ABSTRACT

Title of Dissertation: THE GENETIC ARCHITECTURE OF COMPLEX TRAITS AND DISEASES IN DAIRY CATTLE

Ellen Freebern, Doctor of Philosophy, 2022

Dissertation directed by: Associate Professor Li Ma,
Department of Animal and Avian Sciences

Genetic architecture refers to the number and locations of genes that affect a trait, as well as the magnitude and the relative contributions of their effects. A better understanding of the genetic architecture of complex traits and diseases will be beneficial for analyzing genetic contributions to disease risk and for estimating genetic values of agricultural importance. In particular, genetic and genomic selection in dairy cattle populations has been well established and exploited through genome-wide association studies, sequencing studies, and functional studies. The objective of this dissertation is to understand the genetic architecture of complex traits and apply the understanding to investigate the biological relationship between genetics and diseases in dairy cattle. First, we performed GWAS and fine-mapping analyses on livability and six health traits in Holstein-Friesian cattle. From our analyses, we reported significant associations and candidate genes relevant to cattle health. Second, we evaluated genome-wide diversity in cattle over a period of time by running GWAS and proposed a gene dropping simulation program. From this study, we identified candidate variants under selection that are

associated with biological and economically important traits in cattle. Also, we demonstrated that gene dropping is an applicable method to investigate changes in the cattle genome over time. Third, we investigated the effect of maternal age and temperature on recombination rate in cattle. We provided novel results regarding the plasticity of meiotic recombination in cattle. Additionally, we found a positive correlation between environmental temperature at conception and recombination rate in Holstein-Friesian cows. Collectively, these studies advance our understanding of the genetic architecture and the biological relationship between complex traits and diseases in dairy cattle.

THE GENETIC ARCHITECTURE OF COMPLEX TRAITS AND DISEASES IN DAIRY CATTLE

by

Ellen Freebern

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Associate Professor Li Ma, Chair
Professor Jiuzhou Song
Dr. John Cole
Dr. George Liu
Professor Thomas Kocher

Dedication

To my parents, for their love and encouragement.

Acknowledgements

I would like to thank my advisor, Dr. Li Ma, for your guidance and support through the years. I sincerely appreciate you for welcoming me into your laboratory and for the opportunity to explore the fascinating field of biostatistics. You have been a great mentor and instructor, and I truly have enjoyed working in your laboratory during my years of graduate study.

I would like to thank my Ph.D. dissertation committee for all their time, effort, support, and expertise. It has been a pleasure working with Dr. John Cole and Dr. George Liu on my dissertation research projects. Additionally, I want to thank Dr. Jiuzhou Song for his knowledgeable teachings of bioinformatics and for our discussions on my studies and life that have been a source of support. I also would like to thank Dr. Thomas Kocher for his valuable feedback and suggestions from a biological perspective, which has been very helpful for my doctoral research.

I would like to extend a special thanks to my current lab members, particularly Yahui Gao for answering my questions and for his helpful suggestions during our time together in the lab. In addition, I would like to thank the former lab members, Jicai Jiang, Daniel Santos, and Lingzhao Fang for our research collaborations and great times together.

I would also give my appreciation to my parents for their constant support and encouragement as I pursued my passion for science and research at graduate school.

Finally, I am very grateful to the Department of Animal and Avian Sciences, the Graduate School, and the University of Maryland for providing me the opportunity and financial assistance to pursue my Ph.D. degree. Thank you for allowing me to achieve my goals.

Table of Contents

List of Abbreviations

BFMAP	Bayesian Fine-MAPping
bTB	Bovine Tuberculosis
CNV	Copy Number Variants
DGAT	Diglyceride Acyltransferase
eQTL	Expression Quantitative Trait Locus
GFBLUP	Best Linear Unbiased Prediction Model
GWAS	Genome-Wide Association Study
HAL	Histidine Ammonia-lyse Gene
HD	High-density
IBD	Inflammatory Bowel Disease
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
Mb	mega bases pairs = 1000 kb = 1 million base pairs
MHC	Major Histocompatibility Complex
RHM	Regional Heritability Mapping
SMR	Summary Data-based Mendelian Randomization
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variants

Chapter 1: Literature Review

In this chapter, I review recent developments in statistical methods for genetic analysis, insights into the biological basis underlying complex traits, and current studies on complex traits from genomic data. Lastly, I bring together these concepts to propose the study objectives for this dissertation.

Recent Developments in Statistical Methods for Genetic Analysis

Genome-wide association studies

In recent years, genome-wide association studies (GWAS) have been a successful methodology for identifying single nucleotide polymorphisms (SNPs) associated with common traits and diseases. SNPs are single base pair changes in the DNA sequence, and they are classified into functional/non-synonymous types and neutral/synonymous types, which are used as genetic markers in GWAS (Collins *et al*, 1998). More than 3,600 SNP-trait associations have been identified and summarized in The National Human Genome Institute GWAS catalogue (Hindorff *et al*, 2009).

The statistical power of GWAS to detect associations between variants and traits depends on factors such as the population sample size, the allele frequency and distribution of effect sizes of causal genetic variants in the population, and the linkage disequilibrium (LD) that exists between observed genotyped variants and the unknown causal variants. Once genetic associations have been identified, researchers can use that information to develop better strategies to treat and prevent diseases.

Although GWAS has been successful in explaining the variance in complex diseases and traits, a substantial amount of heritability remains unexplained. Also known as missing heritability, one explanation is that complex traits are highly polygenic and are affected by rare variants (Eichler *et al*, 2010). The minor allele frequency (MAF) of rare variants falls below 0.5%, which makes it more difficult to capture these variants by current GWAS genotyping arrays. Thus, much larger sample sizes are necessary to ensure sufficient power for GWAS to detect these associations. Other explanations for missing heritability include overestimation of SNP effects and unaccounted epistatic effects (Makowsky *et al*, 2011). Several approaches have been developed to account for missing heritability, including the integration of copy number variants (CNVs) into GWAS (Manolio *et al*, 2010). It is important to examine potential sources of missing heritability and address such limitations when conducting GWAS.

Single-variant association tests and meta-analysis

A frequent method that is used when working with multiple GWAS is meta-analysis, which combines the results from each independent study. This approach increases the power to detect SNPs that have small effects on the phenotype, as well as reduces false positive findings. Meta-analysis is particularly useful to synthesize results from previous studies with the aim of drawing conclusions about a collection of research (Haidich, 2010).

Prior to conducting a meta-analysis, certain factors should be considered, including the sample size, definition of the trait or disorder being studied, as well as the statistics used to summarize the association result. In an ideal situation, the GWAS

studies in a meta-analysis will be conducted under the same criteria and study design. However, this may be difficult in practice, such as for a cumulative meta-analysis that may be limited by the availability of data from prior studies.

Additionally, tests for heterogeneity and genomic inflation factors can be applied to correct for the presence of false positives. There exists a tradeoff between power and sample size, which arises when research studies that examine different hypotheses become combined. In this case, an excess of heterogeneity could obscure detection of associations. Thus, careful examination of possible biases is needed prior to interpreting results.

As discussed above, to conduct accurate meta-analysis, the studies must make use of the same research study design and hypothesis. In practice, meta-analysis can also make use of imputed SNP results. However, such results may contain uncertainty due to imputation uncertainty for each SNP. To account for this uncertainty, several approaches have been developed. One approach is to remove SNPs that are identified to have poor imputation quality based on the ratio of the empirical observed variance of the allele dosage to the expected binomial variance held at Hardy-Weinberg equilibrium, p(1-p). Alternatively, the SNPTEST software package utilizes a Bayesian approach to evaluate imputation uncertainty (de Bakker *et al*, 2008). This approach first samples genotypes based on the estimated imputation probabilities, and then takes the average of the resulting Bayes Factors. Although this approach is more computationally involved than classical association tests, an advantage is the inclusion of covariates and its wide availability in standard computing platforms. Additionally, Bayesian hierarchical models allow the incorporation of results from various sources to establish informative priors on

the current meta-analysis study. This allows for the generalization of meta-analysis studies, but it can be computationally complex if there are multiple hypothesis in the GWAS to test. Thus, an empirical Bayesian approach can be applied, which uses the marginal likelihood of the GWAS data and covariates to calculate the probability of association (Lewinger *et al*, 2007).

Imputation

A standard process used in GWAS analysis is imputation, which refers to the statistical inference of unobserved genotypes from a reference population. For instance, imputation can make use of LD information from a population reference panel to estimate missing genotypes. This process can increase the number of testable single nucleotide variants (SNVs) in a genome, which facilitates a greater overall genome coverage of an array. Additionally, imputation has been widely used to detect SNPs, and thus effectively increase the association power in GWAS (Marchini *et al*, 2010).

A key consideration when performing imputation is the availability of reference panels that provide comprehensive information regarding allele frequencies and LD patterns. The importance of this arises when imputing low-frequency and rare variants. These variants are often population-specific and may be influenced by demographic events, which can reduce imputation accuracy (Wojcik *et al*, 2018). Thus, increasing the size of reference panels allows for more reference haplotypes to be captured and error rates to be minimized.

A dense reference panel that has been commonly used is the 1000 Genomes Project in humans. This resource characterizes the common and low-frequency variants

in individuals from diverse populations to analyze genetic contribution to disease (The 1000 Genomes Project Consortium, 2012). More recently, the Haplotype Reference Consortium has been constructed using whole-genome sequence data of individuals from predominantly European descent (The Haplotype Reference Consortium, 2016). The development of dense reference panels such as these will facilitate comprehensive imputation studies on individuals from different populations.

For cattle studies, a routinely used reference panel is the 1000 Bull Genomes Project, which is a database containing whole-genome sequence data from ancestors of current cattle breeds. As of 2018, there have been 84 million SNPs identified by this database (Hayes and Daetwyler, 2018). By capturing a significant proportion of diverse cattle, this project aims to better understand and predict cattle traits that are important for milk and meat production. Additionally, it serves as a resource that contains the annotated sequence variants and genotypes of ancestor bulls. A study conducted in 2017 used whole-genome sequence data from the 1000 Bulls Genomes Project to evaluate the accuracy of imputation in Brown Swiss cattle (Frischknecht *et al*, 2017). The authors selected four different imputation programs in their evaluation: Beagle, FImpute, Impute2, and Minimac. While all methods were adequate, they found that Minimac resulted in the highest imputation accuracy in terms of dosage correlation and genotype concordance. Thus, the authors demonstrated that high accuracy of imputation is possible using the 1000 Bull Genomes reference population.

As discussed above, software selection will influence the performance of genotype imputation. Accurate imputation is necessary to reduce computational costs and provides a better ratio of output to input data. A study in 2015 implemented those

considerations, which lead to the development of a new algorithm for findhap (version 4) software (VanRaden *et al*, 2015). One advantage of this software is its speed, which is derived from the fast imputation algorithm. For instance, findhap can process genotypes from multiple low-density arrays to impute more than 60,000 markers for more than 500,000 dairy cattle in national genomic evaluation systems. The efficiency of findhap is compounded by its low computing costs to sequence data. Efforts to improve accuracy of fast imputation are ongoing, such as the inclusion of high-density (HD) genotypes with low-coverage sequence data. Altogether, these results provide guidance for future imputation studies to consider the effects of software as well as reference selection on imputation performance.

Fine-mapping

GWAS has been widely used to identify loci associated with complex traits, yet the process becomes difficult when LD exists among neighboring SNPs. This poses an issue because the presence of LD often obscures the causal variants driving a GWAS association signal. For instance, a study on udder health in dairy cattle in 2014 aimed to identify SNPs relevant to clinical mastitis (Sahana *et al*, 2014). However, only target regions were detected due to high levels of LD that were present between SNPs.

In this situation, fine-mapping can be applied to determine which of the associated variants are causal. One approach to fine-mapping is known as ranking p-value, where variants are ordered based on the strength of marginal association statistics (Faye *et al*, 2013). A limitation of this method is that p-values are not necessarily a comparable measure for variants to be causal across loci. For instance, a low p-value

associated with a noncausal variant may be due to LD with multiple causal SNPs (Stephens and Balding, 2009). An alternative approach, used when multiple causal variants are present, is the computation of posterior probabilities of causality per SNP in a region. This approach makes use of the likelihoods of observed z-scores that are conditional on each potential set of causal variant(s). The resulting posterior probabilities are used to create the smallest set of SNPs that contain the true causal variant(s) from a given probability. More recently, other fine-mapping methods have been proposed, including CAVIAR, which is a software that has the advantage of only using the marginal test statistics and correlation coefficients among SNPs (Chen *et al*, 2015). Moreover, if the correlation coefficients among SNPs are not present, they can be approximately computed from available reference panels, such as the 1000 Genomes Project for humans.

As stated above, there are multiple fine-mapping strategies available for use. When selecting a strategy, some factors that influence fine-mapping performance should be considered. The factors include the number of causal SNPs in a region, local LD structure, and SNP density. For instance, high SNP density is needed to capture a greater number of causal variants. It may not be feasible to increase SNP density due to sequencing costs for large sample sizes. Alternatively, genotyping using specialized arrays may be used to evaluate SNPs associations, which has the advantage of being cost-efficient and increasing SNP density in known genetic regions. Depending on array design and content, this method may be preferable.

Once a fine-mapping strategy has been selected, the associated SNPs are evaluated for their likely function. This information can be derived using genome

annotation from publicly available databases, including Gene Ontology, ENCODE, FAANG-cattle, and cattle GTEx. Integrating annotation provides functional context to GWAS findings and can be utilized to improve fine-mapping resolution. Generally, genome annotations are categorized as protein-coding and non-protein coding. For protein-coding annotations, existing studies (e.g. CADD) have found improved prediction accuracy when diverse genome annotations are combined to a single quantitative score (Kircher *et al*, 2014). Non-protein coding annotations can be analyzed by annotation tools such as FIRE (Ioannidis *et al*, 2017) and RegulomeDB (Boyle *et al*, 2012), which assigns scores to non-coding variants based on their potential to regulate the expression levels of nearby genes.

After integrating genome annotation into fine-mapping, functional variants can be identified and used in follow-up studies. A limitation to this method is that fine-mapping studies depend on the status of genome annotation, which may impact its detection of causal variants with low MAF. Efforts to improve fine-mapping resolution and genomic annotation are ongoing, including a fast Bayesian Fine-MAPping method (BFMAP) that has been shown to address the issue of high LD present in the cattle genome (Jiang *et al*, 2019). This method efficiently integrates fine-mapping with functional annotations in dairy cattle to identify candidate genes of complex traits. Collectively, the insight garnered from fine-mapping studies provides us with a better understanding of the genetic basis of complex traits that will be useful for future genomic prediction studies.

Biological Basis Underlying Complex Traits

The path from GWAS to biology

GWAS has been shown to be a useful experimental design in assessing the contribution of common variants to disease susceptibility and the associations between genetic variants and traits (Price *et al,* 2015). Yet, the path that links GWAS to biology is not straightforward since the associations detected by GWASs are not directly informative of the target gene or mechanism. Additionally, the effects of gene regulation are often tissue-specific, but the tissues corresponding to diseases may be inaccessible or difficult to precisely study. New analytical methods offer a way to decipher this link, and thus provide biological insights into diseases.

Initiatives, such as the ENCODE project, have supplemented the interpretation of associations of non-coding variants from GWAS. In particular, ENCODE has generated maps of regulatory annotation in disease-relevant tissues, which enhances our understanding of the biological basis of gene expression and regulation in the genome (The ENCODE Project Consortium, 2012). New analytical methods have made use of epigenetic marks and 3D maps of chromatin contacts to elucidate regulatory relationships relevant to complex genetic disorders (Won *et al,* 2016). Additionally, data from GWASs can be integrated with expression quantitative trait locus (eQTL) studies to predict gene targets of complex traits. For instance, a method named summary data-based Mendelian randomization (SMR) utilizes summary data from GWAS and data from eQTL studies to identify pleiotropic associations between gene expression and complex traits (Zhu *et al*, 2016).

Biological mechanisms

GWAS generally aims to detect SNPs in the genome that are relevant to traits or diseases. These SNPs may be located in the coding sequence of a gene, a non-coding region, or an intergenic region. Depending on their location in the functional regions of the genome, SNPs will have a different effect on biological function. In coding regions, synonymous SNPs will not affect an amino acid while nonsynonymous SNPs, including missense mutations and nonsense mutations, will change the resulting protein. Specifically, over 90% of GWAS variants are located in non-coding regions (Maurano *et al*, 2012). Deciphering the mechanism of these mutations will facilitate a greater understanding of how SNPs affect biology. We will review the aforementioned mutations and findings from relevant studies.

Missense mutations result in an incorrect amino acid to be incorporated into a protein. For instance, a study using data from a Holstein-Friesian dairy cattle population identified a missense mutation in the *DGAT1* gene (Grisart *et al*, 2002). This gene is known to encode the enzyme diglyceride acyltransferase (DGAT). The authors identified a *K232A* substitution in *DGAT1*, which was found to have a major effect on milk fat content (Grisart *et al*, 2002).

Nonsense mutations cause a premature stop codon that leads to the shortening of a protein. This mutation can be deleterious since a truncated protein generally will lose its function. For example, a study identified a nonsense mutation in *CWC15* of Jersey cattle (Sonstegard *et al*, 2013). This gene is the bovine protein CWC15 homolog of a spliceosome-associated protein. The nonsense mutation reduces the size of the *CWC15*

protein product by 177 amino acids to a length of 54 amino acids (Sonstegard *et al*, 2013). Effects of this mutation include reduced fertility and reproductive efficiency in Jersey cattle, which may have an economic impact on producers as well.

Synonymous mutations do not alter the amino acid sequence of proteins, but recent studies have determined that they can contribute to changes in protein function (Sauna and Kimchi-Sarfaty, 2011). A study in 2004 found a synonymous SNP located in the corneodesmosin (CDSN*TTC) gene, which carries psoriasis-associated SNPs (Capon *et al*, 2004). Using site-directed mutagenesis, the authors determined that the synonymous mutation confers increased mRNA stability. Thus, it is possible for synonymous SNPs to influence protein expression level by modifications to protein expression. Additionally, Wang *et al* (2014) identified two synonymous mutations in the histidine ammonia-lyse gene (*HAL*). This gene functions to encode histidine ammonia-lysate, which is an enzyme used in histidine catabolism. The *HAL* gene is located within reported QTLs for milk production traits, so the authors wanted to investigate relevant genetic variants in Chinese Holstein cows. They found that the synonymous mutations lead to alterations in codon usage frequency, with SNP (ss974768523) having AAT changed to AAC from 14.7 to 21.4 per thousand, and with SNP (ss974768525) having ATC changed to ATT from 23.3 to 14.6 per thousand (Wang *et al*, 2014). These results highlight the potential of these two synonymous mutations to influence gene expression in *HAL*.

Biological insights from GWAS

Biologically causal genes will typically be located close to the most associated SNP from GWAS. Thus, it is likely that the set of genes near an association will be highly enriched for causal genes. Often times, these genes may not have been expected to be candidate genes, which highlights the novel insights that GWAS findings can provide. Jostins *et al* (2012) conducted an imputed-based association analysis using GWAS data and discovered the previously unsuspected importance of autophagy in Crohn's disease. Given this finding, the study successfully expanded on prior understanding of the pathogenesis of inflammatory bowel disease (IBD) and raised awareness of the fundamental biology of this immune-mediated disorder.

Analyses of GWAS data can also provide insight about functional mechanisms. This has been mediated by studies that seek to identify the functional annotations enriched for the association with diseases. For example, Trynka *et al* (2013) conducted a study that demonstrated GWAS loci exhibit specific cell type enrichments. Their study utilized GWAS data to examine 15 chromatin marks and observed statistically significant phenotypic cell type specificity for H3K4me3 marks, which are known to highlight active gene promoters. Additionally, the authors found that chromatin marks not corresponding to active gene regulation were not phenotypically cell type-specific. Their GWAS analyses showed that complex disease and trait alleles can influence gene regulation in a cell type-specific manner and may also be used to identify plausible causal variants with functional importance.

Elucidating the genetic architecture of complex traits

GWAS findings have been used to help clarify the genomic basis of complex traits and provide insight into the complexity of genetic architecture. Yet, the set of associated variants identified from GWAS often accounts for only a small proportion of the genetic variation in the trait. For instance, SNP chips have been used to detect fertility-related variants in dairy cattle, but are limited by their low coverage of genes and the genome (Ma *et al*, 2019). Research methods have been proposed to better understand the genetic architecture of complex traits.

An important consideration that methods in recent years have accounted for was to assess the contribution of SNPs with effect sizes too small to be detected in GWAS. Raphaka *et al* (2017) applied a relevant approach in a study conducted on a British Holstein-Friesian cattle population. The authors used GWAS and regional heritability mapping (RHM) to assess cattle susceptibility to bovine tuberculosis (bTB). The RHM analyses revealed new genomic regions on bovine chromosome 18 (BTA18) and BTA3 for infected phenotypes that GWAS did not previously identify. In addition, RHM identified regions on BTA23, which supported a previous finding of genomic regions on BTA23 associated with paratuberculosis in Jersey cattle (Zare *et al*, 2014). This approach provides a framework that may be used for future studies on biological pathways that are critical to cattle susceptibility to disease.

Additionally, it has been found that highly associated SNPs typically cluster in biological pathways (Lango *et al*, 2010). Current methods can utilize this information to analyze complex trait phenotypes. For instance, a genomic feature best linear unbiased

prediction model (GFBLUP) has been applied to study mastitis and milk production traits in Holstein and Jersey cattle (Fang *et al*, 2017). This model assumes that all genomic markers have equal contribution to the variability of a particular trait. In this study, GFBLUP improved the accuracy of genomic prediction for mastitis and milk production traits by the inclusion of gene expression data in the analysis. Thus, integrating knowledge of functional genomic regions will facilitate a better understanding of the genetic architecture of complex traits.

Current Studies on Complex Traits from Genomic Data

There has been increasing interest in developing novel analytic methods to study complex traits. Recent studies have made use of publicly available databases, such as the 1000 Genomes Project, to analyze associations between complex traits and rare variants (Auer and Lettre, 2015). Additionally, GWAS has provided promising findings regarding the genetics of complex diseases, and it remains an effective method for investigating genomic data. Current initiatives taken to enhance GWAS analyses include expanding studies to account for more diverse diseases (Gurdasani *et al*, 2019), incorporating more precise phenotypes (Höglund *et al*, 2019), further investigation of the X chromosome (Zhang *et al*, 2020), and capturing a larger proportion of variation in genes of interest (Wang *et al*, 2020). Moreover, the use of updated reference panels of genomic variation has the potential to expand GWAS coverage and improve the detection of associations between disease and common SNPs (Witte, 2010).

