

Identification of causal factors for recessive lethals in dairy cattle with special focus on large chromosomal deletions

Md Mesbah Uddin

▶ **To cite this version:**

Md Mesbah Uddin. Identification of causal factors for recessive lethals in dairy cattle with special focus on large chromosomal deletions. Animal genetics. Institut agronomique, vétérinaire et forestier de France; Aarhus universitet (Danemark), 2019. English. NNT : 2019IAVF0018 . tel-02447526

HAL Id: tel-02447526
https://pastel.hal.science/tel-02447526

Submitted on 21 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NNT : 2019 IAVF 0018

THESE DE DOCTORAT

préparée à l'Institut des sciences et industries du vivant et de l'environnement (AgroParisTech)

et

Center for Quantitative Genetics and Genomics (Aarhus University)

pour obtenir le grade de

Docteur de l'Institut agronomique, vétérinaire et forestier de France

Spécialité : Génétique animale

École doctorale n°581
Agriculture, alimentation, biologie, environnement et santé (ABIES)

par

Md Mesbah UDDIN

Identification of causal factors for recessive lethals in dairy cattle with special focus on large chromosomal deletions

Etude de délétions chromosomiques et de variants génétiques responsables de mortalité embryonnaire chez les bovins laitiers

Directeur de thèse : **Didier BOICHARD et Goutam SAHANA**
Co-encadrement de la thèse : **Bernt GULDBRANDTSEN, Mogens Sandø LUND et Aurélien CAPITAN**

Thèse présentée et soutenue à Foulum, (Danemark), le 17 Septembre 2019:

Composition du jury :

M. Just JENSEN, Professor, Aarhus University	Président
M. Didier BOICHARD, Senior Scientist, INRA	Directeur de thèse
M. Göran ANDERSSON, Professor, Swedish University of Agricultural Sciences (SLU)	Rapporteur
Mme Alessandra STELLA, Senior Researcher, National Research Council of Italy (CNR)	Rapporteur
M. Georg THALLER, Professor, Kiel University	Rapporteur
M. Claus Bøttcher JØRGENSEN, Professor, University of Copenhagen	Rapporteur

UMR 1313 Génétique Animale et Biologie Intégrative
AgroParisTech|INRA
78350 Jouy-en-Josas
France

Center for Quantitative Genetics and Genomics
Department of Molecular Biology and Genetics
Aarhus University, 8830 Tjele
Denmark

Identification of causal factors for recessive lethals in dairy cattle with special focus on large chromosomal deletions

Md Mesbah Uddin

PhD Thesis

This PhD thesis is submitted to the Graduate School of Science and Technology (GSST), Aarhus University, Denmark, and the Doctoral School ABIES (Agriculture Food Biology Environment Health), AgroParisTech, France, in fulfilment of requirements for the double PhD degrees under the Erasmus Mundus double degree program EGS-ABG.









Supervisors

Main supervisors

Goutam Sahana
Department of Molecular Biology and Genetics, Aarhus University, Denmark

Didier Boichard
Génétique Animale et Biologie Intégrative, AgroParisTech/INRA, France

Co-supervisors

Bernt Guldbrandtsen
Department of Molecular Biology and Genetics, Aarhus University, Denmark

Mogens Sandø Lund
Department of Molecular Biology and Genetics, Aarhus University, Denmark

Aurélien Capitan
Génétique Animale et Biologie Intégrative, AgroParisTech/INRA, France
Allice, France

Acknowledgement

Firstly, I am grateful to the EGS-ABG consortium for giving me the opportunity in this double doctorate program, hosted by two leading universities in the field—Aarhus University and AgroParisTech. Last four years was a life-changing and eye-opening journey for me with lots of academic, research, cultural and international experiences that I could not imagine without this Erasmus Mundus endeavor. I am proud and feel lucky to be an EGS-ABG graduate. I am thankful to all EGS-ABG colleagues, past and present, for their help throughout this journey.

I am grateful to both of my PhD host institutes, QGG/Aarhus University and INRA/AgroParisTech, for offering this PhD. I am thankful to MBG/AU and GABI/INRA secretaries for their help regarding the administration. I have to mention one name, Karin Smedegaard (may she rest in peace). Her help and assistance started from receiving me from the Viborg train station, setting-up the MBG email, registering for social security, opening bank account … to visiting apartments and renting my first apartment in Viborg. I am grateful to her, will deeply miss her presence.

I was lucky to have both Goutam and Didier as my PhD supervisors. I am thankful to them for their help and guidance throughout my PhD. It was an excellent learning experience for me. I really enjoyed and appreciated their prompt feedback on my work (both research results and writing) with detailed and to-the-point suggestions, and constructive criticisms when needed. I am indebted to Goutam and Didier for their untiring efforts to educate me, to correct my mistakes, and finally to shape this dissertation in a presentable manner (just within one week!).

Specially, I want to thank Goutam under whose direct supervision I started my PhD journey and spent major part of the learning phase of my PhD. From day one, Goutam made it clear that I have to take charge of my PhD studies. No doubt, I was overwhelmed with this freedom and independence, but with his sincere supervision, I was able to navigate through my PhD, and gained enough confidence to pursue research topics from conception to dissemination. Before starting PhD, I had little experience working in Linux environment or with big data. I am indebted to Goutam for bearing with me, especially in the 1st year of my PhD where he allowed me to attend several courses and spend time to learn programming/bioinformatics techniques (even though, I had no apparent progress in my research project). Without that, I could not imagine pursuing a PhD in quantitative genetics (let alone completing it!).

I am thankful to Didier for his supervision during the second half of my PhD. His mentorship was pleasant, delightful and enlightening. Despite his busy schedule, I always had access and undivided attention from him. Every meeting and discussion I had with him, whether research or life in general, ended with enlightenment and wisdom, which greatly helped my research and personal life. Lastly, my stay in GABI was stress-free, and I was always in vacation mood—could be the Paris effect!!!

I am also thankful to my co-supervisors Bernt, Mogens and Aurelien. I am grateful to Bernt from whom I learned the ins and outs of the Linux cluster in QGG, whole-genome sequence analysis, various aspects of genetics (to name few). Particularly, when I was struggling to interpret results in light of population/evolutionary genetics, I cherish the memories of those one-to-one discussions with Bernt (that often lasted several hours in late afternoon), which greatly helped me in finishing my first PhD paper. I am thankful to Mogens for bringing a broader, applied, and real-life problem solving perspective of the research to the supervision team, which was the North Star in my PhD that helped me not to fell prey of hair-splitting type analysis, and guided me to pursue topics that are relevant to livestock breeding and genetics. I am also grateful to Mogens for his keen interest in ensuring an all-

round training for all QGG PhD fellows to prepare us for a better career in academia/research and/or industry. Finally, I am thankful to Aurelien from whom I learned a lot, especially, while working with the recessive lethal project. I admired his enthusiasm, friendly attitude and hands-on supervision approach; I really enjoyed working with him. Besides his guidance on mapping recessive lethals, I learned a lot on storytelling, and precise scientific writing—though it was lot of extra work to remove few of those “may be/could be” words (he would often say we have the data to check that...).

I am thankful to Grum, my QGG Buddy, for his help in social and work life. I also thank my office mates Lingzhao and Bingjie in QGG, and Rabia and Margarita in GABI/INRA for their pleasant company and stimulating work environment. I am grateful to all QGG/Aarhus University and G2B/INRA colleagues for their support throughout this journey.

Lastly, I am grateful to my family, parents, my wife Abeda, and son Abdullah, for their unconditional love, support and sacrifice. Without their amazing supports, this PhD would still be a dream. I dedicate this PhD Thesis to them.

Md Mesbah Uddin

17 September 2019

Résumé in English

Fertility is an economically important trait in dairy cattle. Fertility is defined as the genetic ability of a cow to show oestrus and conceive after insemination, and to resume breeding after calving. This trait had a negative trend in recent years, partly due to intensive selection for production-related traits—which has a negative genetic correlation with fertility—using few bulls with high genetic merits for production traits. Before genomics, it was difficult to achieve faster genetic progress for fertility using pedigree-based breeding scheme due to low heritability of this trait. Genomic selection scheme provides a solution to this hard-to-select problem, especially for traits with low heritability. Identification of causal variants for recessive lethal mutations, when possible, and selection of a set of predictive markers that successfully track such causal variants, is vital for making mating-decision (e.g. to avoid at-risk mating) and for additional increase in genetic gains using genomic selection.

The overall aim of this PhD thesis is to identify causal variants for recessive lethal mutations and select a set of predictive markers that are in high linkage-disequilibrium with the causal variants for female fertility in dairy cattle. We addressed this broad aim under five articles/manuscripts that are presented through **Chapter 2** to **Chapter 6** in this thesis.

Chapter 2 describes a systematic approach of mapping recessive lethals in French Normande cattle using homozygous haplotype deficiency (HHD). This study shows the influence of sample size, quality of genotypes, quality of (genotype) phasing and imputation, age of haplotype (of interest), and last but not the least, multiple testing corrections, on discovery and replicability of HHD results. It also illustrates the importance of fine-mapping with pedigree and whole-genome sequence (WGS) data, (cross-species) integrative annotation to prioritize candidate mutation, and finally, large-scale genotyping of the candidate mutation, to validate or invalidate initial results.

Chapter 3 describes a high-resolution population-scale mapping of large chromosomal deletions from whole-genome sequences of 175 animals from three Nordic dairy breeds. This study employs three different approaches to validate identified deletions. Next, it describes population genetic properties and functional importance of these deletions. Finally, it illustrates deletion formation mechanisms based on the assembled sequence features at breakpoints. This study provides the basis for subsequent genetic studies performed during this PhD thesis.

