller mellemlange ROH områder. Dette mønster adskilte sig påfaldende fra den humane population, hvor skadelige varianter var mere ophobede i lange ROH områder. Konklusionerne bidrager til forståelsen af effekterne af indavl og kunstig selektion på fordelingen af funktionelle, genetiske varianter i ROH områder hos kvægpopulationer.

I **Kapitel 3** sammenlignes ROH-baserede estimater af indavlskoefficienter med andre estimater baseret på forskellige datakilder. De andre estimater er baserede på den genomiske slægtskabsmatrix (GRM), overskud af homozygositet og

korrelationen mellem gameter, og de er alle følsomme overfor allelfrekvenser i basispopulationen. Det ROH-baserede estimat afspejler direkte niveauet af homozygositet på tværs af genomet og har den fordel ikke at være påvirket af allelfrekvenserne i basispopulationen. Indavlskoefficienten beregnet ved brug af stamtavle var moderat korreleret med det ROH-baserede estimat, forudsat at afstamninger gennemsnitligt kunne spores 8 generationer tilbage i tiden. Markørtætheden påvirkede beregningen af ROH; derimod var ROH-baserede beregninger, der brugte forskellige markørtætheder indbyrdes højt korrelerede. Indavlskoefficienten baseret på ROH med brug af 50k data viste sig generelt at være tæt på estimater baseret på helgenomsekvensdata. Det er imidlertid nødvendigt at bruge helgenomsekvensdata for at opnå det fulde billede af ROH områder af forskellige længder.

I **Kapitel 4** bruges data fra Rød Dansk Malkerace (DK-RED) til at påvise indkrydsning på tværs af genomet fra de højtydende racer Holstein (HOL) og Brunkvæg (BSW). Adskillige haplotyper indkrydset i DK-RED overlappede med gener og QTL, der påvirker mælkeproduktion samt mælkens fedt- og proteinindhold, kælvningsegenskaber, eksteriøregenskaber, fodereffektivitet, skelet og hunlig frugtbarhed. De indkrydsede haplotyper fra HOL og BSW blev korreleret med selektionssignaturer i DK-RED populationen. Resultaterne viser klart de genomiske konsekvenser af selektiv indkrydsning i moderne malkekvæg.
I kapitel 5 identificeres helgenomsekvensvarianter, der er associerede med holdbarhed i tre kvægracer: Holstein, Jersey og Nordisk Rød Malkerace. Væsentligst blev associerede SNP identificeret mellem gener med ukendt annotation, eller udenfor kodende genregioner. Den stærke koblingsuligevægt mellem kausale varianter og markører vanskeliggør udvælgelsen af kandidatgener indenfor de stærkest associerede kromosomområder. Identificerede, genomiske områder associeret med holdbarhed rummer sygdomsgener, inklusive KCNK16, PPP1R14C og GCH1, hvis nærmere funktion bør studeres.

I **kapitel 6** sammenlignes forskellige metoder til associationskortlægning for sjældne varianter på tværs af genomet ved at bruge af et simulationsstudie. Ved at samle sjældne varianter indenfor et gen til en meta-variant øgedes den statistiske styrke til at påvise associationer mellem egenskaber og sjældne varianter sammenlignet med en standard enkeltmarkørtest. Når MAF var ekstremt lav, fandtes der mange type I fejl i inear mixed models til kortlægning af sjældne varianter. Resultaterne viser, at metoder såsom load test eller varianskomponenttests bør bruges i stedet for lineære, blandede modeller til

kortlægning af sjældne varianter hos kvæg og andre husdyrarter med en tilsvarende familie- og populationsstruktur.
I kapitel 7 undersøges varianter i forskellige MAF-klassers bidrag til genetisk variation i til komplekse egenskaber. Sjældne og lavfrekvente varianter bidrog relativt mere til sundhedsrelaterede egenskaber sammenlignet med produktionsegenskaber. resultaterne indikerer imidlertid , at større stikprøver med både genotyper og fænotyper er påkrævede for at kunne drage mere generelle konklusioner. Sammenlignet med den varians i egenskaber, som kan forklares ved hjælp af 50k SNP-markører og slægtskabsinformation, var udbyttet af helgenomsekvensvarians begrænset målt ved den forklarede varians. For at forklare den varians, der ikke forklares af de genomiske markører, må den genomiske prædiktion i kvæg inkludere en polygen effekt baseret på slægtskabssinformation.