Given the knowledge acquired from current studies, initiatives have been taken to better uncover the genetic basis of complex traits. For instance, a review of GWAS of dairy fertility traits found that many cattle GWAS tend to have low power, which underscores the need to increase data collection to enable more powerful studies in dairy cattle (Ma *et al*, 2019). This concept was addressed in a related study that utilized a combination of GWAS and comparative epigenomic analyses to detect large-scale genotype-phenotype associations in Holstein cattle (Liu *et al*, 2020). Using this approach, the authors identified novel tissues and cell types for 45 economically important traits and artificial selection in cattle, as well as tissues for 58 complex traits and diseases in

humans that were correlated in cattle. These findings can provide insight to the underlying molecular mechanisms of complex traits in cattle and support the potential for cattle to be a model for human complex traits.

Another research initiative that is currently underway has studies that are predicting complex trait phenotypes from genotype data. Knowledge from the phenotypic prediction of complex traits may facilitate more targeted disease screening programs based on genetic makeup as well as the understanding of disease mechanisms (Schrodi *et al*, 2014). Generally, predictions are made by selecting genomic variants and using their estimated effect sizes as a predictor. This step can be followed-up by validation using a sample of known phenotypes and then application to samples with unknown phenotypes. One study applied this strategy to a cohort of genotyped individuals from the U.K. Biobank (Canela-Xandri *et al*, 2016). The authors obtained prediction accuracies for four obesity-related traits and for height, which is a trait of particular interest due to its unclear genetic basis (Kaiser, 2020). The prediction accuracies of the traits were reported to be significantly improved compared to prior GWAS meta-analyses of similar size (Canela-Xandri *et al*, 2016). Another study made genomic predictions using three dairy cow traits (coat color, milk-fat percentage, and overall type) to investigate their genetic basis in Holstein cattle (Hayes *et al*, 2010). This work showed that many small effect loci are required to capture the genetic variance of the traits, which suggests that large differences exist in their genetic architecture. Looking forward, the genetic insights that current studies provide offer a promising future for the analysis of complex traits from genomic data.

Study Objectives

The overall objective of this study is to understand the genetic architecture of complex traits and apply the understanding to investigate the biological relationship between genetics and disease in dairy cattle.

This dissertation is organized as follows. **Chapter 2** discusses a GWAS analysis of livability and six health traits in Holstein cattle. We then describe a fine-mapping procedure for those traits, and findings from both analyses are summarized. **Chapter 3** describes an evaluation of genome-wide diversity in cattle over a period of time. We then introduce the application of a gene dropping software to visualize systematic changes from the evaluation. **Chapter 4** introduces a study of meiotic recombination and demonstrates the effect of maternal age and temperature on the recombination rate in cattle. **Chapter 5** summarizes the conclusions of those three projects and discusses future perspectives in these areas, especially in relation to complex traits.

References

1. Collins F, Brooks L, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Research, (1998); 8(12): 1229-31.

2. Hindorff L, Sethupathy P, Junkins H, Ramos E, Mehta J, Collins F, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of National Academy of Sciences of the United States of America. (2009); 106(23): 9362-7.

3. Eichler EE, *et al*. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450.

4. Makowsky R., Pajewski N. M., Klimentidis Y. C., Vazquez A. I., Duarte C. W., *et al*., 2011. Beyond missing heritability: Prediction of complex traits. *PLoS Genet*. **7**: e1002051.

5. Manolio, T.A. *et al*. 2009 Finding the missing heritability of complex diseases. *Nature* **461**, 747–753.

6. Haidich, A.B. (2010). Meta-analysis in medical research. *Hippokratia, 14*(Suppl 1), 29–37.

7. de Bakker, P. I., Ferreira, M. A., Jia, X., Neale, B. M., Raychaudhuri, S., & Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human molecular genetics, 17*(R2), R122–R128.

8. Lewinger, J.P., Conti, D. V., Baurley, J. W., Triche, T. J., & Thomas, D. C. (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genetic epidemiology, 31*(8), 871–882.

9. Marchini, J., Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499–511 (2010).

10. Wojcik, G.L., *et al*. (2018). Imputation-Aware Tag SNP Selection to Improve Power for Large-Scale, Multi-ethnic Association Studies. *G3: Genes, Genomes, Genetics, 8*(10), 3255-3267.

11. The 1000 Genomes Project Consortium., Corresponding Author., McVean, G. *et al*. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).

12. The Haplotype Reference Consortium., McCarthy, S., Das, S. *et al*. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283 (2016).

13. Hayes, B.J. & Daetwyler, H. D. 1000 Bull Genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annu. Rev. Anim. Biosci.* **7**, 1 (2018).

14. Frischknecht, M., Pausch, H., Bapst, B. *et al*. Highly accurate sequence imputation enables precise QTL mapping in Brown Swiss cattle. *BMC Genomics* **18**, 999 (2017).

15. VanRaden, P.M., Sun, C. & O'Connell, J.R. Fast imputation using medium or low-coverage sequence data. *BMC Genet* **16**, 82 (2015).

16. Sahana, G., *et al*. (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*(11), 7258–7275.

17. Faye, L. L., Machiela, M. J., Kraft, P., Bull, S. B., & Sun, L. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. *PLoS genetics* **9**, 8 (2013).

18. Stephens, M., Balding, D. Bayesian statistical methods for genetic association studies. *Nat Rev Genet* **10**, 681–690 (2009).

19. Chen, W., *et al*. (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics*, *200*(3), 719–736.

19. Kircher, M., Witten, D., Jain, P. *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).

20. Ioannidis, N. M., *et al*. (2017). FIRE: functional inference of genetic variants that regulate gene expression. *Bioinformatics*, *33*(24), 3895–3901.

21. Boyle, A. P., *et al*. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, *22*(9), 1790–1797.

22. Jiang, J., Cole, J.B., Freebern, E. *et al*. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Commun Biol* **2**, 212 (2019).

23. Price Alkes L., Spencer Chris C. A. and Donnelly Peter. 2015. Progress and promise in understanding the genetic basis of common diseases. *Proc. R. Soc. B.* **282**: 20151684

24. The ENCODE Project Consortium., Overall coordination (data analysis coordination)., Dunham, I. *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

25. Won H, de la Torre-Ubieta L, Stein JL, *et al*. Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*. 2016;538(7626):523-527.

26. Zhu, Z., Zhang, F., Hu, H. *et al*. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481–487 (2016).

27. Maurano MT, Humbert R, Rynes E, *et al*. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337(6099):1190-1195.

28. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, *et al*. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Res* **12**: 222–231.

29. Sonstegard, T. S. *et al*. Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. *PLoS ONE* **8**, e54872 (2013).

30. Sauna, Z., Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* **12**, 683–691 (2011).

31. Capon, Francesca *et al*. "A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups." *Human molecular genetics* vol. 13,20 (2004): 2361-8.

32. Jostins, L., Ripke, S., Weersma, R. *et al*. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).

33. Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., & Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, *45*(2), 124–130.

34. Ma, L., Cole, J. B., Da, Y., & VanRaden, P. M. (2019). Symposium review: Genetics, genome-wide association study, and genetic improvement of dairy fertility traits. *Journal of dairy science*, *102*(4), 3735–3743.

35. Raphaka, K., Matika, O., Sánchez-Molano, E. *et al*. Genomic regions underlying susceptibility to bovine tuberculosis in Holstein-Friesian cattle. *BMC Genet* **18**, 27 (2017).

36. Zare, Y., Shook, G. E., Collins, M. T., & Kirkpatrick, B. W. (2014). Genome-wide association analysis and genomic prediction of Mycobacterium avium subspecies paratuberculosis infection in US Jersey cattle. *PloS one*, *9*(2), e88380.

37. Lango Allen, H., Estrada, K., Lettre, G. *et al*. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).

38. Fang, L., Sahana, G., Ma, P. *et al*. Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet Sel Evol* **49**, 44 (2017).

39. Auer, P.L., Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* **7**, 16 (2015).

40. Gurdasani, D., Barroso, I., Zeggini, E. *et al*. Genomics of disease risk in globally diverse populations. *Nat Rev Genet* **20**, 520–535 (2019).

41. Höglund, J., Rafati, N., Rask-Andersen, M. *et al*. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Sci Rep* **9**, 16844 (2019).

42. Zhang, J., Kadri, N.K., Mullaart, E. *et al*. Genetic architecture of individual variation in recombination rate on the X chromosome in cattle. *Heredity* **125**, 304–316 (2020).

43. Wang, Y., Zhang, F., Mukiibi, R. *et al*. Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: II: carcass merit traits. *BMC Genomics* **21**, 38 (2020).

44. Witte J. S. (2010). Genome-wide association studies and beyond. *Annual review of public health*, *31*, 9–20.

45. Liu, S., Yu, Y., Zhang, S. *et al*. Epigenomics and genotype-phenotype association analyses reveal conserved genetic architecture of complex traits in cattle and human. *BMC Biol* **18**, 80 (2020).

46. Schrodi, S. J., Mukherjee, S., Shan, Y., *et al*. (2014). Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. *Frontiers in genetics*, *5*, 162.

47. Canela-Xandri, O., Rawlik, K., Woolliams, J. A., & Tenesa, A. (2016). Improved Genetic Profiling of Anthropometric Traits Using a Big Data Approach. *PloS one*, *11*(12), e0166755.

48. Kaiser, Jocelyn. "'Landmark' Study Resolves a Major Mystery of How Genes Govern Human Height." *Science*, 3 Nov. 2020.

49. Hayes, B. J., Pryce, J., Chamberlain, A. J., *et al*. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS genetics*, *6*(9), e1001139.

Chapter 2: Genomic Fine-Mapping of Livability and Six Health Traits in Holstein Cattle

Abstract

Background: Health traits are of significant economic importance to the dairy industry due to their effects on milk production and associated treatment costs. Genome-wide association studies (GWAS) provide a means to identify associated genomic variants and thus reveal insights into the genetic architecture of complex traits and diseases. The objective of this study is to investigate the genetic basis of seven health traits in dairy cattle and to identify potential candidate genes associated with cattle health using GWAS, fine mapping, and analyses of multi-tissue transcriptome data.

Results: We studied cow livability and six direct disease traits, mastitis, ketosis, hypocalcemia, displaced abomasum, metritis, and retained placenta, using de-regressed breeding values and more than three million imputed DNA sequence variants. After data edits and filtering on reliability, the number of bulls included in the analyses ranged from 11,880 (hypocalcemia) to 24,699 (livability). GWAS was performed using a mixed-model association test, and a Bayesian fine-mapping procedure was conducted to calculate a posterior probability of causality to each variant and gene in the candidate regions. The GWAS detected a total of eight genome-wide significant associations for three traits, cow livability, ketosis, and hypocalcemia, including the bovine Major Histocompatibility Complex (MHC) region associated with livability. Our fine-mapping of associated regions reported 20 candidate genes with the highest posterior probabilities of causality for cattle health. Combined with transcriptome data across multiple tissues in

cattle, we further exploited these candidate genes to identify specific expression patterns in disease-related tissues and relevant biological explanations such as the expression of Group-specific Component (GC) in the liver and association with mastitis as well as the Coiled-Coil Domain Containing 88C (CCDC88C) expression in CD8 cells and association with cow livability.

Conclusions: Collectively, our analyses report six significant associations and 20 candidate genes of cattle health. With the integration of multi-tissue transcriptome data, our results provide useful information for future functional studies and better understanding of the biological relationship between genetics and disease susceptibility in cattle.

Keywords: GWAS, Fine mapping, Health trait, Gene expression, Dairy cattle

Note: This chapter was previously published by BMC Genomics. Ellen Freebern made the main contribution and is the first author. The original citation is as follows:

GWAS and fine-mapping of livability and six disease traits in Holstein cattle. Freebern E, Santos DJA, Fang L, Jiang J, Parker Gaddis KL, Liu GE, VanRaden PM, Maltecca C, Cole JB, Ma L. *BMC Genomics*. 2020 Jan 13;**21**(1):41.

Introduction

One of the fundamental goals of animal production is to profitably produce nutritious food for humans from healthy animals. Profitability of the dairy industry is influenced by many factors, including production, reproduction, and animal health (Liang *et al*, 2017). Cattle diseases can cause substantial financial losses to producers as the result of decreased productivity, including milk that must be dumped, and increased costs for labor and veterinary care. Indirect costs associated with reduced fertility, reduced production after recovery, and increased risk of culling also can be substantial. For example, ketosis is a metabolic disease that occurs in cows during early lactation and hinders the cow's energy intake, thus subsequently reduces milk yield and increases the risk of displaced abomasum, which is very costly (Duffield *et al*, 2000). Mastitis is a major endemic disease of dairy cattle that can lead to losses to dairy farmers due to contamination, veterinary care, and decreased milk production (Seegers *et al*, 2003). In addition, cows may develop milk fever, a metabolic disease that is related to a low blood calcium level known as hypocalcemia (Reinhardt *et al*, 2011). Another common disease in cattle is metritis, which is inflammation of the uterus and commonly seen following calving when cows have a suppressed immune system and are vulnerable to bacterial infection (Bartlett *et al*, 1986). Complications during delivery can also result in a retained placenta (Laven *et al*, 1996). Many of the postpartum diseases are caused by the energy imbalance due to onset of lactation, especially in high producing cows. These complex diseases are jointly affected by management, nutrition, and genetics. A better understanding of the underlying genetic components can help the management and genetic improvements of cattle health.

Genome-wide association studies (GWAS) have been successful at interrogating the genetic basis of complex traits and diseases in cattle (Cole *et al*, 2011; Gaddis *et al*, 2018; Jiang *et al*, 2019; Ma *et al*, 2019). Pinpointing the causal variants of complex traits has been challenging because complex traits are influenced by many genes, their interactions, and the environment, and also because of the high level of linkage disequilibrium (LD) between genomic variants (Schaid *et al*, 2018). Fine-mapping is a common post-GWAS analysis, where posterior probabilities of causality are assigned to candidate variants and genes. In humans, fine-mapping of complex traits is currently performed along with or following GWAS studies. The utility of fine-mapping in cattle studies, however, has been limited by data availability and the high levels of LD present in cattle populations (Khatkar *et al*, 2008; McKay *et al*, 2007; Sargolzaei *et al*, 2008). To circumvent this challenge, a recent study developed a fast Bayesian Fine-MAPping method (BFMAP), which performs fine-mapping by integrating various functional annotation data (Jiang *et al*, 2019). Additionally, this method can be exploited to identify biologically meaningful information from candidate genes to enhance the understanding of complex traits (Fang *et al*, 2019).

The U.S. dairy industry has been collecting and evaluating economically important traits in dairy cattle since the late 1800s, when the first dairy improvement programs were formed. Since then, a series of dairy traits have been evaluated, including production, body conformation, reproduction, and health traits. Cow livability was included in the national genomic evaluation system by the Council on Dairy Cattle Breeding (CDCB) in 2016 (Wright *et al*, 2016). This trait reflects a cow's overall ability to stay alive in a milking herd by measuring the percentage of on-farm deaths per

lactation. Cow livability is partially attributable to health and can be selected to provide more milk revenue and less replacement of cows. In 2018, six direct health traits were introduced into the U.S. genomic evaluation, including ketosis, mastitis, hypocalcemia or milk fever, metritis, retained placenta, and displaced abomasum (Garrick *et al*, 2009). These phenotypic records along with genotype data collected from the U.S. dairy industry provide a unique opportunity to investigate the genetic basis of cattle health. The aim of our study is, therefore, to provide a powerful genetic investigation of seven health traits in cattle, to identify candidate disease genes and variants with relevant tissue-specific expression, and to provide insights into the biological relationship between candidate genes and the disease risk they may present on a broad scale.

Results

Genome-wide association study of livability and six direct health traits

We conducted genome-wide association analyses of seven health related traits using imputed sequence data and de-regressed breeding values for 27,214 Holstein bulls that have many daughter records and thus accurate phenotypes. After editing and filtering on reliability, we included 11,880 to 24,699 Holstein bulls across the seven traits (Table 1). Compared to the analysis using predicted transmitting ability (PTA) as phenotype (Additional file 1), GWAS on de-regressed PTA values produced more consistent and reliable results (Garrick *et al*, 2009). While different results between analyses of raw and de-regressed PTAs were obtained for the six health traits, little difference was observed for cow livability, which had more records and higher reliability (Table 1 and Additional

file 2). Therefore, we only considered association results obtained with de-regressed PTAs in all subsequent analyses.

Out of the seven health traits, we detected significantly associated genomic regions for only three traits after Bonferroni correction: hypocalcemia, ketosis, and livability (Fig. 1). In total, we had one associated region on BTA 6 for hypocalcemia, one region on BTA 14 for ketosis, and six regions for cow livability on BTA 5, 6, 14, 18, 21, and 23, respectively (Table 2). Notably, the bovine Major Histocompatibility Complex (MHC) region on BTA 23 (Takeshima *et al*, 2006) is associated with cow livability. Additionally, association signals on BTA 16 for ketosis (*P*-value = 1.9×10^{-8}) and BTA 6 for mastitis (*P*-value = 4.2×10^{-8}) almost reached the Bonferroni significance level. Other traits had prominent signals, but their top associations were below the Bonferroni threshold. Since sequence data have the highest coverage of functional variants in our study, we included all these regions to query the Cattle QTLdb for a comparative analysis.

When compared to existing studies, many of these health-related regions have been previously associated with milk production or disease related traits in cattle (Table 2) (Hu *et al*, 2012). The top associated region for hypocalcemia is around 10,521,824 bp on BTA 6, where QTLs were reported for body/carcass weight and reproduction traits with nearby genes being Translocation Associated Membrane Protein 1 Like (*TRAM1L1*) and N-Deacetylase And N-Sulfotransferase (*NDST4*). The region around 2,762,595 bp on BTA 14 for ketosis is involved with milk and fat metabolism and contains the well-known Diacylglycerol O-Acyltransferase 1 (*DGAT1*) gene. The region around 7,048,452 bp on BTA 16 for ketosis was also previously associated with fat metabolism. The region

around 88,868,886 bp on BTA 6 associated with mastitis is close to the GC gene with many reported QTLs associated with mastitis (Jiang *et al*, 2019; Olsen *et al*, 2016; Sahana *et al*, 2013; Wu *et al*, 2015). This region was also associated with cow livability in this study, with QTLs involved with the length of productive life (Nayeri *et al*, 2017). For the six regions associated with cow livability (Table 2), we found reported QTLs related to productive life, somatic cell count, immune response, reproduction, and body conformation traits (Nayeri *et al*, 2017). The top associated regions for displaced abomasum on BTA 4 and BTA 8 have been previously associated with cattle reproduction and body conformation traits (Nalaila *et al*, 2012; Pryce *et al*, 2011; Snelling *et al*, 2010). For metritis, the top associated variant, 3,662,486 bp on BTA4, is close to small nucleolar RNA MBI161 (*SNORA31*), and around ±1 Mb upstream and downstream were QTLs associated with production, reproduction, and dystocia (Olsen *et al*, 2010). Genes RUN Domain Containing 3B (*RUNDC3B*; BTA 4), Quinoid Dihydropteridine Reductase (*QDPR*; BTA 6), Transmembrane Protein 182 (*TMEM182*; BTA 11), and Zinc Finger Protein (*ZFP28*; BTA 18) are the closest genes to the retained placenta signals with previous associations related to milk production, productive life, health and reproduction traits, including calving ease and stillbirth (Cole *et al*, 2011).

Association of livability QTL with other disease traits

Cow livability is a health-related trait that measures the overall robustness of a cow. As the GWAS of cow livability was the most powerful among the seven traits and detected six QTL regions, we evaluated whether these livability QTLs were also associated with other disease traits. Out of the six livability QTLs, four of them were related to at least one disease trait at the nominal significance level (Table 3). All these overlapping

associations exhibited consistent directions of effect: alleles related to longer productive life were more resistant to diseases. The most significant QTL of livability on BTA 18 is associated with displaced abomasum and metritis, both of which can occur after abnormal birth. This QTL has been associated with gestation length, calving traits, and other gestation and birth related traits (Fang *et al,* 2019). The QTL on BTA 6 is associated with hypocalcemia, ketosis, and mastitis. The BTA 21 QTL is associated with hypocalcemia and mastitis. The BTA 5 QTL is related to displaced abomasum and ketosis. Interestingly, the bovine MHC region on BTA 23 is not associated with the immune-related disease traits, which suggests that those genes do not explain substantial variation for the presence or absence of a disease during a lactation or that we do not have enough power to detect the association.

Fine-mapping analyses and validation from tissue-specific expression

Focusing on the candidate QTL regions in Table 2, the fine-mapping analysis calculated posterior probabilities of causalities (PPC) for individual variants and genes to identify candidates (Table 4), which were largely consistent with the GWAS results. A total of eight genes detected in GWAS signals were also successfully fine-mapped, including Plexin A4 (*PLXNA4*), FA Complementation Group C (*FANCC*), Neurotrimin (*NTM*) for displaced abomasum, GC for mastitis and livability, ATP Binding Cassette Subfamily C Member 9 (*ABCC9*) for livability, QDPR for retained placenta, Zinc Finger And AT-Hook Domain Containing (*ZFAT*) and *CCDC88C* for livability. In addition, fine-mapping identified new candidate genes, including Cordon-Bleu WH2 Repeat Protein (*COBL*) on BTA 4 for metritis, *LOC783947* on BTA 16 for ketosis, *LOC783493* on BTA 18 for retained placenta, and *LOC618463* on BTA 18 and *LOC101908667* on BTA 23 for

livability. The genes LOC107133096 on BTA 14 and *LOC100296627* on BTA 4 detected respectively for ketosis and retained placenta by fine mapping were close to two genes (*DGAT1* and *ABCB1* or ATP Binding Cassette Subfamily B Member 1) that have known biological association with milk production and other traits. In addition to the detected genes in these two cases, we further investigated genes with a potential biological link with disease, and genes with the highest PPC (*PARP10* or PolyADP-ribose polymerase 10 and *MALSU1* or Mitochondrial Assembly Of Ribosomal Large Subunit 1) that were located between these two references (Table 4). No genes were detected by fine-mapping in the signal on BTA 6 for hypocalcemia (Fig. 1), given that the nearest genes were beyond a 1 Mb window boundary.

In addition, we investigated the expression levels of fine-mapped candidate genes across cattle tissues using existing RNA-Seq data from public databases. While many genes are ubiquitously expressed in multiple tissues, several fine-mapped genes were specifically expressed in a few tissues relevant to cattle health (Table 4). Interesting examples of tissue-specific expression and candidate genes included liver with mastitis and livability (*GC*), and CD8 cells with livability (*CCDC88C*). Although this analysis is preliminary, these results provide additional support for these candidate genes of cattle health and help the understanding of how and where their expression is related with dairy disease resistance.

Discussion

In this study, we performed powerful GWAS analyses of seven health and related traits in Holstein bulls. The resulting GWAS signals were further investigated by a Bayesian fine-mapping approach to identify candidate genes and variants. Additionally, we included tissue-specific expression data for candidate genes to reveal a potential biological relationship between genes, tissues and cattle diseases. Finally, we provide a list of candidate genes of cattle health with associated tissue-specific expression that can be readily tested in future functional validation studies.