Chapter 4 deals with three main issues related to imputation of structural variants, in this case, large chromosomal deletions, e.g. availability of deletion genotypes, size of haplotype reference panel, and finally, imputation itself. To address the first two issues, this study describes a Gaussian mixture model-based approach where read-depth data from the variant call format (VCF) file is used to genotype a known deletion locus, without the need for raw sequence (BAM) file. Finally, it presents a pipeline for joint imputation of WGS variants along with large chromosomal deletions.

Chapter 5 describes genome-wide association studies for female fertility in three Nordic dairy cattle breeds using imputed WGS variants including large chromosomal deletions. This study is based on the analyses of eight fertility related traits using single-marker association, conditional and joint analyses. This study illustrates that inflation in association test-statistics could be seen even after correcting for population stratification using (genomic) principal components, and relatedness among the samples using genomic relationship matrices; however, this was known for traits with strong polygenic effects, among other factors. Finally, mapping of several new quantitative trait loci (QTL), along with the previously known ones, are reported in this study. This study also highlights the importance of including (imputed) large deletions for association mapping of fertility traits.

Chapter 6 describes prediction of genomic breeding values for fertility using SNP array-chip genotypes, selected QTL and large chromosomal deletion. Using genomic best linear unbiased prediction (GBLUP) method with one or several genomic-relationship matrices derived from a set of selected markers, this study reports higher prediction accuracy compared with previous report. This study also highlights the influence of selecting markers with best predictability, especially for a breed with small training population, in accuracy of genomic prediction. The results demonstrate that large deletions in general have a high predictive performance.

Finally, **Chapter 7** provides a general discussion, conclusion, and perspective based on findings of this PhD study.

Résumé på Dansk

Frugtbarhed er en økonomisk vigtig egenskab hos malkekvæg. Frugtbarhed defineres som en koens evne til at viser brunsttegn, bliver drægtige efter insemination, og som kan genoptage avl efter kælvning. Denne egenskab har de seneste år været i tilbagegang på grund af kraftig selektion for produktionsrelaterede egenskaber, der har en negativ sammenhæng med frugtbarheden. Det er svært at selektere for (bedre) frugtbarhed ved hjælp af stamtavlebaserede avlsprogrammer, da frugtbarhed har en lav arvelighed. Genomisk information kan muligvis løse dette problem. Identifikation af kausale markører, hvis muligt, og udvælgelse af et sæt prædiktive markører, der med succes sporer de sande kausale alleler, er afgørende for at foretage parringsbeslutninger (f.eks. for at undgå parring mellem to bærere) og for at øge genetisk fremgang med genomisk selektion.

Det overordnede formål med denne ph.d.-afhandling er at identificere kausale alleler og at vælge et sæt prædiktive markører, der er i høj koblingsuligevægt med de kausale alleler for hunlig frugtbarhed hos malkekvæg. Vi forholder os til det overordnede formål gennem fem artikler / manuskripter, der præsenteres i kapitel 2 til kapitel 6 i denne afhandling.

Kapitel 2 beskriver en systematisk metode til kortlægning af recessive letale alleler hos fransk Normande kvæg ved hjælp af ”homozygous haplotype deficiency” (HHD). Dette studie viser hvordan prøvestørrelse, kvalitet af genotyper, kvalitet af (haplotype) fasning og imputation, alder af haplotype, og sidst men ikke mindst, ”multiple testing” korrektioner har indflydelse på opdagelse og replikation af HHD-resultater. Dette studie viser også vigtigheden af kortlægning med stamtavle og helgenom-sekvensering (WGS) data, integreret notation for at prioritere kandidatmutation på tværs af arter, og indsamling af store mængder genotypeinformation om kandidatmutationen for at validere foreløbige resultater.

Kapitel 3 beskriver en kortlægning af store kromosomale deletioner ved brug af helgenomsekvenser fra 175 dyr fra tre nordiske malkekvægsracer. Denne undersøgelse anvender tre forskellige tilgange til validering af kortlægningsresultater. Dernæst beskrives populationen genetiske egenskaber og de funktionelle betydninger af de identificerede deletioner. Til sidst forklares mekanismer, som kan danne deletioner, baseret på de samlede sekvensfunktioner hvor deletioner er opståede. Denne undersøgelse danner basis for de efterfølgende genetiske analyser, som er præsenteret i denne ph.d.-afhandling.

Kapitel 4 omhandler tre hovedproblemer relateret til imputering af strukturelle varianter, som i dette tilfælde er store kromosomale deletioner. Disse er tilgængelighed af reference deletionsgenotyper, størrelse af referencepanel for haplotyper og imputation metodologi. For at løse de to første problemer beskriver dette studie en gaussisk mixed-model tilgang, hvor læsningsdata fra VCF-filen (VCF) bruges til at genotype et kendt locus med deletioner uden behov for rå sekvensfil. Endelig præsenterer den fremgang til fælles imputation af WGS varianter samtidig med store kromosomale deletioner.

Kapitel 5 beskriver associeringsundersøgelser af alleler på tværs af hele genomet med hunlig frugtbarhed i tre nordiske mælkekvægracer ved anvendelse af imputerede WGS-varianter og store kromosomale deletioner. Denne undersøgelse er baseret på analyser af otte fertilitetsrelaterede egenskaber som anvender single-marker association, både betingede og fælles analyser. Dette studie viser, at der er inflation i teststatistikken for samling af testresultater selv efter korrektion for populationsopdeling ved anvendelse af egenvektorer fra den genomiske slægtskabsmatrice for prøverne; Dette var dog kendt for egenskaber med stærke polygeniske virkninger, med mere. Til sidst rapporteres kortlægning af flere nye loci, der koder for kvantitative egenskaber (QTL), sammen med de tidligere kendte i denne undersøgelse. Denne undersøgelse fremhæver også vigtigheden af at inkludere (imputerede) store deletioner når fertilitetsegenskaber kortlægges.

Kapitel 6 beskriver en prædiktion af genomiske avlsværdier for frugtbarhed (fertilitetsindeks) ved anvendelse af SNP array-chip genotyper, valgt QTL og store kromosomale deletioner. Ved anvendelse af genomisk ”Best Linear Unbiased Prediction” (GBLUP) med en eller flere genomiske slægtskabsmatricer afledt af et sæt markerede markører, viser denne undersøgelse højere prædiktionsevne sammenlignet med tidligere undersøgelser. Denne undersøgelse fremhæver også indflydelsen af at vælge markører med den bedste forudsigelighed, især for en race med en lille træningspopulation, i forhold til nøjagtighed af genomisk prædiktion. Til sidst vises at store deletioner generelt har bedre prædiktionsevne.

Til sidst giver kapitel 7 en generel diskussion, konklusion og perspektivering baseret på denne ph.d.-afhandling.

Résumé en Français

La fertilité est un caractère important du point de vue économique chez les vaches laitières. La fertilité au sens large est définie comme la capacité de la vache à exprimer l'oestrus, à concevoir après l'insémination, et à restaurer sa cyclicité sexuelle après le vêlage. Ce caractère a subi une dégradation au cours de ces 30 dernières années, en particulier du fait de la sélection intense sur la production – caractère présentant une corrélation génétique négative avec la fertilité – basée sur l'utilisation d'un nombre limité de taureaux aux valeurs génétiques élevées. La fertilité est un caractère difficile à sélectionner par sélection généalogique en raison de sa faible héritabilité. L'utilisation d'information génomique dans le schéma de sélection peut apporter une solution à ce problème des caractères à faible héritabilité. L'identification des variants causaux, lorsque cela est possible, et / ou la sélection à l'aide d'un ensemble de marqueurs prédictifs eux-mêmes associés aux variants causaux, est essentielle pour pour accroître l'efficacité de la sélection génomique sur ces caractères mais également pour définir les plans d'accouplement (par exemple, pour éviter les accouplements entre porteurs de variants létaux).

L'objectif général de cette thèse est d'identifier les variants causaux ou, à défaut, un ensemble de marqueurs prédictifs - qui présentent un déséquilibre de liaison élevé avec les variants causaux - pour la fertilité des vaches laitières. Nous avons abordé cet objectif général dans cinq articles / manuscrits présentés aux chapitres 2 à 6 de cette thèse.

Le chapitre 2 décrit une approche systématique de cartographie des variants létaux récessifs chez les bovins Normands français basée sur la recherche de déficit en haplotypes homozygotes (HHD). Cette étude montre l'influence de la taille de l'échantillon, de la qualité des génotypes, de la qualité du phasage des génotypes en haplotypes et de l'imputation, de l'âge de l'haplotype et enfin, de la définition des seuils de signification prenant en compte les tests multiples, sur la découverte et la reproductibilité des résultats de HHD. Elle illustre également l'importance de la cartographie fine avec les données de généalogie et de séquence de génome entier (WGS), l'annotation intégrative (entre espèces) pour hiérarchiser les mutations candidates et, enfin, le génotypage à grande échelle de la mutation candidate, pour valider ou invalider les mutations initiales. Cette étude met en évidence une mutation létale dans le gène CAD dont la fréquence dans la population est de 3%.

Le chapitre 3 décrit une cartographie à haute résolution de grandes délétions chromosomiques de séquences du génome dans une population de 175 animaux appartenant à trois races laitières nordiques. Cette étude utilise trois approches différentes pour valider les résultats de la

cartographie. Le chapitre décrit les propriétés génétiques des populations et l'importance fonctionnelle des délétions identifiées. Enfin, il illustre les mécanismes de formation des délétions basés sur les caractéristiques des séquences d'ADN aux points de cassure. Cette étude fournit les bases pour les études génétiques ultérieures réalisées au cours de cette thèse sur le délétions.