I **kapitel 8** undersøges effekten af sjældne og lavfrekvente varianter på sikkerheden af genomisk prædiktion. Den genomiske prædiktions sikkerhed, der opnås ved brug af hhv. 50k-markører, både 50k markører og sjældne og lavfrekvente varianter (RLFV) og udvalgte sekvensdata ved at bruge forskellige strategier, blev sammenlignet. Den genomiske prædiktions sikkerhed blev kun marginalt forbedret ved at bruge udvalgte, sjældne og lavfrekvente varianter baseret på associationer og annotationer, eller ved kun at bruge sjældne og lavfrekvente varianter. Der er flere grunde til dette resultat, herunder lav sikkerhed for RLFVs, lav styrke til at kortlægge RLFVs med den eksisterende stikprøvestørrelse, og at almindelige varianter på 50k-chippen er i stand til at forklare en stor del af den varians, der skyldes RLFVs. Resultaterne i simulationen indikerer imidlertid, at den genomiske prædiktions sikkerhed ved brug af 50k-data og udvalgte, sjældne og lavfrekvente varianter kan forbedres, forudsat at de kausale, sjældne og lavfrekvente varianter forklarer en tilstrækkelig del af den genetiske varians, og kan identificeres præcist på forhånd.

I den afsluttende, overordnede diskussion sætter jeg denne afhandlings resultater i et bredere perspektiv ved at diskutere de vigtigste resultater og deres bidrag til den eksisterende litteratur. Forskellige perspektiver på brugen af helgenomsekvensvarianter i kvægavl bliver diskuteret. Emnerne inkluderer genomisk indavl og indkrydsning detekteret ved hjælp af helgenomsekvensdata, effekten af selektion på fordelingen af genetiske varianter i indavlede og indkrydsede områder, samt sjældne varianters rolle i genomisk evaluering og de afsluttende bemærkninger.

Acknowledgements

Acknowledgement

Although there is only my name on the cover, many people have contributed to the research in their ways and for that, I want to take this opportunity to give all of them my gratitude. Without their support, I may not have gotten to where I am today.

Undoubtedly, first and foremost my thanks would go to my supervisors: Goutam, Mario, Bernt and Mogens.

Goutam, I feel so lucky to have you as my supervisor and you are such a fantastic, inspirational and gracious scientist and supervisor. Thank you very much for your support and help both scientifically and personally during these years of my PhD study. I could not do as good as now I have done without your guidance. I really appreciate the freedom you have given me in the beginning of my PhD so that I could conduct the researches that I am interested in. I am looking forward generating more fruitful results under your guidance later.

Mario, it is simply a great experience to have you as my supervisor! Your guidance and support has gone through the whole journey of my PhD study, although for most of time you are far away from me in Wageningen. You have taught me not only how to solve the scientific problems also enlightened me how to think clearly, insightfully and deeply about the research question and in different perspectives. The ability of how to think I have learned is so important for me to become more independent and even develop my own research idea in the future, and I really appreciate it! I gained not only knowledge also a lot of interest on doing research in genomic prediction under your guidance during my stay in Wageningen. Thank you very much for all nice discussions with you both scientifically and personally.

Bernt, I am one of the luckiest to have you as my co-supervisor, although you have taught me much more than a co-supervisor should have done. I am very grateful to get your generous help on programming in Linux and Perl, and have fruitful discussions with you in population genetics and statistical models. Your help is always there every time when I encounter a technical or scientific problem. I also got a lot of encouragement and inspiration on how to make things 'perfect' from your guidance. Thank you very much for your valuable time to discuss with me.