In our GWAS analysis, we used de-regressed PTA as phenotype and incorporated the reliabilities of the deregressed PTAs of livability and six disease traits. Three traits were found to have significant association signals, hypocalcemia, ketosis, and livability, which demonstrated the power of our GWAS study. For example, we also observed regions associated with livability, in particular, with the region around 58,194,319 on BTA 18 to possess a large effect on dairy and body traits. Our finding was corroborated by a BLAST analysis that identified a related molecule, Siglec-6, which is expressed in tissues such as the human placenta (Cole *et al*, 2009). Further analyses can be performed to characterize the functional implications of these association regions for the seven health and related traits in cattle.

When using PTA values as phenotype in GWAS, we observed different regions to be associated, compared to the GWAS with de-regressed PTA (Fig. 1 and Additional file 2). For example, a genomic region larger than 4 Mb on BTA 12 was associated with most of the health traits (Additional file 2). Although these generally appeared as clear

association signals, we observed only a few HD SNP markers to be associated, which may be due to poor imputation. Additionally, this region was reported by VanRaden *et al*. (2017) as having low imputation accuracy. The lower imputation accuracy on BTA 12 was determined to be caused by a gap between the 72.4 and 75.2 Mb region where no SNPs were present on the HD SNP array (VanRaden *et al*, 2017). Additional studies are needed to address this imputation issue in order to improve the accuracy and power of future analysis on this region. Since different family relationship will affect the GWAS results when using direct versus deregressed PTAs, these differences in relatedness can lead to false positive GWAS results, especially for low-quality imputed data. In sum, this comparison of GWAS using PTA and de-regressed PTA supports the use of de-regressed PTA values with reliabilities accounted for in future GWAS studies in cattle.

Application of BFMAP for fine-mapping allowed us to identify 20 promising candidate genes (Table 4) and a list of candidate variants (Additional file 3) for health traits in dairy cattle. We found that most of the genes possess tissue-specific expression, notably the detected gene *LOC107133096* on BTA 14 for ketosis. This gene is located close to the *DGAT1* gene that affects milk fat composition. A previous candidate gene association study by Tetens *et al.* (2013) proposed *DGAT1* to be an indicator of ketosis. In that study, the *DGAT1* gene was determined to be involved in cholesterol metabolism, which is known to be an indicator of a ketogenic diet in humans (Tetens *et al*, 2013). This result highlights a potential pathway in the pathogenesis of ketosis that may be an area for future research. Additionally, ketosis is a multifactorial disease that is likely influenced by multiple loci. Therefore, implementation of a functional genomics approach would allow identification of more genetic markers, and in doing so, improve

resistance to this disease. For displaced abomasum, the gene *PLXNA4* was observed to have an association with the variant 97,101,981 bp on BTA 4 (Table 4 and Additional file 3). Our analysis also detected tissue-specific expression for *PLXNA4* in the aorta. A previous study on atherosclerosis found that Plexin-A4 knockout mice exhibited incomplete aortic septation (Toyofuku *et al*, 2008). These findings provide some support for the potential association of *PLXNA4* with cattle health.

Six signals were observed as clear association peaks for livability (Fig. 1). The associated variant at 8,144,774 – 8,305,775 bp on BTA 14 was close to the gene *ZFAT*, which is known to be expressed in the human placenta (Barbaux *et al*, 2012). In particular, the expression of this gene is downregulated in placentas from complicated pregnancies. Additionally, a GWAS study performed in three French dairy cattle populations found the *ZFAT* gene to be the top variant associated with fertility (Marete *et al*, 2018). Since calving and other fertility issues could be risk factors to cause animal death, these results lend support of this candidate gene with the livability. On BTA18, the associated variant at 57,587,990 – 57,594,549 bp was near the gene *LOC618463*, which has been previously identified as a candidate gene associated with calving difficulty in three different dairy populations (Gowane *et al*, 2015). The associated variant at 56,645,629 – 56,773,438 bp on BTA21 is located close to the *CCDC88C* gene (Table 4). In addition to our detection of tissue-specific expression with the CD8 cell, this gene has been associated with traits such as dairy form and days to first breeding in cattle (Jiang *et al*, 2019).

It is notable that our GWAS signal for livability at 25,904,084 – 25,909,461 bp on BTA 23 is located in the bovine MHC region (Table 4). The gene we detected was

LOC101908667, which is one of the immune genes of the MHC. This is of considerable interest because MHC genes have a role in immune regulation. The MHC complex of cattle located on BTA 23 is called the bovine leukocyte antigen (*BoLA*) region. This complex of genes has been extensively studied in research investigating the polymorphism of genes in *BoLA* and their association with disease resistance (Gowane *et al*, 2013). Therefore, our research highlights a gene of considerable interest that should be further explored to understand its importance in breeding programs and its potential role in resistance to infectious diseases.

Additionally, we identified an associated variant for livability at 88,687,845 - 88,739,292 bp on BTA6 close to the gene *GC*, which was specifically expressed in tissues such as the liver (Table 4). This gene has been previously studied in an association analysis that investigated the role of *GC* on milk production (Olsen *et al*, 2016). It found that the gene expression of *GC* in cattle is predominantly expressed in the liver. Moreover, affected animals displayed decreased levels of the vitamin D binding protein (DBP) encoded by *GC*, highlighting the importance of *GC* for a cow's production. Additionally, liver-specific *GC* expression has been identified in humans, specifically regulated through binding sites for the liver-specific factor HNF1 (Huroki *et al*, 2007). Collectively, these results offer evidence for *GC* expression in the liver, which may be an important factor for determining cow livability.

Interestingly, the *GC* gene was also detected to have tissue-specific expression in the liver for mastitis (Table 4). This is corroborated by a study on cattle infected with mastitis to possess limited DBP concentration (Olsen *et al*, 2016). Vitamin D plays a key part in maintaining serum levels of calcium when it is secreted into the milk (Horst *et al*,

2003). Since *GC* encodes DBP, it was suggested that the GC gene has a role in regulating milk production and the incidence of mastitis infection in dairy cattle. It is important to note that bovine mastitis pathogens, such as *Staphylococcus aureus* and *Escherichia coli*, also commonly occur as pathogens of humans. Therefore, development of molecular methods to contain these pathogens is of considerable interest for use in human medicine to prevent the spread of illness and disease. For instance, the use of enterobacterial repetitive intergenic consensus typing enables trace back of clinical episodes of *E. coli* mastitis, thus allowing for an evaluation of antimicrobial products for the prevention of mastitis (Zadoks *et al,* 2011). Continued investigation using molecular methods are needed to understand the pathogenesis of mastitis and its comparative relevance to human medicine. Based on the fine mapping for metritis, the new gene assigned was *COBL* on BTA 6 (Table 4). However, this candidate gene was found to have variants only passing the nominal significance level for causality and for GWAS. Further exploration of this candidate gene is needed to contribute to our understanding of its function and potential tissue-specific expression.

For retained placenta, the gene *TMEM182* was observed to have an association with a variant between 7,449,519 – 7,492,871 bp on BTA11 (Table 4). Our tissue-specific analysis identified *TMEM182* to have an association in muscle tissues. A study performed in Canchim beef cattle investigated genes for male and female reproductive traits and identified *TMEM182* on BTA 11 as a candidate gene that could act on fertility (Buzanskas *et al,* 2017). Additionally, the gene *TMEM182* has been found to be upregulated in brown adipose tissue in mice during adipogenesis, which suggests a role in the development of muscle tissue (Wu *et al,* 2008). One important factor that causes

retention of fetal membranes in cattle is the impaired muscular tone of organs such as the uterus and abdomen (Schlafer *et al*, 2000). This suggests the importance of the *TMEM182* gene and the need for future studies to better understand its role in the cattle breeding program.

Conclusions

In this study, we reported eight significant associations for seven health and related traits in dairy cattle. In total, we identified 20 candidate genes of cattle health with the highest posterior probability, which are readily testable in future functional studies. Several candidate genes exhibited tissue-specific expression related to immune function, muscle growth and development, and neurological pathways. The identification of a novel association for cow livability in the bovine MHC region also represents an insight into the biology of disease resistance. Overall, our study offers a promising resource of candidate genes associated with complex diseases in cattle that can be applied to breeding programs and future studies of disease genes for clinical utility.

Methods

Ethics statement

This study did not require the approval of the ethics committee, as no biological materials were collected.

Genotype data

Using 444 ancestor Holstein bulls from the 1000 Bull Genomes Project as reference, we previously imputed sequence variants for 27,214 progeny-tested Holstein bulls that have highly reliable phenotypes via FindHap version 3 (VanRaden *et al*, 2014). We applied stringent quality-control procedures before and after imputation to ensure the data quality. The original 777,962 HD SNPs were reduced to 312,614 by removing highly correlated SNP markers with a |r| value higher than 0.95 and by prior editing. Variants with a minor allelic frequency (MAF) lower than 0.01, incorrect map locations (UMD3.1 bovine reference assembly), an excess of heterozygotes, or low correlations ($|r| < 0.95$) between sequence and HD genotypes for the same variant were removed. The final imputed data was composed of 3,148,506 sequence variants for 27,214 Holstein bulls. Details about the genomic data and imputation procedure are described by VanRaden et al. (VanRaden *et al*, 2017). After imputation, we only retained autosomal variants with MAF ≥0.01 and *P*-value of Hardy-Weinberg equilibrium test > 10^{-6}.

Phenotype data

The data used were part of the 2018 U.S. genomic evaluations from the Council on Dairy Cattle Breeding (CDCB), consisting of 1,922,996 Holstein cattle from the national dairy

cattle database. Genomic predicted transmitting ability (PTA) values were routinely calculated for these animals and were included in this study. Deregressed PTA values according to Garrick *et al.* (2009) were analyzed in GWAS for livability, hypocalcemia, displaced abomasum, ketosis, mastitis, metritis, and retained placenta. We restricted the de-regression procedure to those bulls with PTA reliability greater than parent average reliability, thus reducing the total number of animals from 27,214 to 11,880, 13,229, 12,468, 14,382, 13,653, 13,541, and 24,699 for the seven traits, respectively (Table 1).

Genome-wide association study (GWAS)

A mixed-model GWAS was performed using MMAP, a comprehensive mixed model program for analysis of pedigree and population data (O'Connell *et al*, 2015). The additive effect was divided into a random polygenic effect and a fixed effect of the candidate SNP. The variance components for the polygenic effect and random residuals were estimated using the restricted maximum likelihood (REML) approach. MMAP has been widely used in human and cattle GWAS studies (Backman *et al*, 2017; Ma *et al*, 2015; Santos *et al*, 2018). The model can be generally presented as:

$$\boldsymbol{y} = \mu + \boldsymbol{m}b + \boldsymbol{a} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector with de-regressed PTAs; $\boldsymbol{\mu}$ is the global mean; $\boldsymbol{m}$ is the candidate SNP genotype (allelic dosage coded as 0, 1 or 2) for each animal; $\boldsymbol{b}$ is the solution effect of the candidate SNP; $\boldsymbol{a}$ is a solution vector of polygenic effect accounting for the population structure assuming $\boldsymbol{a} \sim N(0, \boldsymbol{G}\sigma_a^2)$, where **G** is a relationship matrix; and $\boldsymbol{e}$ is a vector of residuals assuming $\boldsymbol{e} \sim N(0, \boldsymbol{R}\sigma_e^2)$, where **R** is a diagonal matrix with diagonal elements weighted by the individual de-regressed reliability ($R_{ii=}1/(r_i^2 - 1)$). For each candidate

variant, a Wald test was applied to evaluate the alternative hypothesis, H_1: $b \neq 0$, against the null hypothesis H_0: $b = 0$. Bonferroni correction for multiple comparisons was applied to control the type-I error rate. Gene coordinates in the UMD v3.1 assembly were obtained from the Ensembl Genes 90 database using the BioMart tool (Zimin *et al*, 2009). The cattle QTLdb database was examined to check if any associated genomic region was previously reported as a cattle quantitative trait locus (QTL) (Hu *et al*, 2012).

Fine-mapping association study

In order to identify potential candidate genes and their causal variants, GWAS signals were investigated through a fine-mapping procedure using a Bayesian approach with the software BFMAP v.1 (https://github.com/jiang18/bfmap) (Jiang *et al*, 2019). BFMAP is a software tool for genomic analysis of quantitative traits, with a focus on fine-mapping, SNP-set association, and functional enrichment. It can handle samples with population structure and relatedness and calculate posterior probability of causality (PPC) to each variant and its causality p-value for independent association signals within candidate QTL regions. The minimal region covered by each lead variant was determined as ±1 Mb upstream and downstream (candidate region ≥2 Mb). This extension allowed the region to cover most variants that have an LD r^2 of > 0.3 with the lead variants. The employed fine-mapping approach included three steps: forward selection to add independent signals in the additive Bayesian model, repositioning signals, and generating credible variant sets for each signal. Details about the BFMAP algorithm and its procedure are described by Jiang *et al*. (2019).

Tissue-specific expression of candidate genes

From publicly available resources including the NCBI GEO database, we have assembled RNA-seq data of 723 samples that involves 91 tissues and cell types in Holstein cattle. We processed all the 732 RNA-seq data uniformly using a rigorous bioinformatics pipeline with stringent quality control procedures. After data cleaning and processing, we fit all data into one model to estimate the tissue specificity of gene expression. We then calculated the t-statistics for differential expression for each gene in a tissue using a previous method (Finucane *et al*, 2018). Specifically, the log2-transformed expression (i.e., log2FPKM) of genes was standardized with mean of 0 and variance of 1 within each tissue or cell type,

$$y_i = \mu_i + x_{is} + x_{iage} + x_{isex} + x_{istudy} + e_i$$

where y_i is the standardized log2-transformed expression level (i.e., log2FPKM) of *i*th gene; μ_i is the overall mean of the *i*th gene; x_{is} is the tissue effect, where samples of the tested tissue were denoted as '1', while other samples as '-1'; x_{iage}, x_{isex}, x_{istudy} were age, sex, and study effects for the *i*th gene, respectively; e_i is residual effect. We fit this model for each gene in each tissue using the ordinary least-square approach and then obtained the t-statistics for the tissue effect to measure the expression specificity of this gene in the corresponding tissue. Using this approach, we evaluated the expression levels for each of the candidate genes that were fine-mapped in this study across the 91 tissues and cell types and identified the most relevant tissue or cell type for a disease trait of interest.

References

1. Liang D, Arnold L, Stowe C, Harmon R, Bewley J: Estimating US dairy clinical disease costs with a stochastic simulation model. *Journal of dairy science* 2017, **100**(2):1472-1486.

2. Duffield T: Subclinical ketosis in lactating dairy cattle. *Veterinary clinics of north america: Food animal practice* 2000, **16**(2):231-253.

3. Seegers H, Fourichon C, Beaudeau F: Production effects related to mastitis and mastitis economics in dairy cattle herds. *Veterinary research* 2003, **34**(5):475-491.

4. Reinhardt TA, Lippolis JD, McCluskey BJ, Goff JP, Horst RL: Prevalence of subclinical hypocalcemia in dairy herds. *The Veterinary Journal* 2011, **188**(1):122-124.

5. Bartlett PC, Kirk JH, Wilke MA, Kaneene JB, Mather EC: Metritis complex in Michigan Holstein-Friesian cattle: incidence, descriptive epidemiology and estimated economic impact. *Preventive veterinary medicine* 1986, **4**(3):235-248.

6. Laven R, Peters A: Bovine retained placenta: aetiology, pathogenesis and economic loss. *Veterinary Record* 1996, **139**(19):465-471.

7. Ma L, Cole J, Da Y, VanRaden P: Symposium review: Genetics, genome-wide association study, and genetic improvement of dairy fertility traits. *J Dairy Sci* 2019; **102**(4):3735-43.

8. Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Crooker BA, Van Tassell CP, Yang J, Wang S, Matukumalli LK: Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. *BMC genomics* 2011, **12**(1):408.

9. Gaddis KP, Megonigal J Jr, Clay J, Wolfe C. Genome-wide association study for ketosis in US jerseys using producer-recorded data. J Dairy Sci. 2018; **101**(1):413–24.

10. Jiang J, Cole JB, Freebern E, Da Y, VanRaden PM, Ma L. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. Commun Biol. 2019; **2**(1):212.

11. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018; **19**(8):491–504.

12. Sargolzaei M, Schenkel F, Jansen G, Schaeffer L. Extent of linkage disequilibrium in Holstein cattle in North America. J Dairy Sci. 2008; **91**(5):2106–17.

13. Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Cavanagh JA, Barris W, Schnabel RD, Taylor JF, Raadsma HW. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics. 2008; **9**(1):187.

14. McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, Crews D, Neto ED, Gill CA, Gao C. Whole genome linkage disequilibrium maps in cattle. BMC Genet. 2007; **8**(1):74.

15. Fang L, Jiang J, Li B, Zhou Y, Freebern E, VanRaden PM, Cole JB, Liu GE, Ma L. Genetic and epigenetic architecture of paternal origin contribute to gestation length in cattle. Commun Biol. 2019; **2**(1):100.

16. Wright J, VanRaden P. Genetic evaluation of dairy cow livability. J Anim Sci. 2016; 94:178.

17. Parker Gaddis K, Tooker M, Wright J, Megonigal J, Clay J, Cole J, VanRaden P: Development of national genomic evaluations for health traits in U.S. Holsteins. Proc 11th World Congr Genet Appl Livest Prod, Auckland, New Zealand, Feb 11–16 2018, Vol. Biol. & Species–Bovine (dairy) 1, p. 594.

18. Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet Sel Evol. 2009; **41**(1):55.

19. Hu Z-L, Park CA, Wu X-L, Reecy JM. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the postgenome era. Nucleic Acids Res. 2012; **41**(D1):D871–9.

20. Takeshima SN, Aida Y. Structure, function and disease susceptibility of the bovine major histocompatibility complex. Anim Sci J. 2006; **77**(2):138–50.

21. Olsen HG, Knutsen TM, Lewandowska-Sabat AM, Grove H, Nome T, Svendsen M, Arnyasi M, Sodeland M, Sundsaasen KK, Dahl SR. Fine mapping of a QTL on bovine chromosome 6 using imputed full sequence data suggests a key role for the group-specific component (GC) gene in clinical mastitis and milk production. Genet Sel Evol. 2016; **48**(1):79.

22. Sahana G, Guldbrandtsen B, Thomsen B, Lund MS. Confirmation and finemapping of clinical mastitis and somatic cell score QTL in Nordic Holstein cattle. Anim Genet. 2013; **44**(6):620–6.

23. Wu X, Lund MS, Sahana G, Guldbrandtsen B, Sun D, Zhang Q, Su G. Association analysis for udder health based on SNP-panel and sequence data in Danish Holsteins. Genet Sel Evol. 2015; **47**(1):50.

24. Nayeri S, Sargolzaei M, Abo-Ismail M, Miller S, Schenkel F, Moore S, Stothard P. Genome-wide association study for lactation persistency, female fertility, longevity, and lifetime profit index traits in Holstein dairy cattle. J Dairy Sci. 2017; **100**(2):1246–58.

25. Snelling W, Allan M, Keele J, Kuehn L, Mcdaneld T, Smith T, Sonstegard T, Thallman R, Bennett G. Genome-wide association study of growth in crossbred beef cattle. J Anim Sci. 2010; **88**(3):837–48.

26. Pryce JE, Hayes BJ, Bolormaa S, Goddard ME. Polymorphic regions affecting human height also control stature in cattle. Genetics. 2011; **187**(3):981–4.

27. Nalaila S, Stothard P, Moore S, Li C, Wang Z. Whole-genome QTL scan for ultrasound and carcass merit traits in beef cattle using Bayesian shrinkage method. J Anim Breed Genet. 2012; **129**(2):107–19.

28. Olsen H, Hayes B, Kent M, Nome T, Svendsen M, Lien S. A genome wide association study for QTL affecting direct and maternal effects of stillbirth and dystocia in cattle. Anim Genet. 2010; **41**(3):273–80.

29. Cole J, VanRaden P, O'Connell J, Van Tassell C, Sonstegard T, Schnabel R, Taylor J, Wiggans G. Distribution and location of genetic effects for dairy traits. J Dairy Sci. 2009; **92**(6):2931–46.

30. VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol. 2017; **49**(1):32.

31. Tetens J, Seidenspinner T, Buttchereit N, Thaller G. Whole-genome association study for energy balance and fat/protein ratio in German Holstein bull dams. Anim Genet. 2013; **44**(1):1–8.

32. Toyofuku T, Yoshida J, Sugimoto T, Yamamoto M, Makino N, Takamatsu H, Takegahara N, Suto F, Hori M, Fujisawa H. Repulsive and attractive semaphorins cooperate to direct the navigation of cardiac neural crest cells. Dev Biol. 2008; **321**(1):251–62.

33. Barbaux S, Gascoin-Lachambre G, Buffat C, Monnier P, Mondon F, Tonanny M-B, Pinard A, Auer J, Bessières B, Barlier A. A genome-wide approach reveals novel imprinted genes expressed in the human placenta. Epigenetics. 2012; **7**(9):1079–90.

34. Marete AG, Guldbrandtsen B, Lund MS, Fritz S, Sahana G, Boichard D. A metaanalysis including pre-selected sequence variants associated with seven traits in three French dairy cattle populations. Front Genet. 2018; 9:522.

35. Purfield DC, Bradley DG, Evans RD, Kearney FJ, Berry DP. Genome-wide association study for calving performance using high-density genotypes in dairy and beef cattle. Genet Sel Evol. 2015; **47**(1):47.

36. Gowane G, Vandre R, Nangre M, Sharma A. Major histocompatibility complex (MHC) of bovines: an insight into infectious disease resistance. Livestock Res Int. 2013; **1**(2):46–57.

37. Hiroki T, Liebhaber SA, Cooke NE. An intronic locus control region plays an essential role in the establishment of an autonomous hepatic chromatin domain for the human vitamin D-binding protein gene. Mol Cell Biol. 2007; **27**(21):7365–80.

38. Horst R, Goff J, Reinhardt T. Role of vitamin D in calcium homeostasis and its use in prevention of bovine periparturient paresis. Acta Vet Scand Suppl. 2003; **97**:35–50.

39. Zadoks RN, Middleton JR, McDougall S, Katholm J, Schukken YH. Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. J Mammary Gland Biol Neoplasia. 2011; **16**(4):357–72.

40. Buzanskas ME, do Amaral Grossi D, Ventura RV, Schenkel FS, TCS C, Stafuzza NB, Rola LD, SLC M, Mokry FB, de Alvarenga Mudadu M. Candidate genes for male and female reproductive traits in Canchim beef cattle. J Anim Sci Biotechnol. 2017; **8**(1):67.