Le chapitre 4 traite de trois questions liées à l'imputation de variants structuraux, ici de délétions chromosomiques importantes: la disponibilité des génotypes de délétion, la taille du panel de référence d'haplotypes et, enfin, l'imputation elle-même. Pour aborder les deux premières questions, cette étude décrit une approche basée sur un modèle de mélange gaussien dans laquelle les données de profondeur de lecture provenant de fichiers au format VCF (variant call format) sont utilisées pour génotyper un locus de délétion connu, en l'absence d'information sur la séquence brute. Enfin, il présente un pipeline pour l'imputation conjointe de variants WGS et de grandes délétions chromosomiques.

Le chapitre 5 décrit des études d'association pangénomiques de la fertilité femelle dans trois races de bovins laitiers nordiques à l'aide de variants WGS imputés et de grandes délétions chromosomiques. Cette étude concerne huit caractères de fertilité et utilise des analyses d'association mono-marqueur, conditionnelles et conjointes. Cette étude montre qu'une surestimation, ou « inflation », des statistiques de test peut être observée même après correction pour la stratification de la population à l'aide de composantes principales génomiques et pour les structures familiales à l'aide de matrices de relations génomiques. Ce biais était connu pour les caractères très polygéniques. Enfin, cette étude présente plusieurs locus de traits quantitatifs (QTL) nouveaux et confirme plusieurs autres déjà connus. Elle souligne également l'importance d'inclure les grandes délétions (imputées) pour la cartographie par association des caractères de fertilité.

Le chapitre 6 décrit la prédiction des valeurs génomiques de fertilité (ou indice de fertilité) à l'aide de génotypes à puces SNP, de QTL sélectionnés et de délétions chromosomiques importantes. En utilisant la méthode de meilleure prédiction linéaire sans biais génomique (GBLUP) avec une ou plusieurs matrices de relations génomiques dérivées d'un ensemble de marqueurs sélectionnés, cette étude rapporte une précision de prédiction améliorée. Cette étude met également en évidence l'influence de la sélection des marqueurs les plus prédictifs, en particulier pour une race ayant une population d'apprentissage réduite, sur la précision des

prédictions génomiques. Enfin, les résultats démontrent que les grandes délétions ont en général un pouvoir prédictif élevé.

Enfin, **le chapitre 7** propose une discussion générale, une conclusion et des perspective basées sur cette thèse.

Table of Contents

Acknowledgement i

Résumé in English iii

Résumé på Dansk v

Résumé en Français vii

Table of Contents x

List of Publications xii

Abbreviations xiii

Chapter 1. General Introduction 1

1.1 Large-scale sequencing and genotyping 3

1.2 Genetic Markers 4

1.3 Homozygous Haplotype Deficiency (HHD) 7

1.4 Genotype Imputation 8

1.5 Genome-wide association study (GWAS) 9

1.6 Genomic Prediction 11

1.7 Aim and objectives of this PhD study 13

1.8 References 14

Chapter 2. A missense mutation (p.Tyr452Cys) in the CAD gene compromises reproductive success in French Normande cattle 19

2.1 Abstract 20

2.2 Introduction 20

2.3 Methods 21

2.4 Results and Discussion 26

2.5 Conclusion 38

2.6 References 39

Chapter 3. Genome-wide mapping of large deletions and their population-genetic properties in dairy cattle 44

3.1 Abstract 45

3.2 Introduction 45

3.3 Materials and methods 46

3.4 Results and discussion 50

3.5 Conclusions 59

3.6 References 60

Chapter 4. Joint imputation of whole-genome sequence variants and large chromosomal deletions in cattle 66

4.1 Abstract 67

4.2 Introduction 67

4.3 Methods 68

4.4 Results and discussion 72
4.5 Conclusion 76
4.6 References 81
Chapter 5. Genome-wide association study with imputed whole-genome sequence variants including large deletions for female fertility in three Nordic dairy breeds 84
5.1 Abstract 85
5.2 Introduction 85
5.3 Methods 86
5.4 Results and Discussion 87
5.5 Conclusion 90
5.6 References 108
Chapter 6. Genomic prediction for female fertility using imputed whole-genome sequence variants including large chromosomal deletions 110
6.1 Abstract 111
6.2 Introduction 111
6.3 Methods 112
6.4 Results and Discussion 113
6.5 Conclusion 115
6.6 References 118
Chapter 7. General Discussion 119
7.1 Recessive Lethals 120
7.2 Use of Whole-Genome Sequence Variants 121
7.3 Genomic Prediction 123
7.4 Evolutionary Conservation as a Tool for Identifying Lethal Genes 125
7.5 Conclusions 126
7.6 Perspectives 126
7.7 References 128
Individual Training Plan 130

List of Publications

Publications/Manuscripts included in this thesis

1. **Mesbah-Uddin, M.**, C. Hozé, P. Michot, A. Barbat, R. Lefebvre, M. Boussaha, G. Sahana, S. Fritz, D. Boichard, and A. Capitan. 2019. *A missense mutation (p.Tyr452Cys) in the CAD gene compromises reproductive success in French Normande cattle.* Journal of Dairy Science 102:6340-6356 [doi: https://doi.org/10.3168/jds.2018-16100] (**Chapter 2**)
2. **Mesbah-Uddin, M.**, B. Guldbrandtsen, T. Iso-Touru, J. Vilkki, D. J. De Koning, D. Boichard, M. S. Lund, and G. Sahana. 2018. *Genome-wide mapping of large deletions and their population-genetic properties in dairy cattle.* DNA Research 25:49-59 [doi: https://doi.org/10.1093/dnares/dsx037] (**Chapter 3**)
3. **Mesbah-Uddin, M.**, B. Guldbrandtsen, M. S. Lund, D. Boichard, and G. Sahana. 2019. *Joint imputation of whole-genome sequence variants and large chromosomal deletions in cattle.* Journal of Dairy Science 102:11193-11206 [doi: https://doi.org/10.3168/jds.2019-16946] (**Chapter 4**)
4. **Mesbah-Uddin, M.**, B. Guldbrandtsen, A. Capitan, M. S. Lund, D. Boichard, and G. Sahana. *Genome-wide association study with imputed whole-genome sequence variants including large deletions for female fertility in three Nordic dairy breeds.* [**Manuscript in preparation**] (**Chapter 5**)
5. **Mesbah-Uddin, M.**, A. Capitan, B. Guldbrandtsen, M. S. Lund, G. Sahana, and D. Boichard. *Genomic prediction for female fertility using imputed whole-genome sequence variants including large chromosomal deletions.* [**Manuscript in preparation**] (**Chapter 6**)

Other Publication/Manuscripts (not included in the thesis)

1. **Mesbah-Uddin, M.**, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2018. *Genotype call for chromosomal deletions using read-depth from whole genome sequence variants in cattle.* Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.662, Auckland, New Zealand.
2. Wu, X., **M. Mesbah-Uddin**, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2019. *Haplotypes responsible for early embryonic lethality detected in Nordic Holstein.* Journal of Dairy Science 102:11116-11123 [doi: https://doi.org/10.3168/jds.2019-16651]
3. Hoze, C., C. Escouflaire, **M. Mesbah-Uddin**, A. Barbat, M. Boussaha, M. C. Deloche, D. Boichard, S. Fritz, and A. Capitan. 2019. *Short Communication: A splice site mutation in CENPU is associated with recessive embryonic lethality in Holstein cattle.* Journal of Dairy Science [doi: https://doi.org/10.3168/jds.2019-17056]

Abbreviations

AISc	number of inseminations per conception in cows
AISh	number of inseminations per conception in heifers
CNV	Copy-number variation
DEL	Deletion (size >50 bp)
FI	Fertility index
GBLUP	Genomic best linear unbiased prediction
GEBV	Genomic estimated breeding value
GWAS	Genome-wide association study
HH	Holstein haplotype
HHD	Homozygous haplotype deficiency
HOL	Holstein
ICF	interval (number of days) from calving to first insemination in cows
IFLc	interval (number of days) from first to last insemination in cows
IFLh	interval (number of days) from first to last insemination in heifers
INDEL	Small insertion/deletion (size <50 bp)
JER	Jersey
LD	Linkage-disequilibrium
MAF	Minor allele frequency
MLM	Mixed linear model
NGS	next-generation sequencing
NH	Normande haplotype
NRRc	non-return rate in cows
NRRh	non-return rate in heifers
QTL	Quantitative trait loci
RDC	Nordic Red Dairy cattle
SAM/BAM	Sequence alignment file/binary SAM
SNP	Single-nucleotide polymorphism
SV	Structural variant
VCF	Variant call format file
WGS	Whole-genome sequence

Chapter 1.
General Introduction

Female fertility is an economically important trait in dairy cattle. Due to low heritability (0.01 to 0.04) of this trait, it is relatively hard to improve using traditional pedigree-based selection scheme. Fertility is negatively correlated with production traits such as milk and protein yield (Pryce et al., 1997, Rauw et al., 1998, Roxström et al., 2001). Therefore, an intensive selection for production traits, over the years, using few bulls with high breeding values resulted in a decrease in cows' fertility. The indirect response to selection for production traits explains only half of the reduction seen in fertility. Factors such as higher inbreeding rate (Howard et al., 2017) also contributes to this decreasing trend, primarily due to embryonic lethality (VanRaden et al., 2011, Cole et al., 2016). However, selection based on genetic markers, with or without pedigree information, in a scheme known as 'genomic selection' (Meuwissen et al., 2001), provides an opportunity to overcome this negative trend in female fertility (Garcia-Ruiz et al., 2016). For genomic selection, large reference population were assembled for major dairy breeds with high quality of phenotypes and genotypes. Due to their large size and their family structure (artificial insemination bulls with many daughters with fertility records), these reference populations allow for high fertility evaluation reliability, in spite of this low heritability. Noteworthy, these reference populations are also a unique resource for identifying the causal factors for recessive genetic defects and mapping quantitative trait loci (QTL).