Mogens, your guidance always sheds light on a broader perspective of my research and it makes me think a lot on how to apply my research results in the practical

breeding. The discussions with you are always enjoyable and insightful. I am very grateful for your support on both scientific and personal matters about difficulties of moving around, saving me from the depression in the end of my PhD. Thank you very much for giving me the opportunity to work in such a great group.

I would like to express my gratitude to those who helped me along my PhD journey. My sincere thanks goes to Mirte, for all the great suggestions for my analysis and always answering my emails quickly; Guosheng, for helping me solve problems in my prediction models; Louise, for your kind help on my different kinds of questions and problems; Elise, thank you for offering the opportunity to join the fantastic "EGS-ABG" program; Jette, Bernt and Pernille, for helping me to translate my summary and resume to Danish; Karin, Tina, Cindie and Birgit in QGG, Lissette in ABGC for helping me deal with the practical issues in my daily work; Huiming and Xiaoping, for teaching me how to simulate data, do genomic prediction, and encourage me during my difficult time; Henk, for your valuable suggestions for my propositions; Mandy and Shuwen, for offering me accommodation during my short stays in Wageningen; Qiuyu, for helping me arrange accommodation before I went to Wageningen.

My sincere thanks would go to all my colleagues in QGG Aarhus University, ABGC Wageningen University and "EGS-ABG" program. You have made my PhD life more colorful and thank you for offering the opportunity to be part of you. Special thanks goes to Hadi, Xiaowei, Wossenie, Naveen, Lei Zhou, Gabriel and Berihu, , for being my office mates, sharing happiness and sadness with me; Lei Wang, Aoxing, Lu, Zhe in QGG and Rodrigo, Shuwen, Mandy, Sanne, Biaty in ABGC, for having lunch together almost every day and you have offered me a lot of happiness.

I would like to acknowledge my Chinese friends in Denmark and the Netherlands. Thank you for accompanying me for this PhD journey and you have made my journey happy and wonderful. Please forgive me for not listing all your names.

Lastly, my deepest thanks goes to my beloved family: my parents, my husband-Hao and my son-Erik (Yangyang). None of these can be achieved without your endless love, understanding and support.

Qianqian Zhang
December, 2017

Curriculum Vitae

About the author

Qianqian Zhang was born on August 30, 1989 in Xianyang, China. She obtained her B.Sc. in Animal Science from Northwest A&F University, China in 2011. During her bachelor, she has developed her interests in animal breeding and genetics by attending research programs such as developing software for calculating breeding values using BLUP, performing lab work on detecting nucleotide polymorphisms, gene cloning and expression in different Chinese cattle populations. After her bachelor study, she got Erasmus Mundus scholarship to participate in European Master of Animal Breeding and Genetics program in 2011. During her Master thesis, she investigated the accuracy of imputation utilizing pedigree relationship information, and signature of selection in European and Asian pigs. She also did internship in University of Veterinary Medicine, Vienna to study the genomic response on cold and hot environments in drosophila experimental evolution lines. These experiences have strengthened her interest in animal breeding and genetics especially in population genetics and statistical genomics. She believes that genomic tools will play an important role in the next generation of animal breeding. In 2013, she obtained her M.Sc. in Animal breeding and genetics from University of Natural Resources and Life Sciences and Wageningen University. She deliberately has chosen to work with next generation sequencing data in her PhD project under Erasmus Mundus scholarship collaborating between Aarhus University and Wageningen University. Her PhD research has focused on exploiting whole genome sequence variants in the context of cattle breeding. The results of this research are described in this thesis. After her PhD, she continues her scientific career in animal breeding and genetics as a post-doctoral researcher.