41. Wu Y, Smas CM. Expression and regulation of transcript for the novel transmembrane protein Tmem182 in the adipocyte and muscle lineage. BMC Res Notes. 2008; **1**(1):85.

42. Schlafer D, Fisher P, Davies C. The bovine placenta before and after birth: placental development and function in health and disease. Anim Reprod Sci. 2000; 60:145–60.

43. VanRaden PM, Sun C: Fast Imputation Using Medium- or Low-Coverage Sequence Data. Proceedings, 10th World Congress of Genetics Applied to Livestock Production 2014.

44. O'Connell JR: MMAP User Guide. Available: http://edn.som.umaryland.edu/mmap/index.php. Accessed 8 Oct 2015. 2015.

45. Backman JD, O'Connell JR, Tanner K, Peer CJ, Figg WD, Spencer SD, Mitchell BD, Shuldiner AR, Yerges-Armstrong LM, Horenstein RB. Genome-wide analysis of clopidogrel active metabolite levels identifies novel variants that influence antiplatelet response. Pharmacogenet Genomics. 2017; **27**(4):159.

46. Santos D, Cole J, Null D, Byrem T, Ma L. Genetic and nongenetic profiling of milk pregnancy-associated glycoproteins in Holstein cattle. J Dairy Sci. 2018; **101**(11):9987–10000.

47. Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, Bickhart DM, Cole JB, Null DJ, Liu GE. Cattle sex-specific recombination and genetic control from a large pedigree analysis. PLoS Genet. 2015; **11**(11):e1005387.

48. Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol. 2009; **10**(4):R42.

49. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, Gazal S, Loh P-R, Lareau C, Shoresh N. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet. 2018; **50**(4):621.

Tables

Table 1. Number of Holstein bulls, reliability of PTA, and heritability (h^2) for six health traits and cow livability.

Trait	N	h^2	Average Reliability
Hypocalcemia	11,880	0.006	0.228
Displaced Abomasum	13,229	0.011	0.269
Ketosis	12,468	0.012	0.260
Mastitis	14,382	0.031	0.338
Metritis	13,653	0.014	0.281
Retained Placenta	13,541	0.001	0.266
Livability	14,699	0.040	0.397

Table 2. Top SNPs and candidate genes associated with hypocalcemia (CALC), displaced abomasum (DSAB), ketosis (KETO), mastitis (MAST), metritis (METR), retained placenta (RETP) and cow livability.

Trait	Chr	Position	MAF	*P*-value	Genes Nearby	Traits Previously Associated[b]
CALC	6	10,521,824	0.014	8.3×10^{-10}[a]	*TRAM1L1, NDST4*	Subcutaneous fat
DSAB	4	97,101,981	0.021	4.4×10^{-7}	*PLXNA4, CHCHD3*	Milk protein yield
DSAB	8	83,052,202	0.109	1.3×10^{-7}	*FANCC*	Stature
DSAB	29	35,977,236	0.073	1.3×10^{-7}	*NTM*	Milk kappa-casein percentage
KETO	14	2,762,595	0.033	1.8×10^{-9}[a]	*LY6K*	Milk protein percentage
KETO	16	7,048,452	0.019	1.9×10^{-8}	*KCNT2*	Milk fat percentage
MAST	6	88,868,886	0.460	4.2×10^{-8}	*GC*	Clinical mastitis
METR	4	3,662,486	0.011	2.7×10^{-7}	*RF00322*	Milk protein yield
RETP	4	32,578,298	0.218	7.4×10^{-7}	*RUNDC3B*	Calving ease
RETP	6	117,620,548	0.026	7.2×10^{-7}	*QDPR*	Milk kappa-casein percentage
RETP	11	7,465,110	0.060	9.1×10^{-8}	*TMEM182*	Abomasum displacement
RETP	18	64,492,219	0.012	1.6×10^{-7}	*ZFP28*	Still birth
Livability	5	88,823,164	0.472	1.5×10^{-10}[a]	*ABCC9*	Productive life
Livability	6	88,801,999	0.454	1.7×10^{-18}[a]	*GC*	Clinical mastitis
Livability	14	8,536,538	0.020	5.3×10^{-10}[a]	*ZFAT*	Productive life
Livability	18	58,194,319	0.075	1.1×10^{-20}[a]	*ZNF614*	Bovine respiratory disease
Livability	21	56,700,449	0.013	8.6×10^{-11}[a]	*CCDC88C*	Type
Livability	23	26,131,593	0.017	3.8×10^{-9}[a]	*BLA-DQB*	Antibody-mediated immune response

[a]Genome-wide significance after Bonferroni correction
[b]Information obtained from the Animal QTLdb for cattle [19]

Table 3. Association results of the top SNPs associated with cow livability for hypocalcemia, displaced abomasum, ketosis, mastitis, and metritis. *P*-values larger than 0.05 and their Beta coefficients were excluded.

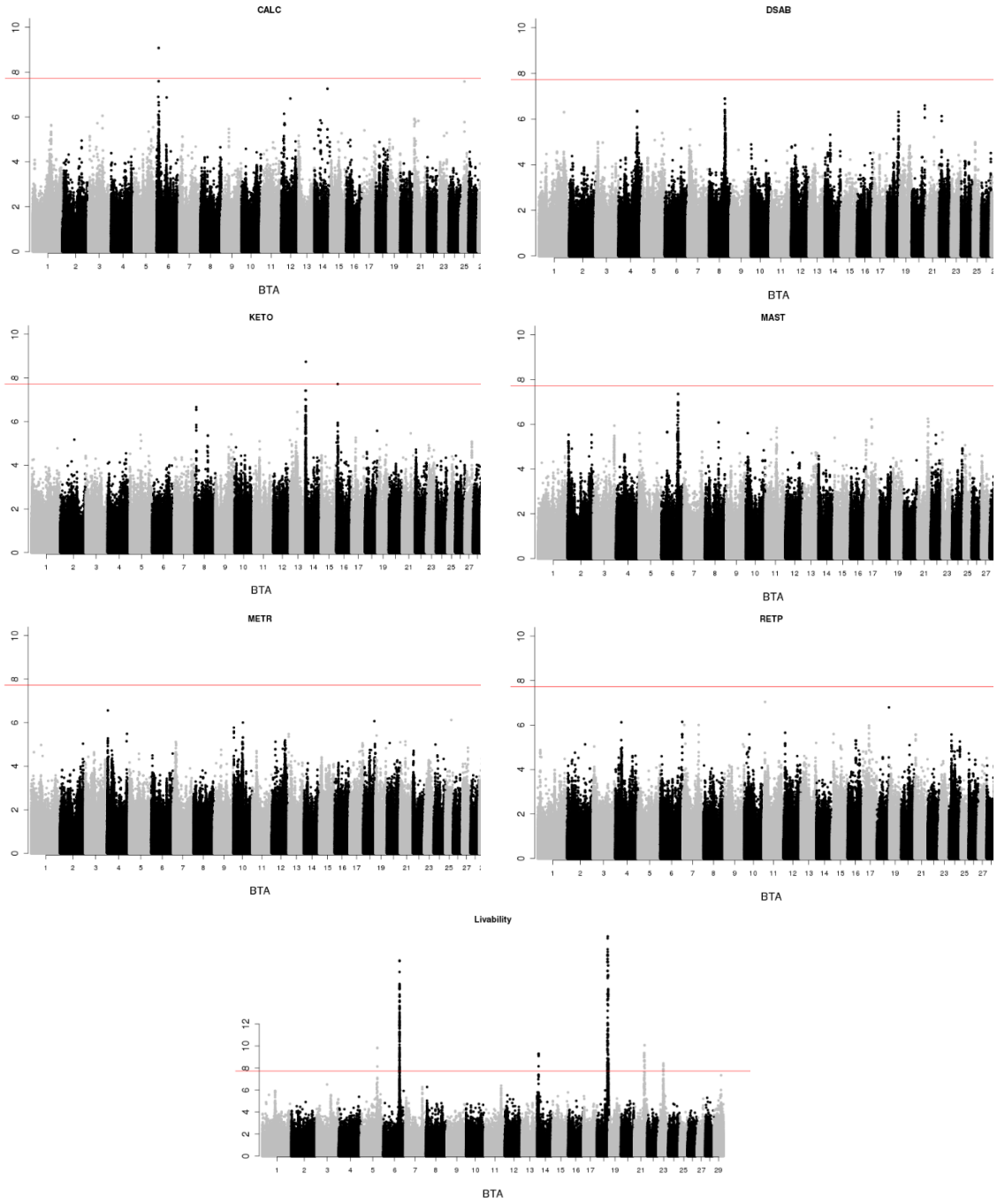
Chr	Position	Livability		Hypocalcemia		Displaced Abomasum		Ketosis		Mastitis		Metritis	
		***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**	***P*-value**	**Beta**
5	88,823,164	1.5×10^{-10}	-0.43	-	-	0.04	-0.14	0.04	-0.21	-	-	-	-
6	88,801,999	1.7×10^{-18}	-0.66	5.0×10^{-3}	-0.2	-	-	2.1×10^{-3}	-0.35	4.2×10^{-7}	-0.75	-	-
14	8,536,538	5.3×10^{-10}	-1.1	-	-	-	-	-	-	-	-	-	-
18	58,194,319	1.1×10^{-20}	-1.0	-	-	1.1×10^{-4}	-0.47	-	-	-	-	0.01	-0.51
21	56,700,449	8.6×10^{-11}	-1.5	0.03	-0.58	-	-	-	-	9.1×10^{-3}	-1.43	-	-
23	26,131,593	3.8×10^{-9}	0.71	-	-	-	-	-	-	-	-	-	-

Table 4. List of candidate genes with highest posterior probability of causality (PPC) and their minimum *P*-values for causality (M_causality) and GWAS (M_GWAS) associated with hypocalcemia (CALC), displaced abomasum (DSAB), ketosis (KETO), mastitis (MAST), metritis (METR), retained placenta (RETP) and cow livability and their tissue specific expression.

Trait	Gene	Chr	Start	End	M_GWAS P-value	M_Causality P-value	PPC	Type	Tissue-Specific Expression
DSAB	*PLXNA4*	4	96,574,369	97,120,718	4.5×10^{-7}	6.5×10^{-7}	0.49	protein_coding	Aorta, Liver, Trachea
DSAB	*FANCC*	8	83,022,522	83,228,696	1.3×10^{-7}	2.1×10^{-9}	0.79	protein_coding	Thyroid
DSAB	*NTM*	29	35,153,012	36,117,726	1.3×10^{-7}	1.7×10^{-7}	0.99	protein_coding	Central Nervous System
	LOC107133096	14	2,760,093	2,762,878	2.0×10^{-9}	5.9×10^{-8}	0.92	lncRNA	-
KETO	*PARP10*	14	2,024,509	2,031,477	7.0×10^{-7}	1.7×10^{-5}	0.16	protein_coding	-
	DGAT1	14	1,795,425	1,804,838	1.0×10^{-6}	1.7×10^{-5}	0.08	protein_coding	Bone Marrow
KETO	*LOC783947*	16	7,050,445	7,055,021	1.9×10^{-8}	1.3×10^{-8}	1.00	lncRNA	-
MAST	*GC*	6	88,687,845	88,739,292	2.0×10^{-7}	1.2×10^{-7}	0.15	protein_coding	Kidney, Cortex, Liver
METR	*COBL*	4	4,494,925	4,795,904	4.3×10^{-3}	7.7×10^{-4}	1.00	protein_coding	-
	LOC100296627	4	32,573,079	32,613,237	7.6×10^{-7}	4.0×10^{-13}	1.00	protein_coding	-
REPT	*MALSU1*	4	32,051,590	32,077,036	7.5×10^{-4}	$1,1\times10^{-13}$	0.98	protein_coding	-
	ABCB1	4	33,013,208	33,095,708	6.3×10^{-1}	8.4×10^{-3}	0.28	protein_coding	-
REPT	*TMEM182*	11	7,449,519	7,492,871	9.0×10^{-8}	9.9×10^{-8}	0.96	protein_coding	Heart, Muscle, Tongue
REPT	*LOC783493*	18	63,799,608	63,803,213	8.3×10^{-3}	1.2×10^{-5}	0.94	Pseudogene	-
Livability	*ABCC9*	5	8,867,2047	88,834,491	1.5×10^{-10}	1.5×10^{-10}	1.00	protein_coding	Aorta, Atrium, Lung, Muscle Uterine myometrium, Ventricle
Livability	*GC*	6	88,687,845	88,739,292	1.9×10^{-17}	1.4×10^{-19}	0.03	protein_coding	Kidney, Cortex, Liver
Livability	*ZFAT*	14	8,144,774	8,305,775	2.1×10^{-5}	3.2×10^{-5}	0.23	protein_coding	-
Livability	*LOC618463*	18	57,587,990	57,594,549	1.7×10^{-20}	3.1×10^{-20}	0.20	protein_coding	-
Livability	*CCDC88C*	21	56,645,629	56,773,438	8.6×10^{-11}	8.9×10^{-11}	0.95	protein_coding	CD8_cell
Livability	*LOC101908667*	23	25,904,084	25,909,461	2.1×10^{-8}	7.9×10^{-9}	0.31	lncRNA	-

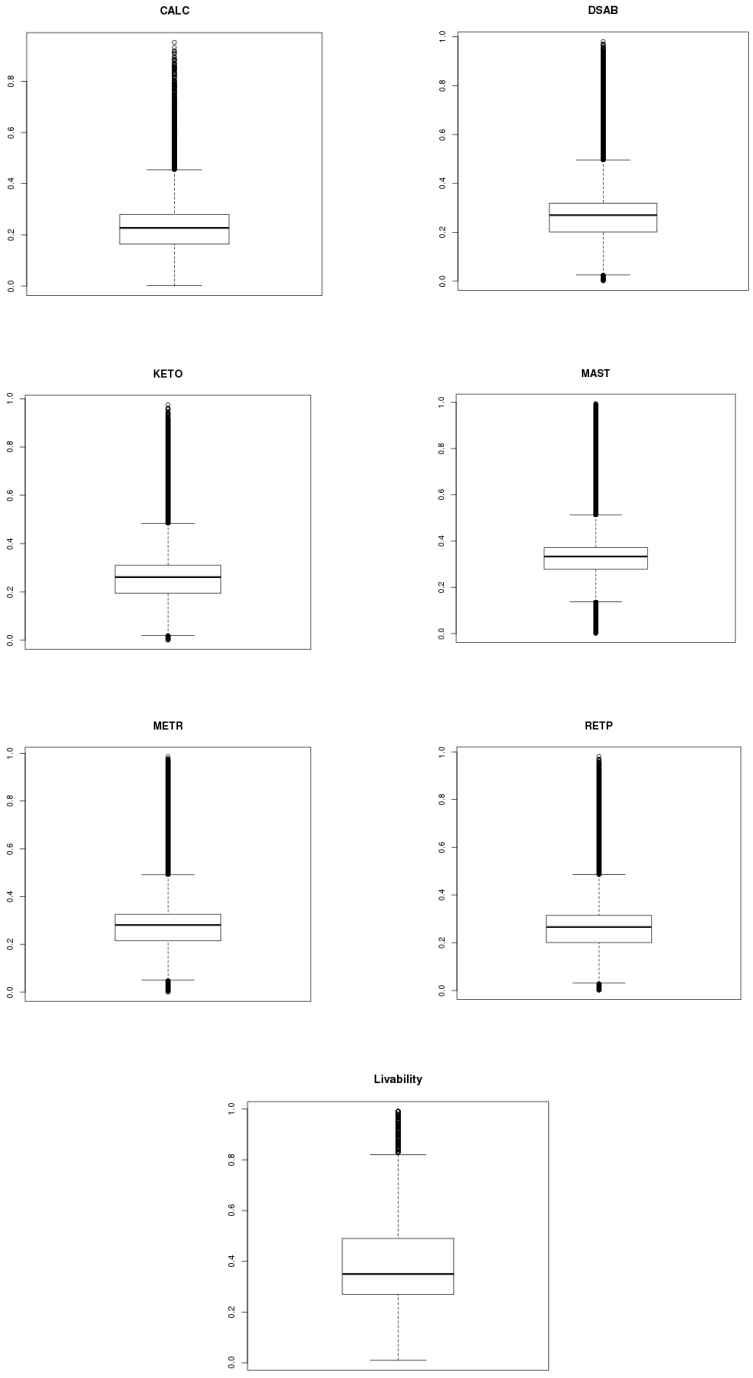
Figures

Figure 1. Manhattan plot for hypocalcemia (CALC), displaced abomasum (DSAB), ketosis (KETO), mastitis (MAST), metritis (METR), retained placenta (RETP) and cow livability. The genome-wide threshold (red line) corresponds to the Bonferroni correction.

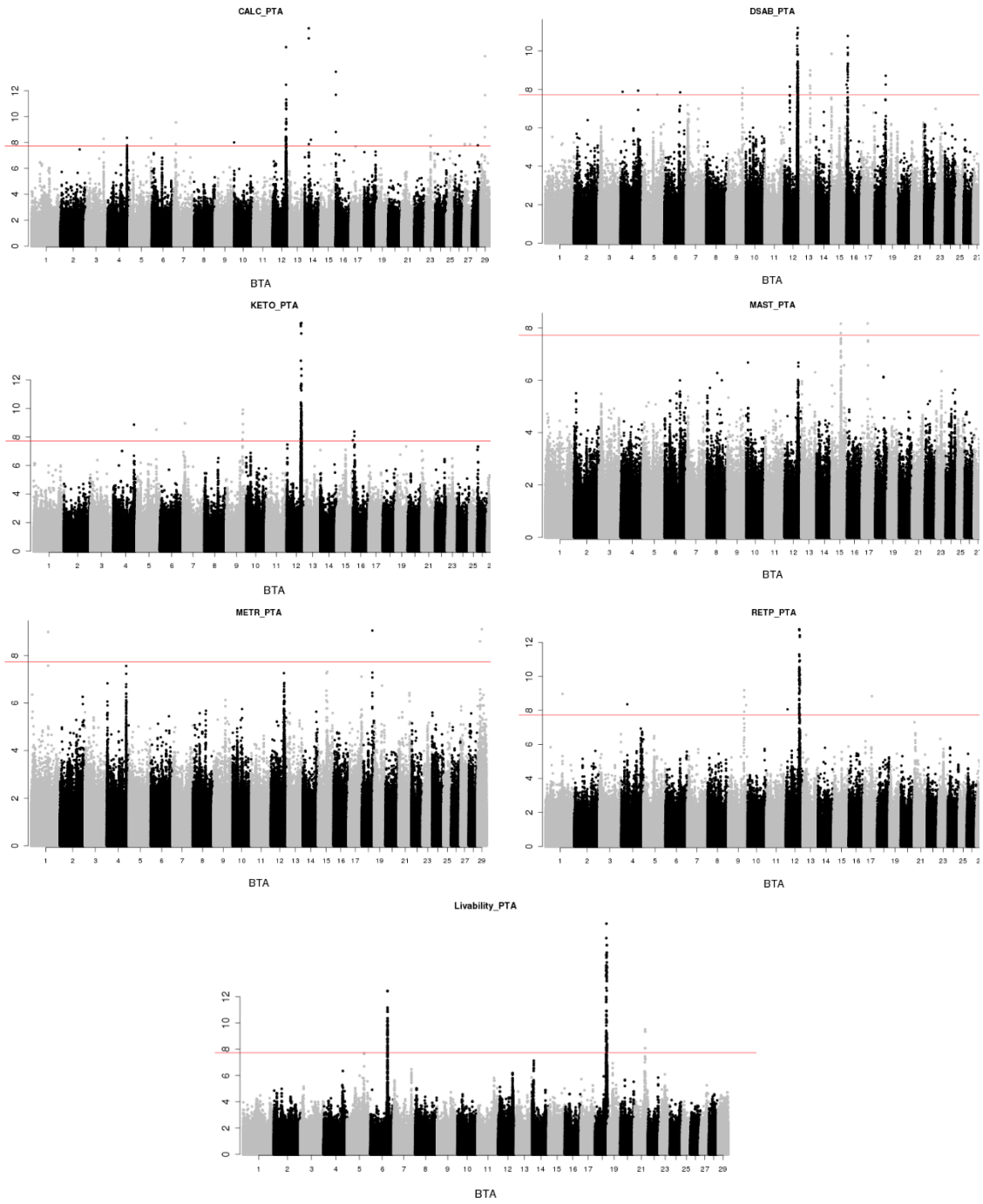


Additional Files

Additional File 1: Boxplot with PTA reliability for hypocalcemia (CALC), displaced abomasum (DSAB), ketosis (KETO), mastitis (MAST), metritis (METR), retained placenta (RETP) and cow livability.



Additional File 2: Manhattan plot using the PTA as phenotype for hypocalcemia (CALC), displaced abomasum (DSAB), ketosis (KETO), mastitis (MAST), metritis (METR), retained placenta (RETP) and cow livability. The genome-wide threshold (red line) corresponds to the Bonferroni correction for a nominal *P*-value = 0.05.



Additional File 3: List of variants into genes with highest posterior probability of causality that are mostly associated with displaced abomasum (DSAB), ketosis (KETO), mastitis (MAST), metritis (METR), retained placenta (RETP) and cow livability.