Recessive lethals are responsible for a substantial economic loss in dairy cattle (Cole et al., 2016). To identify causal factors/QTL for recessive lethals, two approaches could be used. First, searching for homozygous haplotype deficiency (HHD) in the population. This approach does not require phenotype. Here, when homozygotes for a common haplotype are absent or less frequent in the population than expected, may indicate embryonic lethality. This HHD approach has been shown to be very successful in cattle where genotyping is widely used and several discoveries were made in last few years. For example, seven recessive haplotypes in Holstein were discovered using HHD, namely, HH1 to HH5 (VanRaden et al., 2011, Fritz et al., 2013, Sahana et al., 2013, Daetwyler et al., 2014, McClure et al., 2014), HH6 (Fritz et al., 2018) and HH7 (Hoze et al., 2019). The second approach is performing genome-wide association studies (GWAS) for fertility traits. For example, in Nordic Red Dairy cattle, two large chromosomal deletions, one affecting fertility (Schulman et al., 2011, Kadri et al., 2014) and the other causing stillbirth (Sahana et al., 2016), were mapped using GWAS approach. In recent years, a third approach, next-generation sequencing (NGS)-based reverse-genetics also becoming a popular choice with lot of success both in humans (MacArthur et al., 2012, Rivas

et al., 2015) and in animals (Charlier et al., 2016, Li et al., 2016, Michot et al., 2016). However, in this PhD thesis, we used former two approaches.

Although identification of causal variants is very challenging, the identification of recessive deleterious alleles is very important to manage their frequencies in a population. In genomic selection, use of causal variants information could theoretically provide higher accuracy of evaluation (van den Berg et al., 2016), and persistence of prediction accuracy over generations (Meuwissen and Goddard, 2010). In mating plans, this information can also be used to avoid at-risk mating and thus can reduce embryonic mortality in the population (Cole et al., 2016). As whole-genome sequencing (WGS) provides a nearly complete list of variants, it may facilitate to identify causal variants. This is one of the major advantage of using WGS data over standardized genotyping panels—that include only a small proportion of the variants and are unlikely to include the causal ones. NGS is rapidly becoming the primary focus to characterize the genetic basis of quantitative traits. WGS could then be used to fine-map the causal mutations. This information could subsequently be implemented in the breeding plan to avoid carrier-to-carrier mating, thus improving fertility in dairy cattle.

The organization of this PhD thesis is as follows. In **Chapter 1**, I have introduced the topic of this thesis with relevant backgrounds followed by the overall aim of the PhD and five study specific objectives. In subsequent five chapters (**Chapter 2** to **Chapter 6**), I have presented the main results and discussions of this PhD thesis under five separate articles. Finally, in **Chapter 7**, I have presented general discussions and future perspectives based on this PhD study.

1.1 Large-scale sequencing and genotyping

This PhD study utilized WGS and large-scale genotyping data that was generated/gathered through two primary international collaborations, namely, the 1000 Bull Genomes Project (1KBGP) (Daetwyler et al., 2014) and the EuroGenomics consortium (Lund et al., 2011, Boichard et al., 2018). The 1KBGP is a large-scale WGS consortium that provided a (nearly) full catalog of small sequence variants (e.g. SNPs and indels) for all modern cattle breeds, which is a valuable resource for identifying candidate mutation. The 1KBGP also provides a large reference panel for imputation of WGS variants into the SNP array-typed cattle populations. The EuroGenomics consortium was initiated with a goal of enlarging reference population for genomic selection. This consortium also designed a low-density custom SNP array, called EuroG10k, for large-scale genotyping. EuroG10k has one common part that

contain SNPs from BovineLD, Bovine50k and BovineHD BeadChips to assure better imputation, and several private parts that contain various classes WGS markers, such as, QTL detected through GWAS, structural variants, and putative causal variants selected using NGS-based reverse-genetics approach (Boichard et al., 2018).

1.2 Genetic Markers

Today, most genetic studies, in humans (Genomes Project et al., 2015), mouse (Li et al., 2015, Bult et al., 2019), and major agricultural species, such as cattle (Daetwyler et al., 2014, Bouwman et al., 2018), sheep (Jiang et al., 2014), pig (Li et al., 2013), chicken (Wong et al., 2004), wheat (He et al., 2019, Pont et al., 2019), rice (Zhao et al., 2018), and many other species, are based on genetic markers derived from whole-genome sequencing (WGS). This revolution in genomics is due to a rapid advancement of WGS techniques coupled with gradual decline in both sequencing and genotyping cost (Goodwin et al., 2016). In most of these studies, these markers include small sequence variants such as single-nucleotide polymorphism (SNP) and small (usually <50 bp) insertion and/or deletion (indel). Large (>50 bp) DNA polymorphisms known as structural variants (SVs) are studied less frequently, primarily because genotype calling is not straightforward and requires complex procedures. These SVs include deletions, duplications, insertions, inversions, and translocations. Unbalanced SVs, i.e. fewer or more copies of some sequences in an organism compared with the reference genome, are called copy-number variants (CNVs) (Alkan et al., 2011). Figure 1.1 illustrates some of these DNA polymorphisms and Figure 1.2 shows a generic pipeline for mapping and genotyping WGS variants.

In short, different methods can be used to call SV genotypes. Some methods rely on WGS data and use different kinds of information: variation in read depth, for insertions and deletions, unexpected read pairs (too close or too far to each other), split reads (when two parts of the same read map to different regions of the genome assembly). Some other methods rely on SNP arrays signals: different intensities may reflect copy number variants (Kadri et al., 2012); Mendelian incompatibilities for close markers are often the consequences of deletions (Capitan et al., 2012); finally, properly designed tests may be used to test breakpoints of structural variants (Boichard et al., 2018). This latter approach is very powerful as it can generate a very large number of reliable genotypes at a given variant and is a method of choice for targeted SV.

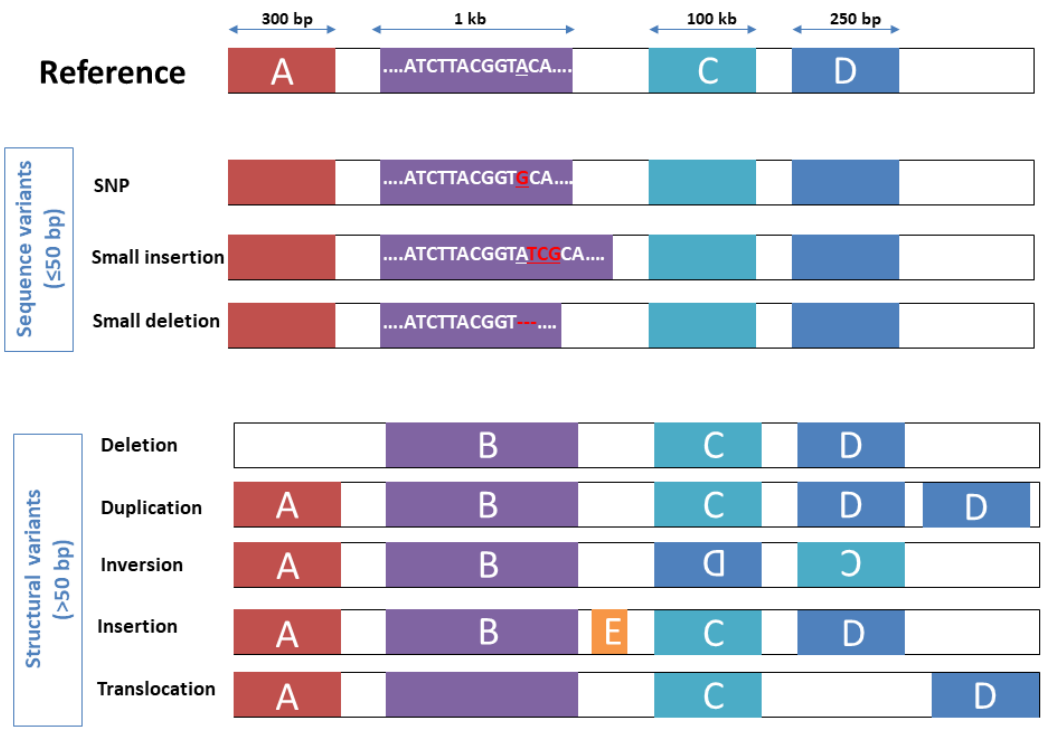


Figure 1.1. An illustration of various classes of whole-genome sequence variants. (Adopted from presentation slides by Mesbah-Uddin, M., GenSAP Annual Meeting 2017, http://mbg.au.dk/fileadmin/gensap/events/2017/Session1/Md_Mesbah_Uddin_GenSAP2017.pdf)

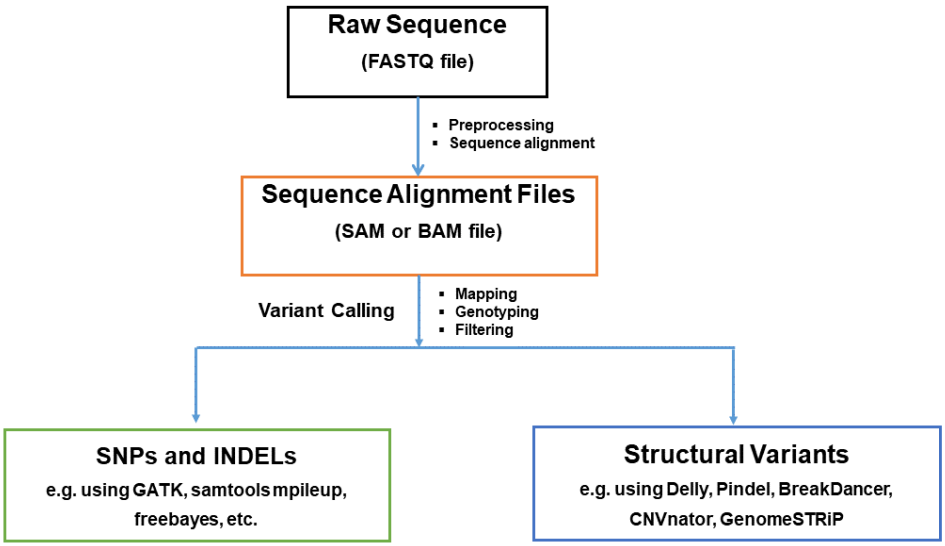


Figure 1.2. A generic pipeline for whole-genome sequence variants calling.