Peer reviewed publications

1. **Zhang, Q.**, M. P. L. Calus, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2017. Contribution of rare and low-frequency whole-genome sequence variants to complex traits variation in dairy cattle. Genetics Selection Evolution, 49(1): 60.
2. **Zhang, Q.**, B. Guldbrandtsen, M. P. L. Calus, M. S. Lund, and G. Sahana. 2016. Comparison of rare variant association mapping methods for quantitative traits in cattle population with complex familial relationship. Genetics Selection Evolution, 48(1): 60.
3. **Zhang, Q.**, B. Guldbrandtsen, J. R. Thomasen, M. S. Lund, and G. Sahana. 2016. Genome-wide association study for longevity with whole genome sequence in 3 cattle breeds. Journal of Dairy Science, DOI: http://dx.doi.org/10.3168/jds.2015-10697.
4. **Zhang, Q.**, B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. 2015. Runs of homozygosity and distribution of functional variants in the cattle genome. B M C Genomics, 16(1):542.
5. **Zhang, Q.**, M. P. L. Calus, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. B M C Genetics, 16(1):88.

Conference proceedings, abstracts and presentations

1. **Zhang, Q.**, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2016. Rare variants from imputed whole genome sequence explain a small fraction of the total genetic variance in different traits in cattle. 5th International Conference on Quantitative Genetics (ICQG), Madison, United States
2. Sahana, G., **Q. Zhang**, B. Guldbrandtsen, and M. S. Lund. 2016. Rare variants' impact on female fertility in dairy cattle. 5th International Conference on Quantitative Genetics (ICQG), Madison, United States
3. Boitard, S., M. Dolezal, B. Servin, D. Fischer, J. Decker, I. Macleod, **Q. Zhang**, B. Guldbrandtsen, M.S. Lund, A. Bagnato, J. Vilkki, and the 1000 bull genomes project. 2015. Disentangling demography and selection effects of cattle domestication - new insights from the 1000 bull genomes project. Annual Meeting of the Society for Molecular Biology and Evolution (SMBE), Vienna, Austria.
4. Schönherz, A., **Q. Zhang**, V. Nielsen, and B. Guldbrandtsen. Genetic Diversity and Population Structure of the Red Danish Dairy Cattle Breed. 2015.

Annual Meeting of the Society for Molecular Biology and Evolution (SMBE), Vienna, Austria.

5. **Zhang, Q.**, B. Guldbrandtsen, M. Bosse, M. S. Lund, and G. Sahana. 2014. Runs of homozygosity and distribution of functional variants in cattle genome. 10th World Congress on Genetics Applied to Livestock Production (WCGALP), Vancouver, Canada.
6. **Zhang, Q.**, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2014. Inbreeding, admixture and selection signature in Danish Cattle breeds. 34th International Society for Animal Genetics Conference (ISAG), Xi'an, China.

Training and education

Mandatory courses (7 ECTS)	Year
Welcome to EGS-ABG	2013
EGS-ABG Summer Research School	2013
EGS-ABG Summer Research School	2017
Ethics course: Ethics and Philosophy in Life Sciences	2014
Advanced scientific courses (20 ECTS)	
Computing data with R	2014
Quantitative genomics	2014
Genome-wide association and genomic prediction in the era of whole genome sequencing	2014
Introduction to programming for animal science	2016
Advanced quantitative genetics for animal breeding	2014
Professional Skills support courses (8 ECTS)	
Academic English for non-Danish speaking PhD students	2014
QGG research skills	2014
Responsible Research Innovation	2015
How to get published	2016
QGG course in scientific writing	2017
Dissemination of knowledge (21.5 ECTS)	
Teaching	
Teaching assistant in Quantitative genetics course	2013-2015
Teaching assistant in Genetics course	2014
International conferences	
The 10th World Congress on Genetics Applied to Livestock Production (WCGALP)	2014
The 34th conference of the international society of animal genetics (ISAG)	2014
5th International Conference on Quantitative Genetics	2016

Seminars and workshop

Annual Gensap meeting	2014, 2017
Annual meeting of the Danish society for genetic epidemiology	2013
Elixir workshop	2017

Colophon

The research described in this thesis was financially supported by the European Commission within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG", and the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B).

The phenotypic data used in this thesis were provided by the Nordic Cattle Genetic Evaluation (NAV, Aarhus, Denmark) and samples for genotyping were provided by Viking Genetics (Randers, Denmark).

The cover of this thesis was designed by Hao Li and Qianqian Zhang.

This thesis was printed by Digisource, Viborg, Denmark.