Trait	Gene	Variant	MAF	GWAS *P*-value	Casualty *P*-value	N_PPC	Annotation
DSAB	*PLXNA4*	4:97101981	0.02	4.5×10^{-7}	6.5×10^{-7}	0.26	Intergenic
DSAB	*FANCC*	8:83052202	0.11	1.3×10^{-7}	2.1×10^{-9}	0.19	Intron
DSAB	*FANCC*	8:83083934	0.11	1.3×10^{-7}	2.1×10^{-9}	0.19	Intron
DSAB	*FANCC*	8:83123780	0.11	1.3×10^{-7}	2.1×10^{-9}	0.19	Intron
DSAB	*NTM*	29:35977236	0.07	1.3×10^{-7}	1.7×10^{-7}	0.91	Intergenic
KETO	*LOC107133096*	14:2762595	0.03	2.0×10^{-9}	5.9×10^{-8}	0.89	Upstream/ Intergenic
	PARP10	14:2025096	0.23	7.0×10^{-7}	1.7×10^{-5}	0.02	Missense/ Downstream
		14:2026646	0.23	7.0×10^{-7}	1.7×10^{-5}	0.02	Upstream/Intron
	DGAT1	14:1802266	0.22	1.0×10^{-6}	2.1×10^{-5}	0.02	Missense/ Downstream
		14:1804647	0.22	1.0×10^{-6}	2.1×10^{-5}	0.02	Downstream
MAST	*GC*	6:88718227	0.45	2.0×10^{-7}	1.2×10^{-7}	0.13	Intron
METR	*COBL*	4:4643092	0.13	4.3×10^{-3}	7.7×10^{-4}	0.04	Intron
REPT	*LOC100296627*	4:32578298	0.22	7.6×10^{-7}	4.0×10^{-13}	1	Intron
	MALSU1	4:32057434	0.17	7.5×10^{-4}	$1{,}1\times10^{-13}$	0	Upstream/ Intergenic
	ABCB1	4:33063807	0.12	6.3×10^{-1}	8.4×10^{-3}	0	Intron
REPT	*QDPR*	6:117727506	0.024	2.2×10^{-4}	2.3×10^{-5}	0	3_prime_UTR_ variant
REPT	*QDPR*	6:117727635	0.024	2.2×10^{-4}	2.3×10^{-5}	0	3_prime_UTR_ variant
REPT	*QDPR*	6:117727851	0.024	$2{,}2\times10^{-4}$	2.3×10^{-5}	0	3_prime_UTR_ variant
REPT	*QDPR*	6:117728023	0.024	2.2×10^{-4}	2.3×10^{-5}	0	Missense
REPT	*QDPR*	6:117743248	0.024	2.2×10^{-4}	2.3×10^{-5}	0	Missense
REPT	*TMEM182*	11:7465110	0.06	9.0×10^{-8}	9.9×10^{-8}	0.95	Intron
REPT	*LOC783493*	18:63802274	0.24	8.3×10^{-3}	2.3×10^{-3}	0	Intron
Livability	*ABCC9*	5:88823164	0.47	1.5×10^{-10}	1.5×10^{-10}	0.51	Splice/Intron
Livability	*GC*	6:88718227	0.45	1.9×10^{-17}	1.4×10^{-19}	0.03	Intergenic
Livability	*ZFAT*	14:8249314	0.04	2.1×10^{-5}	3.0×10^{-5}	0	Intron
Livability	*LOC618463*	18:57589121	0.07	1.7×10^{-20}	3.1×10^{-20}	0.19	Intron
Livability	*CCDC88C*	21:56700449	0.01	8.6×10^{-11}	8.9×10^{-11}	0.46	Intron
Livability	*LOC101908667*	23:25905667	0.02	2.1×10^{-8}	7.9×10^{-9}	0.08	Intergenic

N_PPC = Normalized posterior probability of causality

Chapter 3: Genome changes due to selection in U.S. dairy cattle

Abstract

Genetic and genomic selection in the U.S. dairy population has achieved successful phenotypic improvement across a comprehensive set of economically important traits that involve production, reproduction, health, and body conformation. While contemporary cows differ phenotypically from their ancestors hundreds of years ago, the changes in the genome, especially those due to selection, remain to be discovered. The aim of this study was to investigate genome-wide and region-specific changes in the U.S. Holstein-Friesian (HF) cattle population between the years of 1950 to 2015. Using the U.S. dairy cooperator's phenotypic and genomic databases hosted at CDCB, we first extracted genotype and phenotype (PTA) data of 27,000 reference bulls and performed GWAS analyses to identify candidate QTLs. We then divided the 27,000 Holstein bulls into nine bins based on birth year, before 1980, 1980-1990, 1990-1995, 1995-2000, 2000-2002, 2002-2005, 2005-2007, 2007-2010, and after 2010. The allele frequency changes between the two extreme time periods were calculated to capture the difference between the earliest and most recent populations. Finally, the genomic regions with the largest allele frequency changes were compared against the QTL regions identified in GWAS analyses.

To identify true genome changes due to selection from those due to random genetic drift, we implemented a gene dropping simulation approach with real pedigree and calculated thresholds of allele frequency change. The process was executed by running a simulation program in R software, which visualizes systemic changes over

individual SNPs and compares them to a distribution under pure genetic drift. Observation of changes above the 99.9% threshold on the distribution may be indicative of selection and affecting dairy traits. From this evaluation of genome-wide and region-specific changes due to selection, we identified candidate QTL regions under selection and that are associated with economically important traits in the U.S. dairy population.

Introduction

Genome selection is a method of marker-assisted selection for quantitative traits that is widely used in breeding programs. This approach offers the opportunity to estimate breeding values without reliance on pedigree and phenotype information, which has been necessary in traditional breeding methods (Doekes *et al*, 2018).

While marker-assisted selection has the potential to predict the genetics of quantitative traits more accurately, and thus increase genetic gain, the implementation of this method has been limited. An alternative method is genetic selection (GS), which uses all of the possible markers simultaneously. This method estimates the marker effects across the whole genome, and in doing so, also provides a reliable estimate of the genomic breeding value (GEBV) (Heidaritabar *et al*, 2014). The use of GEBVs can alleviate the requirement to phenotype an entire population per generation, which will reduce the time, effort, and costs typically needed by breeding programs (Spindel *et al*, 2016).

The current availability of the bovine reference sequence allowed for the systematic assessment of many single nucleotide polymorphisms (SNPs) by whole genome re-sequencing of animals (Nguyen *et al*, 2015). Due to the accessibility of SNP data, it is now possible to monitor and analyze the influence of selection within a population. Thus, studying the genetic diversity of dairy cattle may contribute to future selection responses as well as increase the frequency of beneficial alleles in the population.

The aim of this study was to evaluate genome-wide diversity in Holstein-Friesian (HF) bulls from 1950 to 2015. Using a USDA dairy cattle dataset, we first investigated whether allele frequency changes are attributed to selection. To identify potential regions under selection, we compared observed allele frequency changes to those expected under pure genetic drift. Additionally, we wanted to evaluate whether the regions under selection were linked to known quantitative trait loci (QTL). To compare the allele frequency differences, we applied a Wilcoxon rank-sum test to the SNP and QTL data. From this evaluation, we will have demonstrated the genome changes due to selection that are present in U.S. dairy cattle.

Results and Discussion

Using a USDA dairy cattle dataset, we extracted genotype and phenotype (PTA) data of 27,214 Holstein-Friesian bulls. The sequence data we imputed from the dataset contains a total of 3 million sequence variants. This large dataset allowed us to run our GWAS and gene simulation program with high power and precision.

Mixed model GWAS

We applied the software MMAP (O'Connell, 2013) to run a mixed model GWAS on our dataset. Association signals were detected from GWAS, and we selected the variants with the smallest *P*-value. Four plots with clear trends are shown in Figure 1. In these plots, the frequency of the sequence variant was observed to change consistently over time.

We identified 16 candidate variants to be under selection in a U.S. Holstein-Friesian cattle population. One of the candidate variants was found on chromosome 1 (Chr1:82024433), which corresponds to rump angle from a single-trait cattle GWAS. Its nearby gene of interest is *IGF2BP2*, which is the insulin-like growth factor 2 mRNA binding protein 2. This protein coding gene is a tumor promoter that drives cancer proliferation (Xu *et al*, 2019). Another candidate variant identified to have a consistent trend was on chromosome 4 (Chr4:113758201) that corresponds to sire stillbirth. A reported gene in this region is *GIMAP7*, which is a member of the GTPase of an immunity-associated protein family. It functions to regulate lymphocyte survival and homeostasis (Schwefel *et al*, 2013). A candidate variant observed on chromosome 5 (Chr5:112474824) has QTLs involved in milk and protein traits. The fourth plot is of the variant located on chromosome 29 (Chr29:41819469) that corresponds to daughter

stillbirth trait from a single-trait GWAS. In addition, the region around this variant has the gene *STX5*, which encodes a syntaxin or target-SNAP receptor. The encoded protein serves to regulate endoplasmic reticulum (ER) to Golgi transport and has a critical role in autophagy (Linders *et al*, 2019).

Gene dropping simulation program

We applied a gene dropping simulation program to verify if the candidate variants we detected from GWAS were consistent with selection. As shown in Figure 2, the red points represent the percentile of each MAF in a bin. The individual blue points denote a single gene drop. We observed a large cluster of gene drops below the 99% threshold ($P < 0.001$), which helped us distinguish the gene drops above the line of the empirical gene drop distribution. These blue dots were indicative of selection and were further analyzed in the following procedures.

Manhattan plot

As shown in Figure 3, we created a Manhattan plot of the absolute allele frequency differences of the variants. We overlapped the top 288 QTLs that were previously identified on the plot as green points. These points were included to visualize if there were regions with allele frequency differences that overlapped with QTLs in the GWAS. We observed a few significant peaks in the Manhattan plot, which we then compared to previously reported regions. For instance, there appeared to be a correspondence between the peak on chromosome 11 with the QTL data. A prior study by Lund *et al* (2007) identified a QTL on bovine chromosome 11 (BTA11) for somatic cell score (SCS). Breeding schemes have used SCS as an indicator trait for clinical mastitis (CM) due to

the moderate to high genetic correlation that exists between CM and SCS (Heringstad *et al*, 2006). Thus, this overlap may be of interest for further research.

Wilcoxon rank-sum test

We conducted the Wilcoxon rank-sum test to compare the MAF with the 3 million SNPs data and to detect if there is a statistically significant difference between the two distributions. The mean of the difference in allele frequency for the SNP and QTL bins and the p-values per bin are shown in Table 1. The Wilcoxon rank-sum test shows statistical significance with a p-value of 1.236×10^{-5} and 1.017×10^{-5} for bins 2 and 3, respectively. This finding indicates there is a difference between the two distributions at the *P*-value threshold of 0.05 for significance. Further analysis of the SNPs and QTLs in those bins may provide insight to this study.

Conclusions and Future Directions

Using a USDA dairy cattle dataset, we identified candidate variants that are under selection by monitoring their allele frequency change over time. These candidate variants are associated with biological traits and economically important traits in cattle, and we found their nearby gene of interest. Additionally, we demonstrated that gene dropping is an applicable method to investigate genome-wide and region-specific changes in the cattle genome through time.

For future research directions, we propose the development of a gene dropping simulation program written in Python. The resulting Δp-distribution under pure genetic drift can be compared for consistency to our current gene dropping plot that was run in R software.

After running the proposed gene dropping simulation, genes in the distribution can be identified for significance. We expect those genes to be observed above the 99.9% threshold line ($P < 0.001$) that is shown in our simulation plot. The genes above the threshold line will be indicative of selection and can be further studied for their effects on genomic selection in dairy cattle.

Methods

Genotype and phenotype data

This study used U.S. dairy cattle data hosted in the National Cattle Genomics database at the Council on Dairy Cattle Breeding (CDCB). Specifically, we studied a U.S. Holstein-Friesian (HF) cattle population between the years of 1950 to 2015. The genotype and phenotype (PTA) data of 27,214 Holstein bulls were extracted by imputation. The extracted sequence data contains a total of 3 million variants.

Mixed model GWAS

We used the mixed model approach that is implemented in the software MMAP (O'Connell, 2015) to run GWAS on our dataset. The mixed model used in our GWAS allowed us to identify sequence variants with consistent allele frequency changes over time. In our model, time is the response variable and variant genotypes is the independent variable. An association signal was detected if the frequency of the sequence variant changes consistently over time. After running GWAS, the sequence variants with the smallest p-value were selected. Four plots of those variants are shown in Figure 1. Specifically, the 27,000 Holstein bulls were divided into nine time bins according to birth year. The time ranged from before 1980 to after 2010. In each of the plots, genotype means were plotted over this time range.

Gene dropping simulation program

We implemented a gene dropping simulation approach in R software to identify the genome changes due to selection from those due to random genetic drift. Gene dropping

is a simulation program that visualizes systemic changes over individual SNPs. Figure 2 shows the resulting plot from this simulation. We divided the data into two time bins, from 1985-1990 to 2009-2015, each containing 3,000 animals. Changes in allele frequency over time were computed as: $\Delta p = p_t - p_0$, with p_t and p_0 being the frequency in the first (1985-1990) and last (2010 to 2015) 5-year periods, respectively. Observed Δp-values were compared to those expected from a distribution under pure genetic drift obtained by the gene dropping simulation. The allele frequency differences were plotted on this graph and compared to a 99.9% threshold (P <0.001), which is shown in red. This represented an expected distribution from pure genetic drift. Thus, the absolute Δp-values observed above the red threshold line were indicative of selection.

Manhattan plot

To visualize where these variants were in the genome, we made a Manhattan plot, which is shown in Figure 3. The absolute allele frequency differences were plotted across 29 chromosomes and overlapped with GWAS results, which were highlighted in green. The overlapping points allowed us to detect whether any SNPs were related to milk production or disease. Additionally, the analysis allowed us to investigate whether regions under selection were linked to the top 288 QTLs.

Wilcoxon rank-sum test

We extracted the top 288s QTLs from the 3 million SNPs data and compared the allele frequency differences to the remaining SNPs using the Wilcoxon rank-sum test. The results are shown in Table 1, where we divided the data into 20 MAF bins based off a numerical cutoff. We performed the Wilcoxon rank-sum test within each bin to verify if

the two distributions were different. Additionally, we computed the mean of the difference in allele frequency for the SNP and QTL bins. The p-value per bin was calculated to identify significance ($P < 0.05$).

References

1. Doekes, H.P., Veerkamp, R.F., Bijma, P., Hiemstra, S.J., and J.J. Windig. “Trends in the genome-wide and region specific genetic diversity in the Dutch-Flemish Holstein-Friesian breeding program from 1986 to 2015.” *Genetics Selection Evolution* (2018) 50:15. *BMC*. Web. 26 Sept 2018.

2. Heidaritabar, M., Vereijken, A., Muir, W.M., Meuwissen, T., Cheng, H., Megens, H.J., Groenen, M.A.M., and J.W.M. Bastiaansen. “Systematic differences in the response of genetic variation to pedigree and genome-based selection methods.” *Heredity* (2014) 113: 503-513.

3. Spindel, J., Begum, H., Akdemir, D. *et al*. Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity* **116**, 395–408 (2016).

4. Nguyen T. T., Huang, J.Z., Wu, Q., Nguyen, T.T., and M.J. Li. “Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests.” *BMC Genomics* (2015) 16: 21-23.

5. O’Connell, J. R. 2015. MMAP User Guide. Available: http://edn.som.umaryland.edu/mmap/index.php.

6. Xu, X., Yu, Y., Zong, K. *et al*. Up-regulation of IGF2BP2 by multiple mechanisms in pancreatic cancer promotes cancer proliferation by activating the PI3K/Akt signaling pathway. *J Exp Clin Cancer Res* **38**, 497 (2019).

7. Schwefel, D., B. S. Arasu, S. F. Marino, B. Lamprecht, K. Ko¨chert, E. Rosenbaum, J. Eichhorst, B. Wiesner, J. Behlke, O. Rocks, *et al*. 2013. Structural insights into the mechanism of GTPase activation in the GIMAP family. *Structure* 21: 550–559.

8. Linders PT, van der Horst C, ter Beest M, van den Bogaart G. Stx5-Mediated ER-Golgi Transport in Mammals and Yeast. *Cells*. 2019; 8(8):780.

9. Lund MS, Sahana G, Andersson-Eklund L, *et al*. Joint analysis of quantitative trait loci for clinical mastitis and somatic cell score on five chromosomes in three Nordic dairy cattle breeds. *J Dairy Sci.* 2007 Nov;90(11):5282-90.

10. Heringstad B, Gianola D, Chang YM, Odegård J, Klemetsdal G. Genetic associations between clinical mastitis and somatic cell score in early first-lactation cows. *J Dairy Sci.* 2006 Jun;89(6):2236-44.

Table

Table 1. Mean difference in allele frequency in QTL and control bins. The top 288 QTLs were extracted from the 3 million SNPs data and compared to the remaining SNPs using the Wilcoxon Rank Sum Test. Each dataset was separated into 20 distinct bins using a numerical cutoff. The mean of the difference in allele frequency was calculated per bin. The Wilcoxon test was run within each bin based on allele frequency difference. P-values were calculated using a significance level of 0.05.

Bins	Mean of SNP bin	Mean of QTL bin	P-value per bin
1	0.0028	0.0099	0.0003
2	0.0083	0.0278	1.236e-05
3	0.0128	0.0521	1.017e-05
4	0.0176	0.0315	0.0014
5	0.0221	0.0332	0.0156
6	0.0266	0.0336	0.1482
7	0.0310	0.0288	0.8609
8	0.0361	0.0971	0.0229
9	0.0405	0.0912	0.0023
10	0.0457	0.0953	0.0012
11	0.0496	0.0853	0.0070
12	0.0545	0.0854	0.5718
13	0.0593	0.0656	0.6831
14	0.0638	0.0933	0.1551
15	0.0673	0.0650	0.9017
16	0.0704	0.0845	0.3275
17	0.0731	0.1009	0.1724
18	0.0741	0.0712	0.6594
19	0.0761	0.0822	0.8713
20	0.0775	0.0760	0.9921

Figures

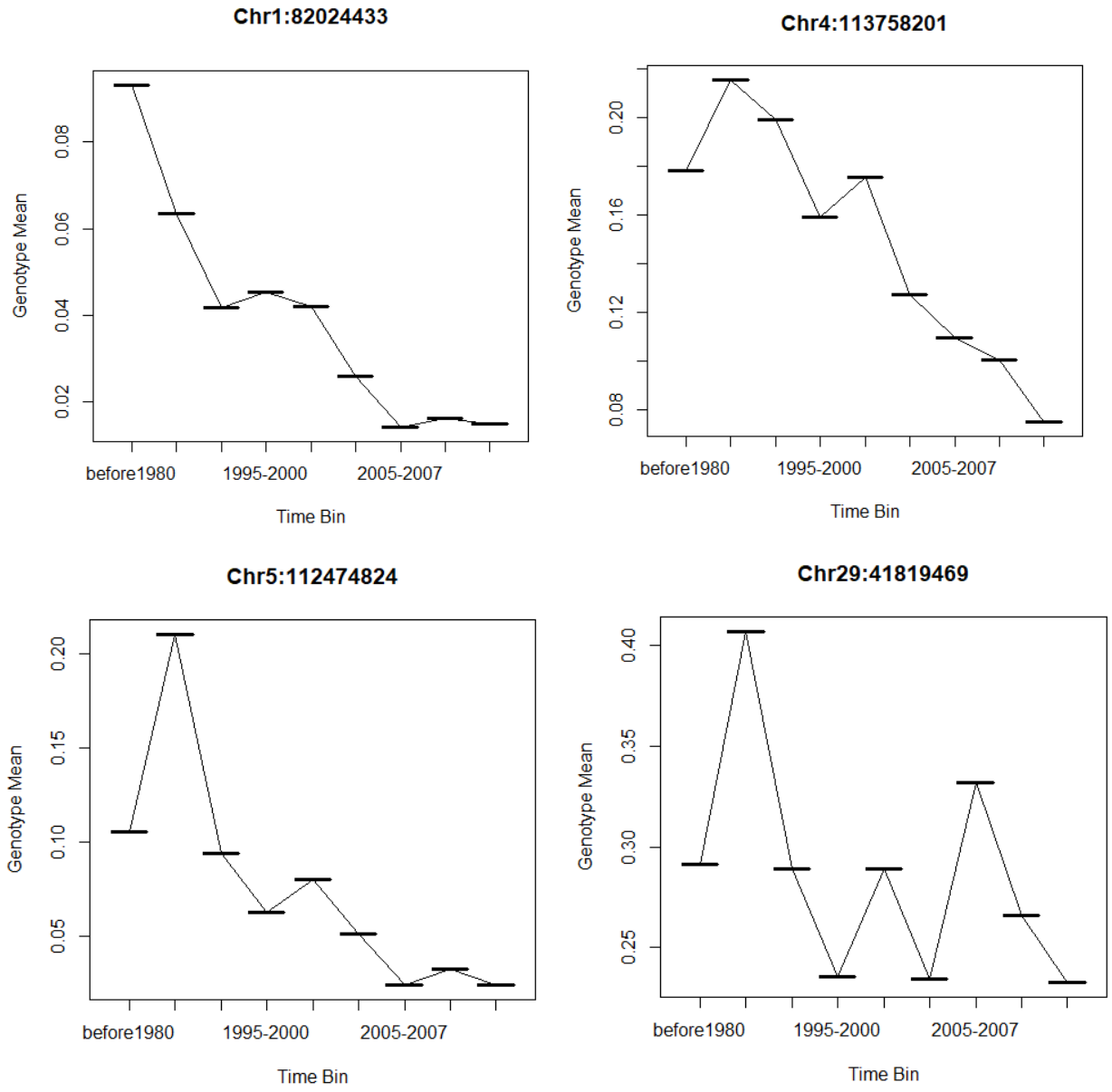


Figure 1. Sequence variants with consistent allele frequency changes over time. 27,000 Holstein bulls were divided into nine timebins based on birth year. Genotype means were plotted over time, which began before 1980 to after 2010.

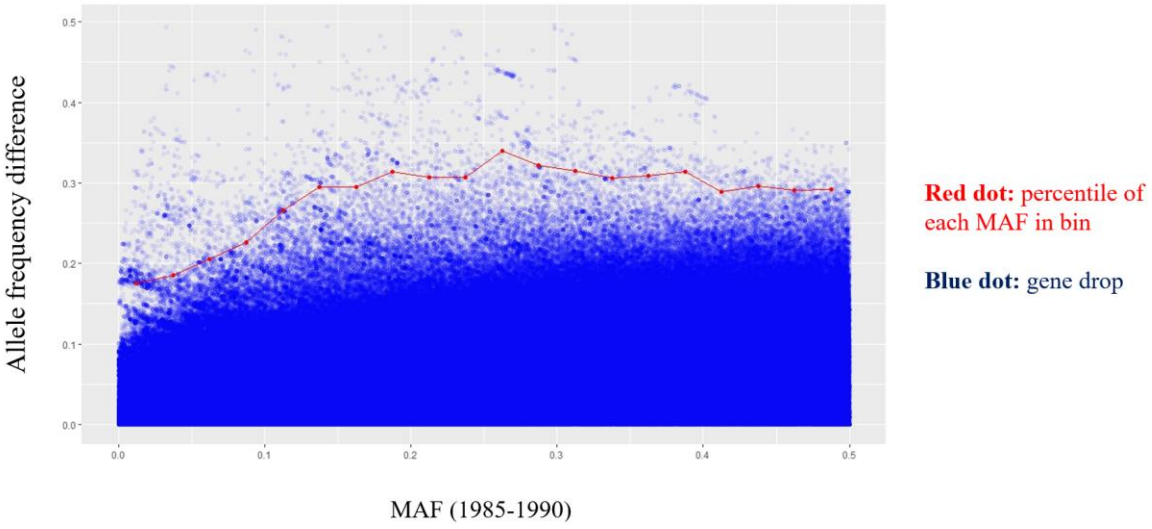


Figure 2. Plot of absolute frequency changes from 1985-1990 to 2009-2015 observed in data and gene dropping. Changes for different minor allele frequencies (MAF) are shown in the 1985-1990 period using MAF classes of 0.5%. The red line is the 99.9% threshold of the gene drop distribution per MAF class.