The focus of several studies of this PhD study was on large chromosomal deletions due to its potential as causal variants especially for fitness related traits. Furthermore, there were fewer studies where large deletions or any other types of SVs were investigated beyond mapping their genomic locations and estimating their population frequencies. However, mapping SVs in itself is a challenging task (Bickhart and Liu, 2014). Moreover, SVs studies also suffers from high false discovery rates, and often requires the use of multiple SV callers to get a consensus (Figure 1.3), or requires population scale sequencing data to boost the discovery signal. Small SV discovery population size, which is typical for WGS dataset in most research labs, is a further constraint on subsequent imputation and genetic study. One solution to small sample size problem could be the use of shared genomic resources available through different collaboration, such as the 1000 Bull Genomes Project where variant call format (VCF) files are shared among the collaborators. It is important to note, however, that BAM files are not always available for these resources and we need methods adapted to VCF files. These files include read-depth data supporting each SNP genotype call and this information can be used to estimate SV genotypes, as shown in **Chapter 4** of this thesis. Of course, the success of this approach will heavily depend on the sequencing read-depth, size of SV region, and more importantly availability of SNPs within the interrogated region. Since, this technique analyses the read-depth variation to estimate SV genotypes, one could only study deletions (copy-number loss) or duplications (copy-number gain). An illustration of SNP read-depth based genotyping of large deletion is presented in Figure 1.4.

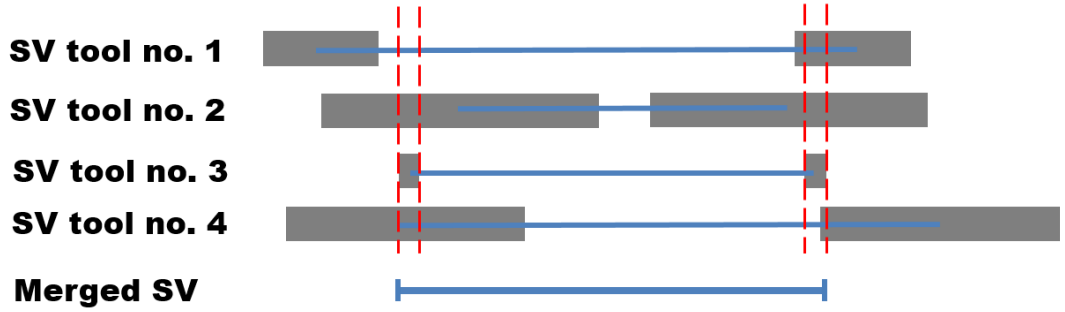


Figure 1.3. Illustration of merging SV calls generated using multiple tools. *Shaded-box represents confidence interval around the SV breakpoints and the straight horizontal line represents start and end of the breakpoint. Here, for breakpoints, SV call with small confidence interval is given preference over other calls, when there is a ~70% reciprocal overlap between the calls.*

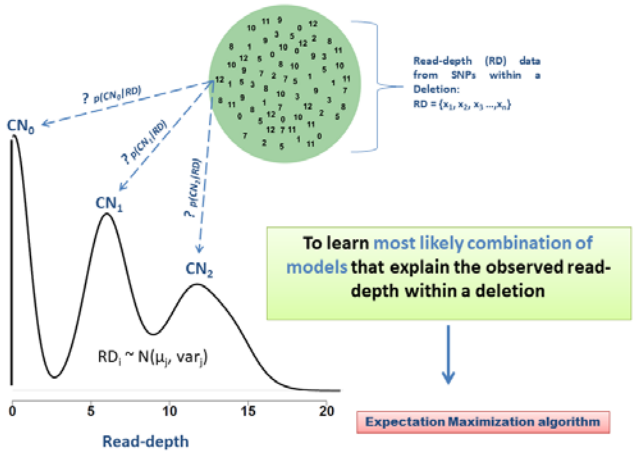


Figure 1.4. Graphical presentation of SNP read-depth based genotyping of large deletion. *CN: Copy-number; RD: read-depth; μ_j and var_j represent mean and variance of j^{th} Gaussian. (adopted from Mesbah-Uddin et al. (2018))*

1.3 Homozygous Haplotype Deficiency (HHD)

Complete absence of a common haplotype in homozygous state or fewer occurrences than expected under neutrality in a large cohort of genotyped animals may indicate lethal or semi-lethal mode of action of the haplotype. Since the introduction of this technique by VanRaden et al. (2011), numerous lethals were discovered in several dairy cattle breeds (Fritz et al., 2013, Sahana et al., 2013, Pausch et al., 2015, Kipp et al., 2016, Schwarzenbacher et al., 2016, Michot et al., 2017, Fritz et al., 2018).

Search for HHD is a vital tool for mapping genetic defects in situations where direct phenotypes are not available, such as for embryonic lethals. This search is quite straightforward. In most routine genomic evaluation procedures, genotypes are imputed and therefore phased. These phased data are used to estimate haplotype frequencies, expected and observed number of homozygous haplotypes, and to infer the status of each individual carrier. In the targeted region, candidate variants and genes can be identified by comparing the carrier-status at the candidate variant for sequenced individuals with their carrier-status at the haplotype. Among the usually limited number of candidate variants, the best one is selected based on its functional annotation, such as predicted functional effect, sequence conservation, and cross-species annotation. In many situations, the prioritized variant locate within the coding sequence of a gene and its annotation is accurate. Finally, this candidate variant (and possibly several others) can be added on the SNP chip used for routine genotyping in genomic selection program, providing large numbers animals with direct genotypes. This information is especially powerful to validate (or invalidate) the candidate variant.

HHD method is especially efficient when the haplotype frequency is moderate to high and linkage disequilibrium (LD) with the causal variant is strong. As an example, a few thousands genotypes are required to map an embryonic lethal at a frequency of 0.1, when 0 homozygote is observed instead several tens (1%) are expected. However, the expected number of homozygotes becomes very limited when the frequency is low (e.g. 1/10,000 when the frequency is 1%) and very large numbers, e.g. hundreds of thousands, of genotyped animals are often required to identify lethal haplotypes with low to moderate frequencies. Due to the availability of large-scale genotyping data in genomic selection, search for HHD is a routine practice for dairy cattle breeds in many countries. Furthermore, as the size of genotyped cohort increases over time, the power of detecting rare and low frequency causal variants with HHD also increases. For example, Hoze et al. (2019) recently mapped a candidate mutation in *CENPU* gene with an allele frequency of 0.8% in French Holstein that was tagged by Holstein haplotype 7 (HH7) with a population frequency of 1.1%. In many situations, these lethal variants were found to be 'untolerated' mutations in the coding sequence, in genomic regions highly conserved in the evolution, and results in a truncated, incomplete, and usually non-functional protein product. For instance, several of these mutations responsible for early embryonic mortality affect cell division (Fritz et al., 2018) or DNA base synthesis pathways (Michot et al., 2017, Mesbah-Uddin et al., 2019), i.e. basic mechanisms of life.

1.4 Genotype Imputation

Genotype imputation is a critical step in most modern genetic studies. Indeed genomic information deeply varies across individuals from low density SNP chip to WGS at various depths, i.e. from several thousand to several tens of millions markers. Inferring the complete information for all individuals is a prerequisite for subsequent genetic analysis such as genome-wide association studies or genomic evaluation. Imputation is a statistical approach to predict genotypes for un-typed markers based on a densely genotyped haplotype reference panel relative to the imputation target population. Because these methods use haplotype information, mostly they also include the genotype phasing process when genotyped are unphased.

There are several tools available for genotype phasing and imputation with very different approaches (Marchini and Howie, 2010, van Leeuwen et al., 2015). The methods derived from human genetics, such as Beagle (Browning and Browning, 2016), Minimac (Das et al., 2016), and IMPUTE2 (Marchini et al., 2007) are the oldest. These tools rely on LD at the population level and use Bayesian algorithms. More recently, algorithm more suited to livestock

population were developed, taking advantage of the small effective population size and the large half-sib family size, i.e. the long-range LD and the small number of haplotypes segregating in the populations. These methods build the library of the haplotypes observed in the reference population and then use this library to fill the missing information. In addition, they efficiently use pedigree information. FImpute (Sargolzaei et al., 2014) is the most widely used software for phasing and imputation in cattle populations. The relative efficiency of imputation strongly varies according to the size of the population, to its structure, extent of LD, and marker density. A combination of methods for phasing and then imputation is sometimes found optimal (Delaneau et al., 2014). FImpute predicts (usually) the most likely genotypes; however, FImpute does not provide imputation accuracy information as an output, contrary to Beagle, Minimac and IMPUTE2. Bayesian methods provide expected genotypes (i.e. allele dosages) which can take into account the uncertainty of the imputed genotypes and therefore, more appropriate than 'best-guess' genotypes for downstream analysis like association studies.