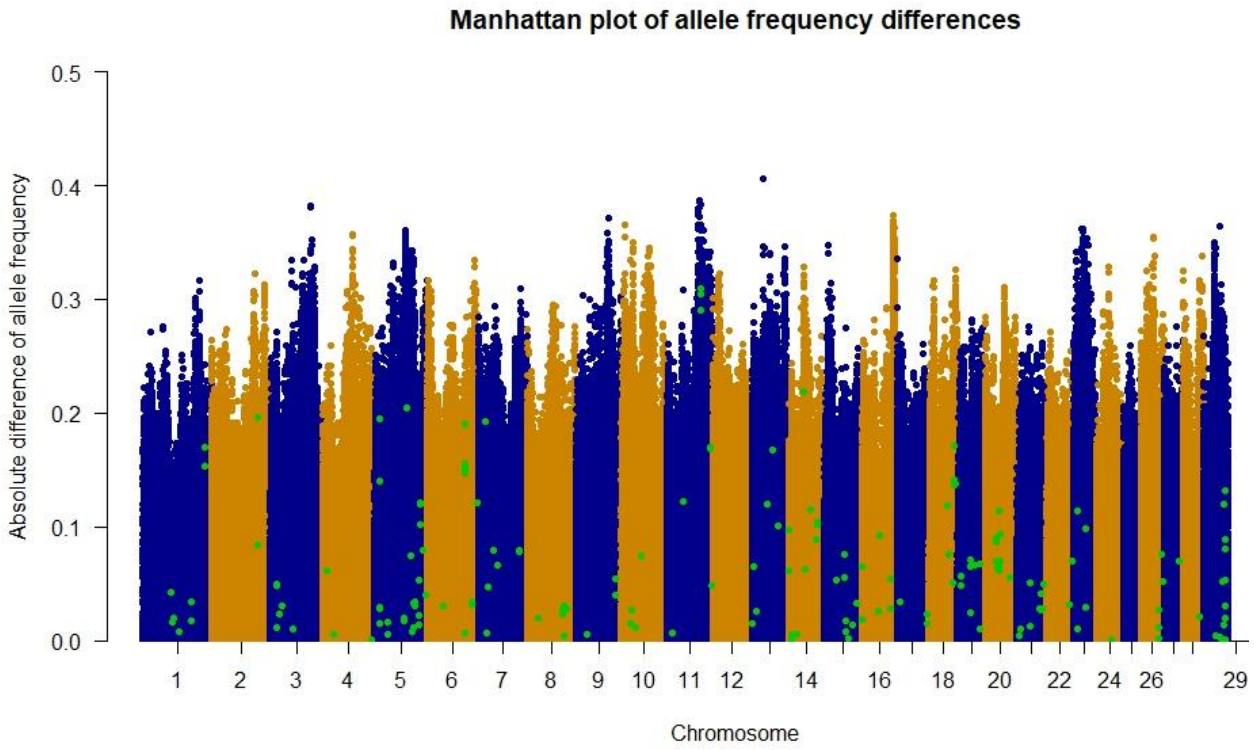


Figure 3. A Manhattan plot of the absolute allele frequency differences across 29 chromosomes. Green points are the top 288 QTLs identified previously.

Chapter 4: Effect of temperature and maternal age on recombination rate

Abstract

Meiotic recombination is a fundamental biological process that facilitates meiotic division and promotes genetic diversity. Recombination is phenotypically plastic and affected by both intrinsic and extrinsic factors. The effect of maternal age on recombination rates has been characterized in a wide range of species, but the effect's direction remains inconclusive. Additionally, the characterization of temperature effects on recombination has been limited to model organisms. Here we seek to comprehensively determine the impact of genetic and environmental factors on recombination rate in dairy cattle. Using a large cattle pedigree, we identified maternal recombination events within 305,545 three-generation families. By comparing recombination rate between parents of different ages, we found a quadratic trend between maternal age and recombination rate in cattle. In contrast to either an increasing or decreasing trend in humans, cattle recombination rate decreased with maternal age until 65 months and then increased afterward. Combining recombination data with temperature information from public databases, we found a positive correlation between environmental temperature during fetal development of offspring and recombination rate in female parents. Finally, we fitted a full recombination rate model on all related factors, including genetics, maternal age, and environmental temperatures. Based on the final model, we confirmed the effect of maternal age and environmental temperature during fetal development of offspring on recombination rate with an estimated heritability of 10% (SE = 0.03) in cattle.

Collectively, we characterized the maternal age and temperature effects on recombination rate and suggested the adaptation of meiotic recombination to environmental stimuli in cattle. Our results provided first-hand information regarding the plastic nature of meiotic recombination in a mammalian species.

Keywords: Recombination, Maternal age, Temperature, Cattle, Genetics

Note: This chapter was previously published by Frontiers in Genetics. Ellen Freebern is a co-first author and made a significant contribution to the analysis and writing of this paper. The original citation is as follows:

Shen B, Freebern E, Jiang J, Maltecca C, Cole JB, Liu GE, Ma L. Effect of Temperature and Maternal Age on Recombination Rate in Cattle. *Front Genet.* 2021 Jul 20;12:682718.

Introduction

Meiotic recombination is an essential process that occurs in all sexually reproducing organisms. This process facilitates the pairing and alignment of homologous chromosomes during prophase, which leads to the formation of crossovers. These crossover events transfer genetic information between the maternal and paternal homologs, and in doing so, ensures that each offspring will receive a unique combination of parental genomes. The extent and pattern of genetic reshuffling has important implications for evolution and population genetics, as well as breeding. However, errors in meiotic recombination can lead to aneuploidy, chromosomal abnormalities, and other deleterious outcomes (Hassold and Hunt, 2001; Lipkin *et al*., 2002). Thus, meiotic recombination must be well-regulated by cellular processes to prevent disturbances in the recombination pathway. It has been found that various factors influence meiotic recombination patterns in human and animal genomes. For instance, genome-wide association studies (GWAS) have identified genes and genetic variants associated with recombination features in humans (Kong *et al*., 2008; Chowdhury *et al*., 2009), mice (Baudat *et al*., 2010), cattle (Sandor *et al*., 2012; Ma *et al*., 2015; Shen *et al*., 2018), and sheep (Johnston *et al*., 2016). Some of the genes from those studies, including *RNF212*, *CPLX1*, and *PRDM9*, have been reported to be associated with individual-level recombination rates across multiple mammalian species.

A variety of intrinsic and external factors affect recombination rates across individuals and populations. These factors can be derived from environmental conditions, such as temperature, or physiological and stressful conditions, such as starvation. The resulting changes in recombination rates pose benefits or consequences to the health and

evolution of a species. Many studies have suggested that maternal age's intrinsic factor has a significant effect on recombination rate, but the direction of the effect is still debatable (Polani and Jagiello, 1976; Hussin *et al*., 2011; Martin *et al*., 2015). A recent multicohort analysis in humans reported a small but significant positive effect of maternal age on recombination rate, which contradicted previous studies in other human population (Martin *et al*., 2015). In mice and hamsters, a negative effect of maternal age was observed (Polani and Jagiello, 1976). However, no effect of maternal age was found for recombination rate in wild sheep (Johnston *et al*., 2016), but an increase was reported in swine (LozadaSoto *et al*., 2021). Also, extensive studies of the maternal age effect have been conducted in *Drosophila*, worms, plants, and yeast, but no consistent conclusion have been reached (Hunter *et al*., 2016; Modliszewski and Copenhaver, 2017). As for the paternal side, many studies reported no effect of paternal age on meiotic recombination (Griffin *et al*., 1995; Hussin *et al*., 2011). Although the biological reason remains unclear for the effect of maternal age on recombination, there are some proposed explanations for both directions of the effect. The positive effect can be explained by a selection hypothesis: the factors related to aneuploidies increase with maternal age, so eggs with more crossovers are more likely to overcome these and give a live birth in older mothers (Kong *et al*., 2004). The negative effect can be explained by another hypothesis that specific meiotic configurations are less likely to be properly processed with increasing maternal age, so recombination rate decreases with maternal age (Hassold *et al*., 1995).

Extrinsic factors, such as temperature and nutrient conditions, have also been found to influence meiotic recombination rates. In *Drosophila*, environmental stressors

such as exposure to Ethylenediaminetetraacetic acid (EDTA) or nutritional deficiency were observed to increase the recombination rate (Levine, 1955). Similarly, in the budding yeast, *Saccharomyces cerevisiae*, a lack of nutritional resources resulted in an increased recombination rate (Abdullah and Borts, 2001). However, the effect of temperature on meiotic recombination rates is more complicated. Some studies reported a positive correlation in *Arabidopsis thaliana, Caenorhabditis elegans, and Melanoplus femurrubrum* (Church and Wimber, 1969; Rose and Baillie, 1979; Francis *et al*., 2007), while other studies found a negative correlation in *Allium ursinum* (Loidl, 1989).

Additionally, both positive and negative correlations were detected in *Drosophila* (Stern, 1926). For instance, a recent study in *Drosophila melanogaster* reported a non-linear increase in meiotic recombination frequency in response to increased exposure to heat shock conditions (Jackson *et al*., 2015). This finding suggests that *Drosophila* can plastically modulate their recombination rate in response to environmental conditions, thus conferring greater adaptive potential to their offspring. In cattle, decreases in fertility rate have occurred due to the major factor of heat stress. In fact, Holstein cattle's conception rate in the summer season is 20–30% less than in the winter season (Cavestany *et al*., 1985).

The formation of the Animal Genomics and Improvement Laboratory (AGIL) has facilitated the development of genetic evaluations for economically important traits in United States dairy cattle. Such studies can enhance research to improve the health and efficiency of cattle, including the study of recombination features across multiple cattle breeds with high statistical power. Using the large cattle genomic database maintained by AGIL and the Council on Dairy Cattle Breeding (CDCB), we have previously

characterized the recombination features and their genetic control in dairy cattle. As mounting evidence has shown, meiotic recombination rates respond to both intrinsic and extrinsic conditions. Therefore, this study aims to determine how recombination rates vary in relation to advancing maternal age and common environment factors, such as temperature.

Results and Discussion

Identification of recombination events in genotyped cattle pedigree

Using a method developed in our previous studies (Ma *et al*., 2015; Wang *et al*., 2016), we identified recombination events by constructing three-generation families from a large cattle pedigree that included an offspring, parents, and grandparents. We phased the SNP genotypes of the focal offspring and its parents within each three-generation family. We then inferred maternal crossover events by comparing phased genotypes between dam-offspring pairs. To ensure optimal data quality, we excluded the X chromosome and used the SNP coordinates that have been corrected for potential mapping errors (Null *et al*., 2019; Rosen *et al*., 2020). In total, we extracted 305,545 three-generation families and identified 6,677,618 maternal crossover events. All the animals have birth dates available, so we used the age of parent at birth of the focal offspring to study the effect of maternal age. Farm location and temperature information were available for 36,999 parents, which were included in the temperature effect analysis.

Effects of maternal age on recombination rate in cattle

Previous studies have suggested a relationship between maternal age and the number of recombination events in plants, mice, and humans. However, even within the same

species, the direction of this correlation remains inconsistent. To investigate how recombination rates are related to maternal age in cattle, we first modeled this relationship with a continuous variable of maternal age and the recombination rate residuals in cattle (Figure 1A). The recombination rate residuals were obtained by adjusting recombination rates with SNP chip density and the number of informative SNP markers within each of the 305,545 three-generation families. As a result, we observed a quadratic trend where recombination rate initially decreased from 20 to 65 months old parents and then increased as the cow grew older than 65 months. Note that we have more statistical support for the decreasing trend of recombination rate from 20 to 65 months since it consists of 91.8% of our records with much smaller standard errors. Parents that gave birth between 65 and 100 months old consist of 21,798 (7.1%) cases of our data, and we only have 3,321 (1.1%) cows giving birth above 100 months age. Still, the increasing trend after 65 months of maternal age is supported with a reasonable sample size (>25,000). This increasing trend of recombination rate in older parents is consistent with maternal age's positive effect on recombination rate in the latest multicohort human study (Martin *et al*., 2015).

Since 65 months is the age that separated two groups of cows by the direction of maternal age effect, we divided the cows into ten age groups starting from 20 months old with an increment of 10 months and plotted the relationship (Figure 1B). Consistently, the same quadratic trend was observed when using maternal age as groups. Note that the last age group consists of all the records of parents giving birth over 120 months of age. To the best of our knowledge, this is the first such study in a mammalian species that reported a U-shaped relationship between maternal age and recombination rate. However,

a quadratic relationship was also identified for the effect of temperature on the recombination rates in plants, fruit flies, and grasshoppers (Church and Wimber, 1969; Phillips *et al*., 2015), although the underlying mechanisms of meiotic recombination are completely different between cattle and these species. And this increasing of recombination after 65 months of age can also be due to culling and data representation issues because only the most fertile cows can survive on a farm for more than 65 months.

Effects of temperature on maternal recombination rates in cattle

Previous studies have shown that temperature affects meiotic recombination rates in many poikilothermic organisms, including yeast, plants, worms, grasshoppers, and large reptiles such as crocodiles (Church and Wimber, 1969; Isberg *et al*., 2006; Phillips *et al*., 2015). However, the direction of the effect remains inconclusive. Utilizing the extensive cattle pedigree data from the United States National Cooperators Database, we characterized meiotic recombination features in dairy cows and integrated them with the environmental temperature information. Using the NOAA National Weather Database, we obtained temperature data for the months when the calves were conceived by the parents of interest. In total, we have 36,999 records with both the environmental temperature and recombination rate data.

We fitted a model to explore the temperature effect on maternal recombination rates in cattle. THI (temperature humidity index) has been widely used to indicate heat or cold stress in cattle (Gaughan *et al*., 2008). An environment with THI exceeding 78 can be considered as a heat stress condition for cattle because both the productive and reproductive performance of cows would be seriously affected (Bohmanova *et al*., 2007).

There is no predetermined THI index for cold stress conditions as cattle's wellness in a cold environment depends on several factors such as management practices and their hair coats. Based on the THI index in cattle, we chose two temperatures as the threshold of hot and cold conditions. Temperatures above 26.67ºC were considered as hot conditions and temperatures below 4.44ºC as cold conditions. In this study, we tested the effect of temperature rather than THI on recombination rate as most of previous studies reported the effect of temperature rather than THI (Church and Wimber, 1969; Loidl, 1989; Phillips *et al*., 2015). Still, it is interesting to investigate whether THI may have a larger impact than temperature only on recombination in future studies.

The cows were divided into three groups based on the temperature condition during the fetal development of the offspring. Over 6 K cows were conceived under hot conditions, over 25 K cows were conceived in a mild temperature environment, and 6 K cows were conceived under cold environment. To characterize the effect of this temperature, we reported box plots of the recombination rate residuals against those three conditions (Figure 2). We observed that cows under hot temperatures during pregnancy showed an elevation of recombination rate while cows under cold conditions showed decreased recombination rate. An increase of recombination events under hot environment is consistent with many previous studies across several species, which found an increase in recombination frequency under heat stress conditions (Lim *et al*., 2008; Jackson *et al*., 2015; Modliszewski and Copenhaver, 2017; Arrieta *et al*., 2021). However, the temperature effect identified here was for the fetal development stage of the offspring, instead of the female parent. Note that it is the fetal development stage of the female parent that is crucial for meiotic recombination in mammals. Therefore, the

temperature effect reported here could possibly be due to an indirect effect on the fitness of the offspring with more or less crossovers, rather than on the meiosis process itself. Finally, the reported temperature effect can be explained by the 'production line' hypothesis for female recombination (Henderson and Edwards, 1968; Kong *et al*., 2004).

Full model analysis of recombination with genetics, maternal age, and temperature

To fully understand the effect of genetic and non-genetic factors on recombination, we fitted a full model on recombination rate with all relevant factors available in our data. We modeled the genetic or animal effect as a random effect and temperature condition as a fixed effect with three levels. We included the temperatures during two important developmental stages of a female meiosis, one during the fetal development for the offspring (first generation in a three-generation family) and the other during the fetal development of the parent (second generation). Finally, we also included maternal age, the parent's birth year, and the quadratic term of these factors in the model (Table 1). Based on the genetic effect, we estimated the recombination rate's heritability to be 10% (*SE* = 0.03), consistent with other studies in livestock animals (Sandor *et al*., 2012; Johnston *et al*., 2016; Zhang *et al*., 2020). Our results from this full model showed that hot temperature during the fetal development of the offspring would increase recombination rate ($P = 0.027$), while cold conditions were associated with decreased recombination rate ($P = 0.019$). However, the temperature (hot and cold conditions) during the fetal development of parents were not significantly associated with recombination rate ($P = 0.271$ and $P = 0.097$, respectively), although that is when female meiosis arrests. We also found that maternal age has a negative effect on recombination

rates ($P = 4.83 \times 10^{-12}$) with a significant quadratic term ($P = 5.68 \times 10^{-4}$), confirming the U-shaped relationship between maternal age and recombination rate. We also noticed that a parent's birth year would positively influence recombination rates ($P = 1.51 \times 10^{-31}$) with a significant quadratic term ($P = 1.01 \times 10^{-40}$), indicating either an increasing trend of recombination rate in the dairy cattle population or some inherent confounding in the data.

Potential Application of Recombination to Animal Breeding

Theoretically, recombination should be beneficial for the long-term efficiency of selection through increasing genetic variation (Otto and Barton, 2001). However, in a short-term period, recombination may also break the combination of beneficial alleles in the haplotypes that were selected for breeding. Recent simulation studies have shown that the effect of modifying recombination rate on the improvement of traditional genomic selection is small (Battagin *et al*., 2016). Still, our recent work on gamete variance provided another way of using recombination on short term selection, especially for bull sires and bull dams (Santos *et al*., 2019). The quadratic effect of maternal age on recombination rate suggests that young bull dams with higher recombination rate will have larger gametic variance and a better chance of producing eggs with extreme genetic merit. Finally, recombination rate does not need to be included as an independent trait in selective breeding because it has no direct economic values. But it will be under indirect, positive selection when breeding program is effective and proper selection indices used because of the long-term benefit of recombination in promoting genetic and gametic variations.

Conclusion

It has been shown that recombination rate can fluctuate in response to environmental changes. In this study, we used large pedigree data of dairy cattle to test the association between recombination rate and genetic and non-genetic factors, including maternal age and temperature. We discovered a non-linear association between maternal age and recombination rate in cattle, which has not been described before. Additionally, we found elevated recombination rates with increasing environmental temperature during conception. Taken together, our study provides clear evidence of an association between meiotic recombination with the non-genetic factors of maternal age and temperature. These results reveal useful insights into both the intrinsic and extrinsic effects on meiotic recombination.

Materials and Methods

Estimation of recombination rate in cattle pedigree

We used an approach similar to the one that was developed in previous studies (Ma *et al*., 2015; Shen *et al*., 2018). First, we identified recombination/crossover events in genotyped cattle pedigree data from the national dairy genomic database hosted at the Council on Dairy Cattle Breeding (CDCB). Based on the millions of animals with genotype and pedigree data available, we extracted 305,545 three-generation families where an offspring (first generation), at least one parent (second generation), and at least one grandparent (third generation) were genotyped. We then phased the two haplotypes of an animal (first and second generations) based on the parental genotypes. We identified crossover events by comparing haplotypes in the first and second generations. Recombination events were assigned to an interval flanked by two informative SNPs (phased heterozygote SNPs in the second generation), and we estimated recombination rate between consecutive SNPs by the average crossover numbers per meiosis. We only used three-generation families that were genotyped by at least 50 K SNP chips. We used the ARS-UCD 1.2 genome assembly (Rosen *et al*., 2020) with updated SNP coordinates1 and removed the loci from problematic regions identified in previous studies (Null *et al*., 2019). We only used autosome data due to the quality issues with the sex chromosomes. We also removed animals with more than 45 genome-wide recombination events based on the distribution of recombination across all animals, which is close to a normal distribution with mean 23.2 and variance 98.3.

Temperature data information from the NOAA database

The National Oceanic and Atmospheric Administration (NOAA) is an American scientific agency that focuses on the conditions of the oceans and atmosphere. It's also the largest database that contains the weather records of most United States cities since 1970s. By accessing the NOAA database using the R package "rnoaa" (Edmund *et al*., 2014), we extracted the weather conditions during two critical periods of a cow's development that may affect the female meiotic recombination process (Table 1). The first temperature was the average temperature in the month prior to the birthdate of the offspring that measures the fetal development environment of the offspring (the first generation in a three-generation family). The second temperature was the average temperature during the month prior to the birthdate of the parent that measures the environment during the fetal development of the parent (the second generation). We then combined the temperature data with our recombination records for our mixed model analysis. By considering both the range of temperature and data availability, we divided the original temperature data into three levels: temperatures above 26.67ºC are considered to be "hot," temperatures below 4.44ºC are considered as "cold," and temperatures in between are "normal.".

A full model analysis of genetics, maternal age, and temperature

From each of the 36,009 three-generation families, we estimated the total number of crossover events per meiosis of the female parent (second generation). We then adjusted the number of crossover events by SNP density and the number of informative markers (phased heterozygote SNPs) of each animal. We first checked the maternal age effect

using a smooth spline and boxplot. The smooth spline was fitted in R using the smooth spline function between recombination rate residuals and maternal age. Using the recombination rate residuals as phenotypes, we also fitted a linear mixed model to test for the effect of all available factors on recombination rate using the MMAP software (O'Connell, 2013). The model equation was fitted as following,

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{T_1} + \boldsymbol{T_2} + \mathbf{A} + \boldsymbol{A^2} + \mathbf{B} + \boldsymbol{B^2} + \mathbf{g} + \boldsymbol{\varepsilon}\,, \qquad \textbf{(1)}$$

where **Y** refers to the recombination rate residuals of individuals, $\mathbf{T_1}$ and $\mathbf{T_2}$ are the fixed effects for low and high temperatures, **A** represents a fixed effect of maternal age, $\mathbf{A^2}$ represents the squared effect of maternal age, **B** and $\mathbf{B^2}$ represents the fixed effect of parent' birth year and its square, and **g** is a random effect for the genetic or animal effect on recombination rate with $\mathbf{g} \sim N(0, \sigma^2\mathbf{G})$ and **G** being a genomic relationship matrix of the individuals calculated using the approach developed by VanRaden (2008). Both the temperatures during the fetal development of the parents and offspring were tested in this model. Statistical differences were declared as significant at $P < 0.05$..

Data availability statement

The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the animal study because no live animals are used.

References

1. Abdullah MF, Borts RH (2001): **Meiotic recombination frequencies are affected by nutritional states in Saccharomycescerevisiae**. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(25):14524-14529.

2. Arrieta, M., Willems, G., DePessemier, J., Colas, I., Burkholz, A., Darracq, A., *et al*. (2021). **The effect of heat stress on sugar beet recombination**. *Theor. Appl. Genet*. 134, 81–93. doi: 10.1007/s00122-020-03683-0.

3. Battagin, M., Gorjanc, G., Faux, A.-M., Johnston, S. E., and Hickey, J. M. (2016). **Effect of manipulating recombination rates on response to selection in livestock breeding programs**. *Genet. Sel. Evol*. 48, 1–12.

4. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B (2010). **PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice**. *Science*, 327(5967):836-840.

5. Bohmanova, J., Misztal, I., and Cole, J. (2007). **Temperature-humidity indices as indicators of milk production losses due to heat stress**. *J. Dairy Sci*. 90, 1947–1956. doi: 10.3168/jds.2006-513.

6. Cavestany D, el-Wishy AB, Foote RH (1985). **Effect of season and high environmental temperature on fertility of Holstein cattle**. *Journal of dairy science*, **68**(6):1471-1478.

7. Chowdhury R, Bois PR, Feingold E, Sherman SL, Cheung VG (2009). **Genetic analysis of variation in human meiotic recombination**. *PLoS genetics*, **5**(9):e1000648.

8. Church, K., and Wimber, D. E. (1969). **Meiosis in the grasshopper: chiasma frequency after elevated temperature and x-rays**. *Can. J. Genet. Cytol*. 11, 209–216. doi: 10.1139/g69-025.

9. Edmund, H., Chamberlain, S., Ram, K., and Edmund, M. H. (2014). rnoaa: **'NOAA' Weather Data from R**. *R Package Version 1.3.2*. Available online at: https://CRAN.R-project.org/package=rnoaa.

10. Francis KE, Lam SY, Harrison BD, Bey AL, Berchowitz LE, Copenhaver GP (2007). **Pollen tetrad-based visual assay for meiotic recombination in *Arabidopsis***. *Proceedings of the National Academy of Sciences of the United States of America*, 104(10):3913-3918.

11. Griffin DK, Abruzzo MA, Millie EA, Sheean LA, Feingold E, Sherman SL, Hassold TJ (1995). **Non-disjunction in human sperm: evidence for an effect of increasing paternal age**. *Human molecular genetics*, **4**(12):2227-2232.

12. Hassold T, Hunt P (2001). **To err (meiotically) is human: the genesis of human aneuploidy**. *Nature reviews Genetics*, **2**(4):280-291.

13. Hassold, T., Merrill, M., Adkins, K., Freeman, S., and Sherman, S. (1995). **Recombination and maternal age-dependent nondisjunction: molecular studies of trisomy** 16. *Am. J. Hum. Genet*. 57, 867–874.

14. Henderson, S., and Edwards, R. (1968). **Chiasma frequency and maternal age in mammals**. *Nature* 218, 22–28. doi: 10.1038/218022a0.

15. Hunter CM, Robinson MC, Aylor DL, Singh ND (2016). **Genetic Background, Maternal Age, and Interaction Effects Mediate Rates of Crossing Over in Drosophila melanogaster Females**. *G3*, **6**(5):1409-1416.

16. Hussin J, Roy-Gagnon MH, Gendron R, Andelfinger G, Awadalla P (2011). **Age-dependent recombination rates in human pedigrees**. *PLoS genetics*, **7**(9):e1002251.

17. Isberg, S. R., Johnston, S. M., Chen, Y., and Moran, C. (2006). **First evidence of higher female recombination in a species with temperature-dependent sex determination: the saltwater crocodile**. *J. Hered*. 97, 599–602. doi: 10.1093/ jhered/esl035.

18. Jackson S, Nielsen DM, Singh ND (2015). **Increased exposure to acute thermal stress is associated with a non-linear increase in recombination frequency and an independent linear decrease in fitness in *Drosophila***. *Bmc Evol Biol*, **15**.

19. Johnston SE, Berenos C, Slate J, Pemberton JM (2016). **Conserved Genetic Architecture Underlying Individual Recombination Rate Variation in a Wild Population of Soay Sheep (*Ovis aries*)**. *Genetics*, **203**(1):583-598.

20. Kong A, Barnard J, Gudbjartsson DF, Thorleifsson G, Jonsdottir G, Sigurdardottir S, Richardsson B, Jonsdottir J, Thorgeirsson T, Frigge ML *et al*. (2004). **Recombination rate and reproductive success in humans**. *Nature genetics* 2004, **36**(11):1203-1206.

21. Kong A, Thorleifsson G, Stefansson H, Masson G, Helgason A, Gudbjartsson DF, Jonsdottir GM, Gudjonsson SA, Sverrisson S, Thorlacius T *et al*. (2008). **Sequence variants in the RNF212 gene associate with genome-wide recombination rate**. *Science*, **319**(5868):1398-1401.

22. Kong, A., Thorleifsson, G., Stefansson, H., Masson, G., Helgason, A., Gudbjartsson, D. F., *et al*. (2008). **Sequence variants in the RNF212 gene associate with genome-wide recombination rate**. *Science* 319, 1398–1401. doi: 10.1126/science.1152422.

23. Levine RP (1955). **Chromosome Structure and the Mechanism of Crossing Over**. *Proceedings of the National Academy of Sciences of the United States of America*, **41**(10):727-730.

24. Lim, J. G., Stine, R. R., and Yanowitz, J. L. (2008). **Domain-specific regulation of recombination in *Caenorhabditis elegans* in response to temperature, age and sex**. Genetics 180, 715–726. doi: 10.1534/genetics.108.090142.

25. Lipkin SM, Moens PB, Wang V, Lenzi M, Shanmugarajah D, Gilgeous A, Thomas J, Cheng J, Touchman JW, Green ED *et al*. (2002). **Meiotic arrest and aneuploidy in MLH3-deficient mice**. *Nature genetics*, **31**(4):385-390.

26. Loidl J (1989). **Effects of Elevated-Temperature on Meiotic Chromosome Synapsis in Allium-Ursinum**. *Chromosoma*, **97**(6):449-458.

27. Lozada-Soto, E. A., Maltecca, C., Wackel, H., Flowers, W., Gray, K., He, Y., *et al*. (2021). **Evidence for recombination variability in purebred swine populations**. *J. Anim. Breed. Genet.* 138, 259–273. doi: 10.1111/jbg.12510.

28. Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, Bickhart DM, Cole JB, Null DJ, Liu GE *et al*. (2015). **Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis**. *PLoS genetics*, **11**(11):e1005387.

29. Martin HC, Christ R, Hussin JG, O'Connell J, Gordon S, Mbarek H, Hottenga JJ, McAloney K, Willemsen G, Gasparini P, *et al*. (2015). **Multicohort analysis of the maternal age effect on recombination**. *Nature communications*, **6**:7846.

30. Modliszewski JL, Copenhaver GP (2017). **Meiotic recombination gets stressed out: CO frequency is plastic under pressure**. *Current opinion in plant biology*, **36**:95-102.

31. Null, D., VanRaden, P. M., Rosen, B., O'Connell, J., and Bickhart, D. (2019). **Using the ARS-UCD1. 2 reference genome in US evaluations**. *Interbull Bull*. 55, 30–34.

32. O'Connell, J. R. (2013). "**MMAP: a comprehensive mixed model program for analysis of pedigree and population data**," in *Proceedings of the 63th Annual Meeting of The American Society of Human Genetics* (Boston, MA).

33. Otto, S. P., and Barton, N. H. (2001). **Selection for recombination in small**

populations. *Evolution* 55, 1921–1931. doi: 10.1554/0014-3820(2001)055[1921:sfrisp]2.0.co;2.

34. Phillips, D., Jenkins, G., Macaulay, M., Nibau, C., Wnetrzak, J., Fallding, D., *et al*. (2015). **The effect of temperature on the male and female recombination landscape of barley**. *New Phytol*. 208, 421–429. doi: 10.1111/nph.13548.

35. Polani PE, Jagiello GM (1976). **Chiasmata, meiotic univalents, and age in relation to aneuploid imbalance in mice**. *Cytogenetics and cell genetics*, **16**(6):505-529.

36. Rose AM, Baillie DL (1979). **The Effect of Temperature and Parental Age on Recombination and Nondisjunction in CAENORHABDITIS ELEGANS**. *Genetics*, **92**(2):409-418.

37. Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., *et al*. (2020). **De novo assembly of the cattle reference genome with single-molecule sequencing**. *Gigascience* 9:giaa021.

38. Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M (2012). **Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle**. *PLoS genetics*, **8**(7):e1002854.

39. Santos, D., Cole, J., Lawlor, T. Jr., VanRaden, P., Tonhati, H., and Ma, L. (2019). **Variance of gametic diversity and its application in selection programs**. *J. Dairy Sci*. 102, 5279–5294. doi: 10.3168/jds.2018-15971.

40. Shen, B., Jiang, J., Seroussi, E., Liu, G. E., and Ma, L. (2018). **Characterization of recombination features and the genetic basis in multiple cattle breeds**. *BMC Genomics* 19:304. doi: 10.1186/s12864-018-4705-y.

41. Stern C (1926). **An Effect of Temperature and Age on Crossing-Over in the First Chromosome of Drosophila Melanogaster**. *Proceedings of the National Academy of Sciences of the United States of America*, **12**(8):530-532.

42. VanRaden, P. M. (2008). **Efficient methods to compute genomic predictions**. *J. Dairy Sci*. 91, 4414–4423. doi: 10.3168/jds.2007-0980.

43. Wang Z, Shen B, Jiang J, Li J, Ma L (2016). **Effect of sex, age and genetics on crossover interference in cattle**. *Scientific reports*, **6**:37698.

44. Zhang, J., Kadri, N. K., Mullaart, E., Spelman, R., Fritz, S., Boichard, D., *et al*. (2020). **Genetic architecture of individual variation in recombination rate on the X chromosome in cattle**. *Heredity* 125, 304–316. doi: 10.1038/s41437-020-0341-9

Table

Table 1. Results of the mixed model analyses of eight factors related to the recombination rates in cattle.

Factor	Beta	*P*-value
Cold temperature during fetal development of offspring	−0.194	0.019
Hot temperature during fetal development of offspring	0.167	0.027
Cold temperature during fetal development of parent	0.139	0.097
Hot temperature during fetal development of parent	0.093	0.271
Maternal Age	−0.082	4.83×10^{-12}
Maternal Age^2	4.69×10^{-4}	5.68×10^{-4}
Parent Birth Year	5.02×10^{-3}	1.51×10^{-31}
Parent Birth Year^2	-4.41×10^{-7}	1.01×10^{-40}

Figures

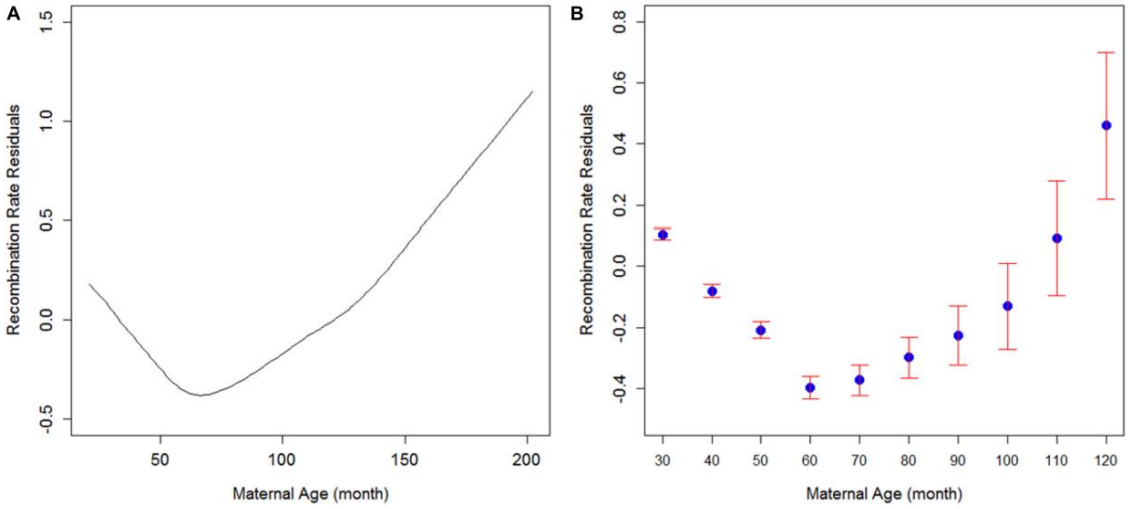


Figure 1. Trend of recombination rate residuals against maternal age in cattle.

(A) Fitted smooth spline of recombination rate residuals along with maternal age. The smooth spline was fitted in R using the smooth spline function between recombination rate residuals and maternal age.

(B) Recombination rate residuals in different maternal age groups. Blue dots are means, and bars are standard errors.

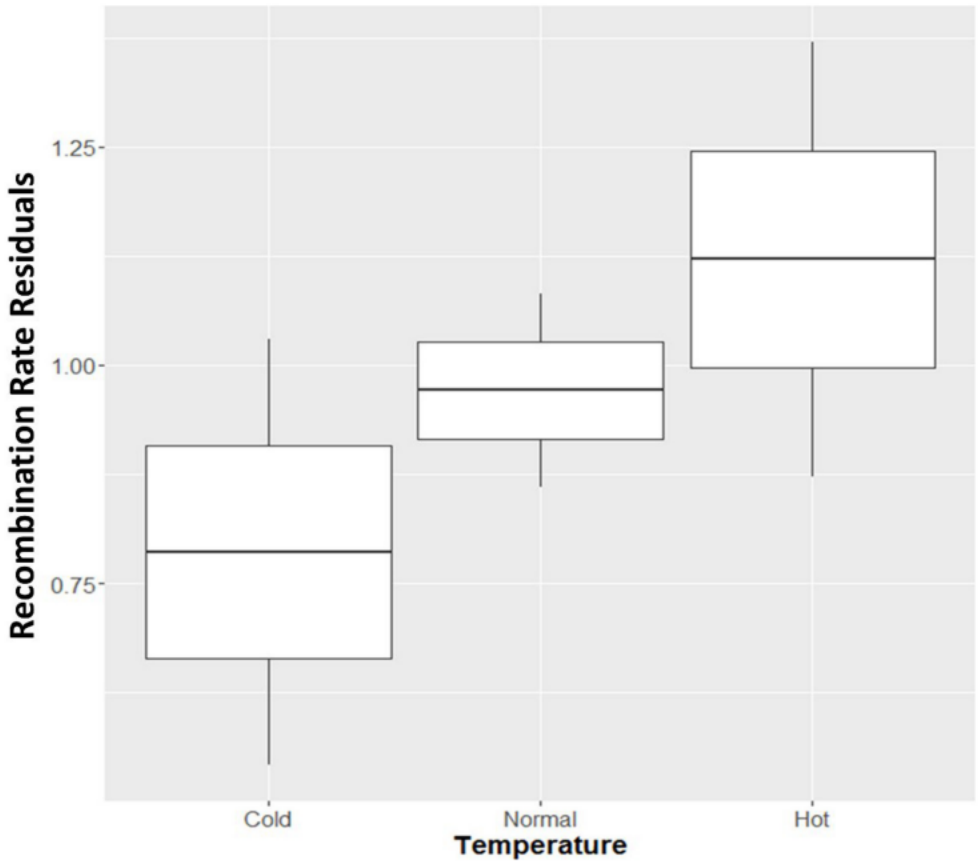


Figure 2. Boxplot of recombination rate residuals in three temperature conditions during the fetal development of offspring. **Cold**: temperatures below 4.44ºC; **Normal**: temperatures between 4.44 and 26.67ºC; **Hot**: temperatures above 26.67ºC

Chapter 5: Conclusions

The objective of this research was to understand the genetic architecture of complex traits and apply the understanding to investigate the biological relationship between genetics and disease in dairy cattle. The studies in Chapters 2-4 were all focused on addressing this objective.

In **Chapter 2**, the aim was to investigate the genetic basis of health and related traits in dairy cattle. We ran a GWAS and fine-mapping analysis on livability and six health traits in Holstein cattle and reported significant associations and candidate genes relevant to cattle health. Additionally, we combined our results with transcriptome data across multiple tissues in cattle, which will facilitate future functional studies on cattle health. Overall, our study provides insight on the biological relationship between genetics and disease susceptibility in cattle.

In **Chapter 3**, our aim was to evaluate genome-wide and region-specific changes in a U.S. dairy cattle population over a period of time. We identified candidate variants under selection, which are associated with biological traits and economically important traits in cattle. In addition, I proposed a gene dropping simulation program in R software to identify the genome changes that occurred due to selection from those due to random genetic drift. I demonstrated that gene dropping is an applicable method to investigate changes in the cattle genome over time. This method and software program will be useful to visualize genes that are significant and study them for their effect on genomic selection in dairy cattle.

In **Chapter 4**, the aim was to study meiotic recombination and demonstrate the effect of maternal age and temperature on recombination rate in cattle. From this

study, we provided novel information regarding the plasticity of meiotic recombination in cattle. We also showed a positive correlation between environmental temperature at conception and recombination rate in Holstein-Friesian cows. Collectively, our results indicate clear evidence of an association between meiotic recombination with maternal age and temperature. This study may facilitate follow-up work on the effects of other factors on meiotic recombination in cattle and other mammalian species.

In summary, the studies in these three chapters specifically focuses on the genetic architecture of complex traits. Future perspectives include conducting functional studies to test the candidate genes associated with complex diseases in cattle that we identified in Chapter 2. Additionally, the fine-mapped regions from our study in Chapter 2 can be analyzed by integrating other functional annotation data to identify biologically meaningful information about those regions and complex traits. From Chapter 3, the development of a gene dropping simulation program in Python will be able to obtain expected absolute allele frequency changes between any two generations. This program can be compared for consistency to our current gene dropping simulation results in U.S. dairy cattle. Furthermore, the results from Chapter 4 can be extended to investigate the association between the genetic architecture of meiotic recombination with other non-genetic factors to determine its implications for genetic studies of complex traits.

Bibliography

Abdullah MF, Borts RH: Meiotic recombination frequencies are affected by nutritional states in Saccharomycescerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(25):14524-14529.

Arrieta, M., Willems, G., DePessemier, J., Colas, I., Burkholz, A., Darracq, A., *et al*. (2021). The effect of heat stress on sugar beet recombination. *Theor. Appl. Genet*. **134**, 81–93. doi: 10.1007/s00122-020-0 3683-0.

Auer, P.L., Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* **7**, 16 (2015).

Backman JD, O'Connell JR, Tanner K, Peer CJ, Figg WD, Spencer SD, Mitchell BD, Shuldiner AR, Yerges-Armstrong LM, Horenstein RB. Genome-wide analysis of clopidogrel active metabolite levels identifies novel variants that influence antiplatelet response. *Pharmacogenet Genomics*. 2017; **27**(4):159.

Barbaux S, Gascoin-Lachambre G, Buffat C, Monnier P, Mondon F, Tonanny M-B, Pinard A, Auer J, Bessières B, Barlier A. A genome-wide approach reveals novel imprinted genes expressed in the human placenta. *Epigenetics*. 2012; **7**(9):1079–90.

Bartlett PC, Kirk JH, Wilke MA, Kaneene JB, Mather EC: Metritis complex in Michigan Holstein-Friesian cattle: incidence, descriptive epidemiology and estimated economic impact. *Preventive veterinary medicine* 1986, **4**(3):235-248.

Battagin, M., Gorjanc, G., Faux, A.-M., Johnston, S. E., and Hickey, J. M. (2016). Effect of manipulating recombination rates on response to selection in livestock breeding programs. *Genet. Sel. Evol*. **48**, 1–12.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B: PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science* 2010, **327**(5967):836-840.

Bohmanova, J., Misztal, I., and Cole, J. (2007). Temperature-humidity indices as indicators of milk production losses due to heat stress. *J. Dairy Sci*. **90**, 1947–1956. doi: 10.3168/jds.2006-513.

Boyle, A. P., *et al*. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome research*, **22**(9), 1790–1797.

Buzanskas ME, do Amaral Grossi D, Ventura RV, Schenkel FS, TCS C, Stafuzza NB, Rola LD, SLC M, Mokry FB, de Alvarenga Mudadu M. Candidate genes for

male and female reproductive traits in Canchim beef cattle. *J Anim Sci Biotechnol*. 2017; **8**(1):67.

Canela-Xandri, O., Rawlik, K., Woolliams, J. A., & Tenesa, A. (2016). Improved Genetic Profiling of Anthropometric Traits Using a Big Data Approach. *PloS one*, **11**(12), e0166755.

Capon, Francesca *et al*. (2004). "A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups." *Human molecular genetics* vol. **13**, 20. 2361-8.

Cavestany D, el-Wishy AB, Foote RH: Effect of season and high environmental temperature on fertility of Holstein cattle. *Journal of dairy science* 1985, **68**(6):1471-1478.

Chen, W., *et al*. (2015). Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics*, **200**(3), 719–736.

Chowdhury R, Bois PR, Feingold E, Sherman SL, Cheung VG: Genetic analysis of variation in human meiotic recombination. *PLoS genetics* 2009, **5**(9):e1000648.

Church K, Wimber DE: Meiosis in the grasshopper: chiasma frequency after elevated temperature and x-rays. *Canadian journal of genetics and cytology Journal canadien de genetique et de cytologie* 1969, **11**(1):209-216.

Cole J, VanRaden P, O'Connell J, Van Tassell C, Sonstegard T, Schnabel R, Taylor J, Wiggans G. Distribution and location of genetic effects for dairy traits. J Dairy Sci. 2009; **92**(6):2931–46.

Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Crooker BA, Van Tassell CP, Yang J, Wang S, Matukumalli LK: Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary US Holstein cows. *BMC genomics* 2011, **12**(1):408.

Collins F, Brooks L, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Research*, (1998); **8**(12): 1229-31.

de Bakker, P. I., Ferreira, M. A., Jia, X., Neale, B. M., Raychaudhuri, S., & Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Human molecular genetics*, **17**(R2), R122–R128.

Doekes, H.P., Veerkamp, R.F., Bijma, P., Hiemstra, S.J., and J.J. Windig. "Trends in the genome-wide and region specific genetic diversity in the Dutch-Flemish Holstein-Friesian breeding program from 1986 to 2015." *Genetics Selection Evolution* (2018) **50**:15. *BMC*. Web.

Duffield T: Subclinical ketosis in lactating dairy cattle. *Veterinary clinics of north america: Food animal practice* 2000, **16**(2):231-253.

Edmund, H., Chamberlain, S., Ram, K., and Edmund, M. H. (2014). rnoaa: 'NOAA' Weather Data from R. *R Package Version 1.3.2*. Available online at: https://CRAN.R-project.org/package=rnoaa.

Eichler EE, *et al*. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**:446–450.

Fang L, Jiang J, Li B, Zhou Y, Freebern E, VanRaden PM, Cole JB, Liu GE, Ma L. Genetic and epigenetic architecture of paternal origin contribute to gestation length in cattle. *Commun Biol*. 2019; **2**(1):100.