1.5 Genome-wide association study (GWAS)

Genome-wide association study (GWAS) is a method of choice for identifying quantitative trait loci (QTL) in virtually all species, due to the advent of high throughput genotyping and sequencing techniques. In GWAS, a dense array of genetic markers, either directly genotyped or imputed, is used to capture a substantial proportion of common genetic variation of the given species (McCarthy et al., 2008). GWAS exploits LD pattern at population level, i.e. the non-random association of alleles at different loci (Slatkin, 2008), to identify association of markers or haplotypes with the trait of interest (and QTL, if any). In ideal situation, these chromosome segments usually carry identical markers or haplotypes, and hence are expected to carry QTL alleles (Hayes, 2013).

Over the years, GWAS has proven to be an efficient approach for gene mapping both in monogenic and polygenic traits. For example, Charlier et al. (2008) first reported fine-mapping of causal variants for five monogenic traits in cattle. Similarly, earlier studies by Barendse et al. (2007), Kolbehdari et al. (2008), Daetwyler et al. (2008) and Druet et al. (2008) also showed success in mapping variants associated with polygenetic traits. Following those early studies, several hundreds of GWASs were performed in cattle over the last decade.

There are numerous statistical methods for association mapping. Today, single-marker association using mixed linear model (MLM) is a method of choice for most GWAS. In MLM analysis, false discoveries could be limited by accounting for both population stratification

(usually by using few principal components (PCs) of the genomic relationship matrix) and relationships among individuals (by using pedigree or genomic kinship matrix) (Zhang et al., 2010). Henderson's matrix notation of the MLM is as follows:

$$y = X\beta + Zu + e$$

where *y* is the vector of observed phenotypes; *u* is a vector of random additive genetic effects; *X* and *Z* are the known design matrices; and *e* is a vector of residual effects. β is a vector containing the fixed effects: in addition to the effect of the genetic marker of interest, it includes other fixed effects such as the population mean, sex, age, PCs, etc. The *u* and *e* vectors are assumed to be normally distributed with a null mean and a variance of: $var\begin{pmatrix} u \\ e \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}$, where $G = 2K\sigma_a^2$ with $\sigma_a{}^2$ as the additive genetic variance and *K* as the kinship matrix. If homogeneous variance is assumed for the residual effect, $R = I\sigma_e{}^2$, where $\sigma_e{}^2$ is the residual variance (it includes all other variances that is not captured by $\sigma_a{}^2$), and *I* is an identify matrix. The proportion of the total variance explained by the additive genetic variance is defined as narrow sense heritability, $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$.

False positive association. Two major sources of false positives in GWAS is population stratification and hidden relatedness in the mapping population. These biases cause spurious association in GWAS. However, in MLM, correction for population structure and relatedness could be done using first few PCs and genomic relationship matrix.

Multiple-testing correction. Typically, hundreds of thousands to several millions of hypothesis testing are performed to identify markers associated with a trait. Therefore, stringent *p-value* threshold is required to account for these multiple testing and reduce false discoveries (type 1 error). Bonferroni correction is most commonly used multiple testing adjustment: here, *p-value* is divided by the total number of tests performed. In GWAS, a commonly used *p-value* threshold is $<5\times10^{-8}$, at 5% level of significance and ~1 million independent tests (i.e. $0.05/10^6$).

Conditional GWAS. Single-marker GWAS often provides hundreds of loci with significant *p-value* (given the study had enough power to capture true QTL, i.e. able to reject null hypothesis when alternative is true). Conditional association analysis could then be performed for a given locus or for a chromosome to identify independent GWAS signal(s). Here, marker with the best association signal is used as a cofactor for subsequent round of GWAS.

Meta-analysis. Meta-analysis is a powerful tool to boost QTL detection power by combining GWAS summary statistics from multiple populations (Willer et al., 2010); examples of this approach in cattle include multi-breed meta-analysis for cattle stature (Bouwman et al., 2018) and milk protein composition in dairy cattle (Sanchez et al., 2017). Following the method by Bolormaa et al. (2014), multi-trait meta-analysis is also used in cattle to identify QTL affecting multiple traits, i.e. pleotropic QTL; example of this approach in cattle includes meta-analysis for mammary gland morphology (Pausch et al., 2016).

1.6 Genomic Prediction

Early prediction of genetic merits of an organism is of vital importance for any breeding program concerning agricultural species. Availability of large-scale SNP array genotypes and WGS data provides tremendous advantage on selecting breeding population based on their breeding values estimated from genomic information, with or without pedigree information, rather than selecting solely based on performances and pedigrees. Genetic gain (yearly), $\Delta G = \frac{ir\sigma_a}{L}$, depends on four parameters: selection intensity (i), selection accuracy (r), generation interval (L), and additive genetic standard deviation (σ_a). Genomic selection can affect at least three of these parameters. In its simplest form, substantial reduction in generation time is possible since genomic evaluation can be performed at birth. In cattle, traditional selection was late due to progeny testing of bulls and performance testing of cows. Schaeffer (2006) showed that genomic selection could double genetic gain by substantial reduction of generation interval in dairy cattle. However, this situation is not general to all species, especially those with shorter generation interval, which explains why genomic selection is so popular in cattle. Genomic selection can also affect selection accuracy if the reference population is large, and selection intensity if genotyping is cheap, allowing large-scale evaluation of many candidates. Nevertheless, genomic selection has brought a revolutionary change in selection and breeding for all major agricultural species (Georges et al., 2019).

The accuracy of prediction is a crucial component for any successful genomic selection program. Numerous studies were performed over the last few years. To improve prediction accuracy, different approaches were tested:

- some studies focused on the prediction methods (Meuwissen et al., 2001, de Los Campos et al., 2013, Kemper et al., 2015, Moser et al., 2015): different priors were used, from a constant Gaussian prior to all SNP (GBLUP) to a prior reflecting an oligogenic determinism;

- few focused on training population characteristics, such as size, genetic relationship with the candidates, and genetic parameters of the trait, such as heritability and genetic correlation (Calus et al., 2018, van den Berg et al., 2019);
- others focused on selecting better predictors (Sørensen et al., 2014, Brøndum et al., 2015, VanRaden et al., 2017, Hay et al., 2018), i.e. not only generic SNP but variants with a higher probability of biological effects. These predictors were selected on the basis of their functional annotation or on the effect on the phenotype, as estimated in independent populations.

However, there are many opportunities yet to be explored to improve prediction accuracy further. Besides SNPs and indels, inclusion of other classes of genetic markers, such as CNVs, also needs further exploration. Feature selection using reverse-genetic approach could also be very useful for prediction purposes, as shown by Charlier et al. (2016), Michot et al. (2016) and Boichard et al. (2018). Last but not the least, automated feature selection using machine learning approaches can also be considered, such as using deep learning approaches as shown by Bellot et al. (2018) for predicting complex traits in humans, although the success so far was limited.

1.7 Aim and objectives of this PhD study

The overall aim of this PhD thesis was to identify causal variants for deleterious recessive mutations and select a set of predictive markers for female fertility in dairy cattle. We addressed this broad aim under five separate studies that I have presented in **Chapter 2** to **Chapter 6** in this thesis with the objectives of each study listed below:-

1. To validate (or invalidate) six previously reported lethal haplotypes in Normande cattle, to identify new loci, and when possible, to identify causal variants in this dairy breed (**Chapter 2**)
2. To map large chromosomal deletions from whole-genome sequences of 67 Holstein, 27 Jersey, and 81 Nordic Red Dairy cattle, and to analyse their (identified deletions) population-genetic properties (**Chapter 3**)
3. To estimate deletion genotype-likelihood using SNP read-depth data from VCF file, to build a haplotype reference panel that include SNPs, indels, and deletions, and to jointly impute these variants into the existing SNP array-typed animals (**Chapter 4**)
4. To perform genome-wide association studies with imputed SNPs, indels and deletions for eight female fertility traits in Holstein, Jersey and Nordic Red Dairy cattle (**Chapter 5**)
5. To investigate the effect of imputed WGS selected markers on accuracy of genomic prediction for female fertility in Holstein, Jersey and Nordic Red Dairy cattle (**Chapter 6**)

1.8 References

Alkan, C., B. P. Coe, and E. E. Eichler. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12:363-376. https://doi.org/10.1038/nrg2958.

Barendse, W., A. Reverter, R. J. Bunch, B. E. Harrison, W. Barris, and M. B. Thomas. 2007. A validated whole-genome association study of efficient food conversion in cattle. Genetics 176:1893-1905. https://doi.org/10.1534/genetics.107.072637.

Bellot, P., G. de Los Campos, and M. Perez-Enciso. 2018. Can Deep Learning Improve Genomic Prediction of Complex Human Traits? Genetics. https://doi.org/10.1534/genetics.118.301298.

Bickhart, D. M. and G. E. Liu. 2014. The challenges and importance of structural variation detection in livestock. Front Genet 5:37. https://doi.org/10.3389/fgene.2014.00037.