Fang, L., Sahana, G., Ma, P. *et al*. Exploring the genetic architecture and improving genomic prediction accuracy for mastitis and milk production traits in dairy cattle by mapping variants to hepatic transcriptomic regions responsive to intra-mammary infection. *Genet Sel Evol* **49**, 44 (2017).

Faye, L. L., Machiela, M. J., Kraft, P., Bull, S. B., & Sun, L. Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. *PLoS genetics* **9**, 8 (2013).

Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, Gazal S, Loh P-R, Lareau C, Shoresh N. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet*. 2018; **50**(4):621.

Francis KE, Lam SY, Harrison BD, Bey AL, Berchowitz LE, Copenhaver GP: Pollen tetrad-based visual assay for meiotic recombination in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(10):3913-3918.

Frischknecht, M., Pausch, H., Bapst, B. *et al*. Highly accurate sequence imputation enables precise QTL mapping in Brown Swiss cattle. *BMC Genomics* **18**, 999 (2017).

Gaddis KP, Megonigal J Jr, Clay J, Wolfe C. Genome-wide association study for ketosis in US jerseys using producer-recorded data. *J Dairy Sci*. 2018; **101**(1):413–24.

Garrick DJ, Taylor JF, Fernando RL. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genet Sel Evol*. 2009; **41**(1):55.

Gaughan, J., Mader, T. L., Holt, S., and Lisle, A. (2008). A new heat load index for feedlot cattle. *J. Anim. Sci*. **86**, 226–234. doi: 10.2527/jas.2007-0305.

Gowane G, Vandre R, Nangre M, Sharma A. Major histocompatibility complex (MHC) of bovines: an insight into infectious disease resistance. Livestock Res Int. 2013; **1**(2):46–57.

Griffin DK, Abruzzo MA, Millie EA, Sheean LA, Feingold E, Sherman SL, Hassold TJ: Non-disjunction in human sperm: evidence for an effect of increasing paternal age. *Human molecular genetics* 1995, **4**(12):2227-2232.

Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, *et al*. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Res* **12**: 222–231.

Gurdasani, D., Barroso, I., Zeggini, E. *et al*. Genomics of disease risk in globally diverse populations. *Nat Rev Genet* **20**, 520–535 (2019).

Haidich, A.B. (2010). Meta-analysis in medical research. *Hippokratia*, **14**(Suppl 1), 29–37.

Hassold T, Hunt P: To err (meiotically) is human: the genesis of human aneuploidy. *Nature reviews Genetics* 2001, **2**(4):280-291.

Hassold, T., Merrill, M., Adkins, K., Freeman, S., and Sherman, S. (1995). Recombination and maternal age-dependent nondisjunction: molecular studies of trisomy 16. *Am. J. Hum. Genet*. **57**, 867–874.

Hayes, B.J. & Daetwyler, H. D. 1000 Bull Genomes project to map simple and complex genetic traits in cattle: applications and outcomes. *Annu. Rev. Anim. Biosci.* **7**, 1 (2018).

Hayes, B. J., Pryce, J., Chamberlain, A. J., *et al*. (2010). Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. *PLoS genetics*, **6**(9), e1001139.

Heidaritabar, M., Vereijken, A., Muir, W.M., Meuwissen, T., Cheng, H., Megens, H.J., Groenen, M.A.M., and J.W.M. Bastiaansen. "Systematic differences in the response of genetic variation to pedigree and genome-based selection methods." *Heredity* (2014) **113**: 503-513.

Henderson, S., and Edwards, R. (1968). Chiasma frequency and maternal age in mammals. *Nature* **218**, 22–28. doi: 10.1038/218022a0.

Heringstad B, Gianola D, Chang YM, Odegård J, Klemetsdal G. Genetic associations between clinical mastitis and somatic cell score in early first-lactation cows. *J Dairy Sci.* 2006 Jun; **89**(6):2236-44.

Hindorff L, Sethupathy P, Junkins H, Ramos E, Mehta J, Collins F, *et al*. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of National Academy of Sciences of the United States of America*. (2009); **106**(23): 9362-7.

Hiroki T, Liebhaber SA, Cooke NE. An intronic locus control region plays an essential role in the establishment of an autonomous hepatic chromatin domain for the human vitamin D-binding protein gene. *Mol Cell Biol*. 2007; **27**(21):7365–80.

Höglund, J., Rafati, N., Rask-Andersen, M. *et al*. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Sci Rep* **9**, 16844 (2019).

Horst R, Goff J, Reinhardt T. Role of vitamin D in calcium homeostasis and its use in prevention of bovine periparturient paresis. *Acta Vet Scand Suppl*. 2003; **97**:35–50.

Hu Z-L, Park CA, Wu X-L, Reecy JM. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the postgenome era. *Nucleic Acids Res*. 2012; **41**(D1):D871–9.

Hunter CM, Robinson MC, Aylor DL, Singh ND: Genetic Background, Maternal Age, and Interaction Effects Mediate Rates of Crossing Over in Drosophila melanogaster Females. *G3* 2016, **6**(5):1409-1416.

Hussin J, Roy-Gagnon MH, Gendron R, Andelfinger G, Awadalla P: Age-dependent recombination rates in human pedigrees. *PLoS genetics* 2011, **7**(9):e1002251.

Ioannidis, N. M., *et al*. (2017). FIRE: functional inference of genetic variants that regulate gene expression. *Bioinformatics*, *33*(24), 3895–3901.

Isberg, S. R., Johnston, S. M., Chen, Y., and Moran, C. (2006). First evidence of higher female recombination in a species with temperature-dependent sex determination: the saltwater crocodile. *J. Hered*. **97**, 599–602. doi: 10.1093/ jhered/esl035.

Jackson S, Nielsen DM, Singh ND: Increased exposure to acute thermal stress is associated with a non-linear increase in recombination frequency and an independent linear decrease in fitness in Drosophila. *BMC Evol Biol* 2015, **15**.

Jiang, J., Cole, J.B., Freebern, E. *et al*. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Commun Biol* **2**, 212 (2019).

Johnston SE, Berenos C, Slate J, Pemberton JM: Conserved Genetic Architecture Underlying Individual Recombination Rate Variation in a Wild Population of Soay Sheep (Ovis aries). *Genetics* 2016, **203**(1):583-598.

Jostins, L., Ripke, S., Weersma, R. *et al*. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).

Kaiser, Jocelyn. "'Landmark' Study Resolves a Major Mystery of How Genes Govern Human Height." *Science*, 3 Nov. 2020.

Khatkar MS, Nicholas FW, Collins AR, Zenger KR, Cavanagh JA, Barris W, Schnabel RD, Taylor JF, Raadsma HW. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics. 2008; **9**(1):187

Kircher, M., Witten, D., Jain, P. *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).

Kong A, Barnard J, Gudbjartsson DF, Thorleifsson G, Jonsdottir G, Sigurdardottir S, Richardsson B, Jonsdottir J, Thorgeirsson T, Frigge ML *et al*: Recombination rate and reproductive success in humans. *Nature genetics* 2004, **36**(11):1203-1206.

Kong A, Thorleifsson G, Stefansson H, Masson G, Helgason A, Gudbjartsson DF, Jonsdottir GM, Gudjonsson SA, Sverrisson S, Thorlacius T *et al*: Sequence variants in the RNF212 gene associate with genome-wide recombination rate. *Science* 2008, **319**(5868):1398-1401.

Lango Allen, H., Estrada, K., Lettre, G. *et al*. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).

Laven R, Peters A: Bovine retained placenta: aetiology, pathogenesis and economic loss. *Veterinary Record* 1996, **139**(19):465-471.

Levine RP: Chromosome Structure and the Mechanism of Crossing Over. *Proceedings of the National Academy of Sciences of the United States of America* 1955, **41**(10):727-730.

Lewinger, J.P., Conti, D. V., Baurley, J. W., Triche, T. J., & Thomas, D. C. (2007). Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genetic epidemiology*, **31**(8), 871–882.

Liang D, Arnold L, Stowe C, Harmon R, Bewley J: Estimating US dairy clinical disease costs with a stochastic simulation model. *Journal of dairy science* 2017, **100**(2):1472-1486.

Lim, J. G., Stine, R. R., and Yanowitz, J. L. (2008). Domain-specific regulation of recombination in Caenorhabditis elegans in response to temperature, age and sex. *Genetics* **180**, 715–726. doi: 10.1534/genetics.108.090142.

Linders PT, van der Horst C, ter Beest M, van den Bogaart G. Stx5-Mediated ER-Golgi Transport in Mammals and Yeast. *Cells*. 2019; **8**(8):780.

Lipkin SM, Moens PB, Wang V, Lenzi M, Shanmugarajah D, Gilgeous A, Thomas J, Cheng J, Touchman JW, Green ED *et al*: Meiotic arrest and aneuploidy in MLH3-deficient mice. *Nature genetics* 2002, **31**(4):385-390.

Liu, S., Yu, Y., Zhang, S. *et al*. Epigenomics and genotype-phenotype association analyses reveal conserved genetic architecture of complex traits in cattle and human. *BMC Biol* **18**, 80 (2020).

Loidl J: Effects of Elevated-Temperature on Meiotic Chromosome Synapsis in Allium-Ursinum. *Chromosoma* 1989, **97**(6):449-458.

Lozada-Soto, E. A., Maltecca, C., Wackel, H., Flowers, W., Gray, K., He, Y., *et al*. (2021). Evidence for recombination variability in purebred swine populations. *J. Anim. Breed. Genet.* **138**, 259–273. doi: 10.1111/jbg.12510.

Lund MS, Sahana G, Andersson-Eklund L, *et al*. Joint analysis of quantitative trait loci for clinical mastitis and somatic cell score on five chromosomes in three Nordic dairy cattle breeds. *J Dairy Sci.* 2007 Nov; **90**(11):5282-90.

Ma, L., Cole, J. B., Da, Y., & VanRaden, P. M. (2019). Symposium review: Genetics, genome-wide association study, and genetic improvement of dairy fertility traits. *Journal of dairy science*, *102*(4), 3735–3743.

Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, Bickhart DM, Cole JB, Null DJ, Liu GE *et al*: Cattle Sex-Specific Recombination and Genetic Control from a Large Pedigree Analysis. *PLoS genetics* 2015, **11**(11):e1005387.

Makowsky R., Pajewski N. M., Klimentidis Y. C., Vazquez A. I., Duarte C. W., *et al*., 2011. Beyond missing heritability: Prediction of complex traits. *PLoS Genet*. **7**: e1002051.

Manolio, T.A. *et al*. 2009 Finding the missing heritability of complex diseases. *Nature* **461**, 747–753.

Marchini, J., Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499–511 (2010).

Marete AG, Guldbrandtsen B, Lund MS, Fritz S, Sahana G, Boichard D. A metaanalysis including pre-selected sequence variants associated with seven traits in three French dairy cattle populations. *Front Genet*. 2018; **9**:522.

Martin HC, Christ R, Hussin JG, O'Connell J, Gordon S, Mbarek H, Hottenga JJ, McAloney K, Willemsen G, Gasparini P *et al*: Multicohort analysis of the maternal age effect on recombination. *Nature communications* 2015, **6**:7846.

Maurano MT, Humbert R, Rynes E, *et al*. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012; **337**(6099):1190-1195.

McKay SD, Schnabel RD, Murdoch BM, Matukumalli LK, Aerts J, Coppieters W, Crews D, Neto ED, Gill CA, Gao C. Whole genome linkage disequilibrium maps in cattle. *BMC Genet*. 2007; **8**(1):74.

Modliszewski JL, Copenhaver GP: Meiotic recombination gets stressed out: CO frequency is plastic under pressure. *Current opinion in plant biology* 2017, **36**:95-102.

Nalaila S, Stothard P, Moore S, Li C, Wang Z. Whole-genome QTL scan for ultrasound and carcass merit traits in beef cattle using Bayesian shrinkage method. *J Anim Breed Genet*. 2012; **129**(2):107–19.

Nayeri S, Sargolzaei M, Abo-Ismail M, Miller S, Schenkel F, Moore S, Stothard P. Genome-wide association study for lactation persistency, female fertility, longevity, and lifetime profit index traits in Holstein dairy cattle. *J Dairy Sci*. 2017; **100**(2):1246–58.

Nguyen T. T., Huang, J.Z., Wu, Q., Nguyen, T.T., and M.J. Li. "Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests." *BMC Genomics* (2015) **16**: 21-23.

Null, D., VanRaden, P. M., Rosen, B., O'Connell, J., and Bickhart, D. (2019). Using the ARS-UCD1. 2 reference genome in US evaluations. *Interbull Bull.* **55**, 30–34.

O'Connell, J. R. (2013). "MMAP: a comprehensive mixed model program for analysis of pedigree and population data," in *Proceedings of the 63th Annual Meeting of The American Society of Human Genetics* (Boston, MA).

Olsen H, Hayes B, Kent M, Nome T, Svendsen M, Lien S. A genome wide association study for QTL affecting direct and maternal effects of stillbirth and dystocia in cattle. *Anim Genet*. 2010; **41**(3):273–80.

Olsen HG, Knutsen TM, Lewandowska-Sabat AM, Grove H, Nome T, Svendsen M, Arnyasi M, Sodeland M, Sundsaasen KK, Dahl SR. Fine mapping of a QTL

on bovine chromosome 6 using imputed full sequence data suggests a key role for the group-specific component (GC) gene in clinical mastitis and milk production. *Genet Sel Evol*. 2016; **48**(1):79.

Otto, S. P., and Barton, N. H. (2001). Selection for recombination in small populations. *Evolution* **55**, 1921–1931. doi: 10.1554/0014-3820(2001)055[1921:sfrisp]2.0.co;2

Parker Gaddis K, Tooker M, Wright J, Megonigal J, Clay J, Cole J, VanRaden P: Development of national genomic evaluations for health traits in U.S. Holsteins. *Proc 11th World Congr Genet Appl Livest Prod*, Auckland, New Zealand, Feb 11–16 2018, Vol. Biol. & Species–Bovine (dairy) **1**, p. 594.

Phillips, D., Jenkins, G., Macaulay, M., Nibau, C., Wnetrzak, J., Fallding, D., *et al*. (2015). The effect of temperature on the male and female recombination landscape of barley. *New Phytol.* **208**, 421–429. doi: 10.1111/nph.13548

Polani PE, Jagiello GM: Chiasmata, meiotic univalents, and age in relation to aneuploid imbalance in mice. *Cytogenetics and cell genetics* 1976, **16**(6):505-529.

Price Alkes L., Spencer Chris C. A. and Donnelly Peter. 2015. Progress and promise in understanding the genetic basis of common diseases. *Proc. R. Soc. B.* **282**: 20151684

Pryce JE, Hayes BJ, Bolormaa S, Goddard ME. Polymorphic regions affecting human height also control stature in cattle. *Genetics*. 2011; **187**(3):981–4.

Purfield DC, Bradley DG, Evans RD, Kearney FJ, Berry DP. Genome-wide association study for calving performance using high-density genotypes in dairy and beef cattle. *Genet Sel Evol*. 2015; **47**(1):47.

Raphaka, K., Matika, O., Sánchez-Molano, E. *et al*. (2017). Genomic regions underlying susceptibility to bovine tuberculosis in Holstein-Friesian cattle. *BMC Genet* **18**, 27.

Reinhardt TA, Lippolis JD, McCluskey BJ, Goff JP, Horst RL: Prevalence of subclinical hypocalcemia in dairy herds. *The Veterinary Journal* 2011, **188**(1):122-124.

Rose AM, Baillie DL: The Effect of Temperature and Parental Age on Recombination and Nondisjunction in CAENORHABDITIS ELEGANS. *Genetics* 1979, **92**(2):409-418.

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., *et al*. (2020). De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9**:giaa021.

Sahana G, Guldbrandtsen B, Thomsen B, Lund MS. Confirmation and finemapping of clinical mastitis and somatic cell score QTL in Nordic Holstein cattle. *Anim Genet*. 2013; **44**(6):620–6.

Sahana, G., *et al*. (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**(11), 7258–7275.

Sandor C, Li W, Coppieters W, Druet T, Charlier C, Georges M: Genetic variants in REC8, RNF212, and PRDM9 influence male recombination in cattle. *PLoS genetics* 2012, **8**(7):e1002854.

Santos, D., Cole, J., Lawlor, T. Jr., VanRaden, P., Tonhati, H., and Ma, L. (2019). Variance of gametic diversity and its application in selection programs. *J. Dairy Sci*. **102**, 5279–5294. doi: 10.3168/jds.2018-15971.

Santos D, Cole J, Null D, Byrem T, Ma L. (2018). Genetic and nongenetic profiling of milk pregnancy-associated glycoproteins in Holstein cattle. *J Dairy Sci*. **101**(11):9987–10000.

Sargolzaei M, Schenkel F, Jansen G, Schaeffer L. (2008). Extent of linkage disequilibrium in Holstein cattle in North America. *J Dairy Sci*. **91**(5):2106–17.

Sauna, Z., Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet* **12**, 683–691.

Schaid DJ, Chen W, Larson NB. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet*. **19**(8):491–504.

Schlafer D, Fisher P, Davies C. (2000). The bovine placenta before and after birth: placental development and function in health and disease. *Anim Reprod Sci* **60**:145–60.

Schrodi, S. J., Mukherjee, S., Shan, Y., *et al*. (2014). Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. *Frontiers in genetics*, **5**, 162.

Schwefel, D., B. S. Arasu, S. F. Marino, B. Lamprecht, K. Ko¨chert, E. Rosenbaum, J. Eichhorst, B. Wiesner, J. Behlke, O. Rocks, *et al*. 2013. Structural insights into the mechanism of GTPase activation in the GIMAP family. *Structure* **21**: 550–559.

Seegers H, Fourichon C, Beaudeau F (2003). Production effects related to mastitis and mastitis economics in dairy cattle herds. *Veterinary research*, **34**(5):475-491.

Shen, B., Jiang, J., Seroussi, E., Liu, G. E., and Ma, L. (2018). Characterization of recombination features and the genetic basis in multiple cattle breeds. *BMC Genomics* **19**:304. doi: 10.1186/s12864-018-4705-y.

Snelling W, Allan M, Keele J, Kuehn L, Mcdaneld T, Smith T, Sonstegard T, Thallman R, Bennett G. (2010). Genome-wide association study of growth in crossbred beef cattle. *J Anim Sci*. **88**(3):837–48.

Sonstegard, T. S. *et al*. (2013). Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. *PLoS ONE* **8**, e54872.

Spindel, J., Begum, H., Akdemir, D. *et al*. (2016). Genome-wide prediction models that incorporate *de novo* GWAS are a powerful new tool for tropical rice improvement. *Heredity* **116**, 395–408.

Stephens, M., Balding, D. (2009). Bayesian statistical methods for genetic association studies. *Nat Rev Genet* **10**, 681–690.

Stern C (1926). An Effect of Temperature and Age on Crossing-Over in the First Chromosome of Drosophila Melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* **12**(8):530-532.

Takeshima SN, Aida Y. (2006). Structure, function and disease susceptibility of the bovine major histocompatibility complex. *Anim Sci J*. **77**(2):138–50.

Tetens J, Seidenspinner T, Buttchereit N, Thaller G. (2013). Whole-genome association study for energy balance and fat/protein ratio in German Holstein bull dams. *Anim Genet*. **44**(1):1–8.

The 1000 Genomes Project Consortium., Corresponding Author., McVean, G. *et al*. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65.

The ENCODE Project Consortium., Overall coordination (data analysis coordination)., Dunham, I. *et al*. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74.

The Haplotype Reference Consortium., McCarthy, S., Das, S. *et al*. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–1283.

Toyofuku T, Yoshida J, Sugimoto T, Yamamoto M, Makino N, Takamatsu H, Takegahara N, Suto F, Hori M, Fujisawa H. (2008). Repulsive and attractive semaphorins cooperate to direct the navigation of cardiac neural crest cells. *Dev Biol*. **321**(1):251–62.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., & Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature genetics*, **45**(2), 124–130.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci*. **91**, 4414–4423. doi: 10.3168/jds.2007-0980

VanRaden PM, Sun C (2014). Fast Imputation Using Medium- or Low-Coverage Sequence Data. Proceedings, *10th World Congress of Genetics Applied to Livestock Production*.

VanRaden, P.M., Sun, C. & O'Connell, J.R. (2015). Fast imputation using medium or low-coverage sequence data. *BMC Genet* **16**, 82.

VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet Sel Evol*. **49**(1):32.

Wang Z, Shen B, Jiang J, Li J, Ma L (2016). Effect of sex, age and genetics on crossover interference in cattle. *Scientific reports* **6**:37698.

Wang, Y., Zhang, F., Mukiibi, R. *et al*. (2020). Genetic architecture of quantitative traits in beef cattle revealed by genome wide association studies of imputed whole genome sequence variants: II: carcass merit traits. *BMC Genomics* **21**, 38.

Witte J. S. (2010). Genome-wide association studies and beyond. *Annual review of public health* **31**, 9–20.

Wojcik, G.L., *et al*. (2018). Imputation-Aware Tag SNP Selection to Improve Power for Large-Scale, Multi-ethnic Association Studies. *G3: Genes, Genomes, Genetics*, **8**(10), 3255-3267.

Won H, de la Torre-Ubieta L, Stein JL, *et al*. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature*. **538**(7626):523-527.

Wright J, VanRaden P. (2016). Genetic evaluation of dairy cow livability. *J Anim Sci*. 94:178.

Wu X, Lund MS, Sahana G, Guldbrandtsen B, Sun D, Zhang Q, Su G. (2015). Association analysis for udder health based on SNP-panel and sequence data in Danish Holsteins. *Genet Sel Evol*. **47**(1):50.

Wu Y, Smas CM. (2008). Expression and regulation of transcript for the novel transmembrane protein Tmem182 in the adipocyte and muscle lineage. *BMC Res Notes*. **1**(1):85.

Xu, X., Yu, Y., Zong, K. *et al*. (2019). Up-regulation of IGF2BP2 by multiple mechanisms in pancreatic cancer promotes cancer proliferation by activating the PI3K/Akt signaling pathway. *J Exp Clin Cancer Res* **38**, 497.

Zadoks RN, Middleton JR, McDougall S, Katholm J, Schukken YH. (2011). Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. *J Mammary Gland Biol Neoplasia* **16**(4):357–72.

Zare, Y., Shook, G. E., Collins, M. T., & Kirkpatrick, B. W. (2014). Genome-wide association analysis and genomic prediction of Mycobacterium avium subspecies paratuberculosis infection in US Jersey cattle. *PloS one*, **9**(2), e88380.

Zhang, J., Kadri, N.K., Mullaart, E. *et al*. (2020). Genetic architecture of individual variation in recombination rate on the X chromosome in cattle. *Heredity* **125**, 304–316.

Zhu, Z., Zhang, F., Hu, H. *et al*. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481–487.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS *et al*. (2009). A whole-genome assembly of the domestic cow, Bos taurus. *Genome biology* **10**(4):R42.