Boichard, D., M. Boussaha, A. Capitan, D. Rocha, C. Hozé, M. P. Sanchez, T. Tribout, R. Letaief, P. Croiseau, C. Grohs, W. Li, C. Harland, C. Charlier, M. S. Lund, G. Sahana, M. Georges, S. Barbier, W. Coppieters, S. Fritz, and B. Guldbrandtsen. 2018. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. Page 675 in Proc. Proc. World Congr. Genet. Appl. Livest. Prod., Auckland, New Zealand. AL Rae Centre for Genetics and Breeding,

Massey University, Palmerston North, New Zealand. http://www.wcgalp.org/system/files/proceedings/2018/experience-large-scale-use-eurogenomics-custom-snp-chip-cattle.pdf.

Bolormaa, S., J. E. Pryce, A. Reverter, Y. Zhang, W. Barendse, K. Kemper, B. Tier, K. Savin, B. J. Hayes, and M. E. Goddard. 2014. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLoS Genet 10:e1004198. https://doi.org/10.1371/journal.pgen.1004198.

Bouwman, A. C., H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, H. Pausch, R. F. Brøndum, P. J. Bowman, B. Thomsen, B. Guldbrandtsen, M. S. Lund, B. Servin, D. J. Garrick, J. Reecy, J. Vilkki, A. Bagnato, M. Wang, J. L. Hoff, R. D. Schnabel, J. F. Taylor, A. A. E. Vinkhuyzen, F. Panitz, C. Bendixen, L. E. Holm, B. Gredler, C. Hozé, M. Boussaha, M. P. Sanchez, D. Rocha, A. Capitan, T. Tribout, A. Barbat, P. Croiseau, C. Drögemüller, V. Jagannathan, C. Vander Jagt, J. J. Crowley, A. Bieber, D. C. Purfield, D. P. Berry, R. Emmerling, K. U. Götz, M. Frischknecht, I. Russ, J. Sölkner, C. P. Van Tassell, R. Fries, P. Stothard, R. F. Veerkamp, D. Boichard, M. E. Goddard, and B. J. Hayes. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet 50:362-367. https://doi.org/10.1038/s41588-018-0056-5.

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci 98:4107-4116. https://doi.org/10.3168/jds.2014-9005.

Browning, B. L. and S. R. Browning. 2016. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 98:116-126. https://doi.org/10.1016/j.ajhg.2015.11.020.

Bult, C. J., J. A. Blake, C. L. Smith, J. A. Kadin, J. E. Richardson, and G. Mouse Genome Database. 2019. Mouse Genome Database (MGD) 2019. Nucleic Acids Res 47:D801-D806. https://doi.org/10.1093/nar/gky1056.

Calus, M. P. L., M. E. Goddard, Y. C. J. Wientjes, P. J. Bowman, and B. J. Hayes. 2018. Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. J Dairy Sci 101:4279-4294. https://doi.org/10.3168/jds.2017-13366.

Capitan, A., A. Allais-Bonnet, A. Pinton, B. Marquant-Le Guienne, D. Le Bourhis, C. Grohs, S. Bouet, L. Clement, L. Salas-Cortes, E. Venot, S. Chaffaux, B. Weiss, A. Delpeuch, G. Noe, M. N. Rossignol, S. Barbey, D. Dozias, E. Cobo, H. Barasc, A. Auguste, M. Pannetier, M. C. Deloche, E. Lhuilier, O. Bouchez, D. Esquerre, G. Salin, C. Klopp, C. Donnadieu, C. Chantry-Darmon, H. Hayes, Y. Gallard, C. Ponsart, D. Boichard, and E. Pailhoux. 2012. A 3.7 Mb deletion encompassing ZEB2 causes a novel polled and multisystemic syndrome in the progeny of a somatic mosaic bull. PLoS One 7:e49084. https://doi.org/10.1371/journal.pone.0049084.

Charlier, C., W. Coppieters, F. Rollin, D. Desmecht, J. S. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, J. C. Frennet, R. Hanset, X. Hubin, C. Jorgensen, L. Karim, M. Kent, K. Harvey, B. R. Pearce, P. Simon, N. Tama, H. Nie, S. Vandeputte, S. Lien, M. Longeri, M. Fredholm, R. J. Harvey, and M. Georges. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet 40:449-454. https://doi.org/10.1038/ng.96.

Charlier, C., W. Li, C. Harland, M. Littlejohn, W. Coppieters, F. Creagh, S. Davis, T. Druet, P. Faux, F. Guillaume, L. Karim, M. Keehan, N. K. Kadri, N. Tamma, R. Spelman, and M. Georges. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. Genome Res 26:1333-1341. https://doi.org/10.1101/gr.207076.116.

Cole, J. B., D. J. Null, and P. M. VanRaden. 2016. Phenotypic and genetic effects of recessive haplotypes on yield, longevity, and fertility. J Dairy Sci 99:7274-7288. https://doi.org/10.3168/jds.2015-10777.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46:858-865. https://doi.org/10.1038/ng.3034.

Daetwyler, H. D., F. S. Schenkel, M. Sargolzaei, and J. A. Robinson. 2008. A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. J Dairy Sci 91:3225-3236. https://doi.org/10.3168/jds.2007-0333.

Das, S., L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P. R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, and C. Fuchsberger. 2016. Next-generation genotype imputation service and methods. Nat Genet 48:1284-1287. https://doi.org/10.1038/ng.3656.

de Los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet 9:e1003608. https://doi.org/10.1371/journal.pgen.1003608.

Delaneau, O., J. Marchini, C. Genomes Project, and C. Genomes Project. 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun 5:3934. https://doi.org/10.1038/ncomms4934.

Druet, T., S. Fritz, M. Boussaha, S. Ben-Jemaa, F. Guillaume, D. Derbala, D. Zelenika, D. Lechner, C. Charon, D. Boichard, I. G. Gut, A. Eggen, and M. Gautier. 2008. Fine mapping of quantitative trait loci affecting female fertility in dairy cattle on BTA03 using a dense single-nucleotide polymorphism map. Genetics 178:2227-2235. https://doi.org/10.1534/genetics.107.085035.

Fritz, S., A. Capitan, A. Djari, S. C. Rodriguez, A. Barbat, A. Baur, C. Grohs, B. Weiss, M. Boussaha, D. Esquerre, C. Klopp, D. Rocha, and D. Boichard. 2013. Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in GART, SHBG and SLC37A2. PLoS One 8:e65550. https://doi.org/10.1371/journal.pone.0065550.

Fritz, S., C. Hoze, E. Rebours, A. Barbat, M. Bizard, A. Chamberlain, C. Escouflaire, C. Vander Jagt, M. Boussaha, C. Grohs, A. Allais-Bonnet, M. Philippe, A. Vallee, Y. Amigues, B. J. Hayes, D. Boichard, and A. Capitan. 2018. An initiator codon mutation in SDE2 causes recessive embryonic lethality in Holstein cattle. J Dairy Sci. https://doi.org/10.3168/jds.2017-14119.

Garcia-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-Lopez, and C. P. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci U S A 113:E3995-4004. https://doi.org/10.1073/pnas.1519061113.

Genomes Project, C., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. 2015. A global reference for human genetic variation. Nature 526:68-74. https://doi.org/10.1038/nature15393.

Georges, M., C. Charlier, and B. Hayes. 2019. Harnessing genomic information for livestock improvement. Nat Rev Genet 20:135-156. https://doi.org/10.1038/s41576-018-0082-2.

Goodwin, S., J. D. McPherson, and W. R. McCombie. 2016. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333-351. https://doi.org/10.1038/nrg.2016.49.

Hay, E. H. A., Y. T. Utsunomiya, L. Xu, Y. Zhou, H. H. R. Neves, R. Carvalheiro, D. M. Bickhart, L. Ma, J. F. Garcia, and G. E. Liu. 2018. Genomic predictions combining SNP markers and copy number variations in Nellore cattle. BMC Genomics 19:441. https://doi.org/10.1186/s12864-018-4787-6.

Hayes, B. 2013. Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). Methods Mol Biol 1019:149-169. https://doi.org/10.1007/978-1-62703-447-0_6.

He, F., R. Pasam, F. Shi, S. Kant, G. Keeble-Gagnere, P. Kay, K. Forrest, A. Fritz, P. Hucl, K. Wiebe, R. Knox, R. Cuthbert, C. Pozniak, A. Akhunova, P. L. Morrell, J. P. Davies, S. R. Webb, G. Spangenberg, B. Hayes, H. Daetwyler, J. Tibbits, M. Hayden, and E. Akhunov. 2019. Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. Nat Genet 51:896-904. https://doi.org/10.1038/s41588-019-0382-2.

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

Hoze, C., C. Escouflaire, M. Mesbah-Uddin, A. Barbat, M. Boussaha, M. C. Deloche, D. Boichard, S. Fritz, and A. Capitan. 2019. Short Communication: A splice site mutation in CENPU is associated with recessive embryonic lethality in Holstein cattle. Manuscript (in review).

Jiang, Y., M. Xie, W. Chen, R. Talbot, J. F. Maddox, T. Faraut, C. Wu, D. M. Muzny, Y. Li, W. Zhang, J. A. Stanton, R. Brauning, W. C. Barris, T. Hourlier, B. L. Aken, S. M. J. Searle, D. L. Adelson, C. Bian, G. R. Cam, Y. Chen, S. Cheng, U. DeSilva, K. Dixen, Y. Dong, G. Fan, I. R. Franklin, S. Fu, R. Guan, M. A. Highland, M. E. Holder, G. Huang, A. B. Ingham,

S. N. Jhangiani, D. Kalra, C. L. Kovar, S. L. Lee, W. Liu, X. Liu, C. Lu, T. Lv, T. Mathew, S. McWilliam, M. Menzies, S. Pan, D. Robelin, B. Servin, D. Townley, W. Wang, B. Wei, S. N. White, X. Yang, C. Ye, Y. Yue, P. Zeng, Q. Zhou, J. B. Hansen, K. Kristensen, R. A. Gibbs, P. Flicek, C. C. Warkup, H. E. Jones, V. H. Oddy, F. W. Nicholas, J. C. McEwan, J. Kijas, J. Wang, K. C. Worley, A. L. Archibald, N. Cockett, X. Xu, W. Wang, and B. P. Dalrymple. 2014. The sheep genome illuminates biology of the rumen and lipid metabolism. Science 344:1168-1173. https://doi.org/10.1126/science.1252806.

Kadri, N. K., P. D. Koks, and T. H. Meuwissen. 2012. Prediction of a deletion copy number variant by a dense SNP panel. Genet Sel Evol 44:7. https://doi.org/10.1186/1297-9686-44-7.

Kadri, N. K., G. Sahana, C. Charlier, T. Iso-Touru, B. Guldbrandtsen, L. Karim, U. S. Nielsen, F. Panitz, G. P. Aamand, N. Schulman, M. Georges, J. Vilkki, M. S. Lund, and T. Druet. 2014. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. PLoS Genet 10:e1004049. https://doi.org/10.1371/journal.pgen.1004049.

Kemper, K. E., C. M. Reich, P. J. Bowman, C. J. Vander Jagt, A. J. Chamberlain, B. A. Mason, B. J. Hayes, and M. E. Goddard. 2015. Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. Genet Sel Evol 47:29. https://doi.org/10.1186/s12711-014-0074-4.

Kipp, S., D. Segelke, S. Schierenbeck, F. Reinhardt, R. Reents, C. Wurmser, H. Pausch, R. Fries, G. Thaller, J. Tetens, J. Pott, D. Haas, B. B. Raddatz, M. Hewicker-Trautwein, I. Proios, M. Schmicke, and W. Grunberg. 2016. Identification of a haplotype associated with cholesterol deficiency and increased juvenile mortality in Holstein cattle. J Dairy Sci 99:8915-8931. https://doi.org/10.3168/jds.2016-11118.

Kolbehdari, D., Z. Wang, J. R. Grant, B. Murdoch, A. Prasad, Z. Xiu, E. Marques, P. Stothard, and S. S. Moore. 2008. A whole-genome scan to map quantitative trait loci for conformation and functional traits in Canadian Holstein bulls. J Dairy Sci 91:2844-2856. https://doi.org/10.3168/jds.2007-0585.

Li, M., S. Tian, L. Jin, G. Zhou, Y. Li, Y. Zhang, T. Wang, C. K. Yeung, L. Chen, J. Ma, J. Zhang, A. Jiang, J. Li, C. Zhou, J. Zhang, Y. Liu, X. Sun, H. Zhao, Z. Niu, P. Lou, L. Xian, X. Shen, S. Liu, S. Zhang, M. Zhang, L. Zhu, S. Shuai, L. Bai, G. Tang, H. Liu, Y. Jiang, M. Mai, J. Xiao, X. Wang, Q. Zhou, Z. Wang, P. Stothard, M. Xue, X. Gao, Z. Luo, Y. Gu, H. Zhu, X. Hu, Y. Zhao, G. S. Plastow, J. Wang, Z. Jiang, K. Li, N. Li, X. Li, and R. Li. 2013. Genomic analyses identify distinct patterns of selection in domesticated pigs and Tibetan wild boars. Nat Genet 45:1431-1438. https://doi.org/10.1038/ng.2811.

Li, W., A. Sartelet, N. Tamma, W. Coppieters, M. Georges, and C. Charlier. 2016. Reverse genetic screen for loss-of-function mutations uncovers a frameshifting deletion in the melanophilin gene accountable for a distinctive coat color in Belgian Blue cattle. Anim Genet 47:110-113. https://doi.org/10.1111/age.12383.

Li, Y., N. T. Klena, G. C. Gabriel, X. Liu, A. J. Kim, K. Lemke, Y. Chen, B. Chatterjee, W. Devine, R. R. Damerla, C. Chang, H. Yagi, J. T. San Agustin, M. Thahir, S. Anderton, C. Lawhead, A. Vescovi, H. Pratt, J. Morgan, L. Haynes, C. L. Smith, J. T. Eppig, L. Reinholdt, R. Francis, L. Leatherbury, M. K. Ganapathiraju, K. Tobita, G. J. Pazour, and C. W. Lo. 2015. Global genetic analysis in mice unveils central role for cilia in congenital heart disease. Nature 521:520-524. https://doi.org/10.1038/nature14269.

Lund, M. S., A. P. Roos, A. G. Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol 43:43. https://doi.org/10.1186/1297-9686-43-43.

MacArthur, D. G., S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M. M. Suner, T. Hunt, I. H. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, C. Genomes Project, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335:823-828. https://doi.org/10.1126/science.1215040.

Marchini, J. and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499-511. https://doi.org/10.1038/nrg2796.

Marchini, J., B. Howie, S. Myers, G. McVean, and P. Donnelly. 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906-913. https://doi.org/10.1038/ng2088.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356-369. https://doi.org/10.1038/nrg2344.

McClure, M. C., D. Bickhart, D. Null, P. Vanraden, L. Xu, G. Wiggans, G. Liu, S. Schroeder, J. Glasscock, J. Armstrong, J. B. Cole, C. P. Van Tassell, and T. S. Sonstegard. 2014. Bovine exome sequence analysis and targeted SNP genotyping of recessive fertility defects BH1, HH2, and HH3 reveal a putative causative mutation in SMC2 for HH3. PLoS One 9:e92769. https://doi.org/10.1371/journal.pone.0092769.

Mesbah-Uddin, M., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2018. Genotype call for chromosomal deletions using read-depth from whole genome sequence variants in cattle. Page 662 in Proc. World Congr. Genet. Appl. Livest. Prod., Auckland, New Zealand. AL Rae Centre for Genetics and Breeding, Massey University, Palmerston North, New Zealand. http://www.wcgalp.org/system/files/proceedings/2018/genotype-call-chromosomal-deletions-using-read-depth-whole-genome-sequence-variants-cattle.pdf.

Mesbah-Uddin, M., C. Hozé, P. Michot, A. Barbat, R. Lefebvre, M. Boussaha, G. Sahana, S. Fritz, D. Boichard, and A. Capitan. 2019. A missense mutation (p.Tyr452Cys) in the CAD gene compromises reproductive success in French Normande cattle. J Dairy Sci 102:6340-6356. https://doi.org/10.3168/jds.2018-16100.

Meuwissen, T. and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623-631. https://doi.org/10.1534/genetics.110.116590.

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

Michot, P., S. Chahory, A. Marete, C. Grohs, D. Dagios, E. Donzel, A. Aboukadiri, M. C. Deloche, A. Allais-Bonnet, M. Chambrial, S. Barbey, L. Genestout, M. Boussaha, C. Danchin-Burge, S. Fritz, D. Boichard, and A. Capitan. 2016. A reverse genetic approach identifies an ancestral frameshift mutation in RP1 causing recessive progressive retinal degeneration in European cattle breeds. Genet Sel Evol 48:56. https://doi.org/10.1186/s12711-016-0232-y.

Michot, P., S. Fritz, A. Barbat, M. Boussaha, M. C. Deloche, C. Grohs, C. Hoze, L. Le Berre, D. Le Bourhis, O. Desnoes, P. Salvetti, L. Schibler, D. Boichard, and A. Capitan. 2017. A missense mutation in PFAS (phosphoribosylformylglycinamidine synthase) is likely causal for embryonic lethality associated with the MH1 haplotype in Montbeliarde dairy cattle. J Dairy Sci 100:8176-8187. https://doi.org/10.3168/jds.2017-12579.

Moser, G., S. H. Lee, B. J. Hayes, M. E. Goddard, N. R. Wray, and P. M. Visscher. 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet 11:e1004969. https://doi.org/10.1371/journal.pgen.1004969.

Pausch, H., R. Emmerling, H. Schwarzenbacher, and R. Fries. 2016. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. Genet Sel Evol 48:14. https://doi.org/10.1186/s12711-016-0190-4.

Pausch, H., H. Schwarzenbacher, J. Burgstaller, K. Flisikowski, C. Wurmser, S. Jansen, S. Jung, A. Schnieke, T. Wittek, and R. Fries. 2015. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. BMC Genomics 16:312. https://doi.org/10.1186/s12864-015-1483-7.

Pont, C., T. Leroy, M. Seidel, A. Tondelli, W. Duchemin, D. Armisen, D. Lang, D. Bustos-Korts, N. Goue, F. Balfourier, M. Molnar-Lang, J. Lage, B. Kilian, H. Ozkan, D. Waite, S. Dyer, T. Letellier, M. Alaux, Wheat, c. Barley Legacy for Breeding Improvement, J. Russell, B. Keller, F. van Eeuwijk, M. Spannagl, K. F. X. Mayer, R. Waugh, N. Stein, L. Cattivelli, G. Haberer, G. Charmet, and J. Salse. 2019. Tracing the ancestry of modern bread wheats. Nat Genet 51:905-911. https://doi.org/10.1038/s41588-019-0393-z.

Pryce, J. E., R. F. Veerkamp, R. Thompson, W. G. Hill, and G. Simm. 1997. Genetic aspects of common health disorders and measures of fertility in Holstein Friesian dairy cattle. Animal Science 65:353-360. https://doi.org/10.1017/S1357729800008559.

Rauw, W. M., E. Kanis, E. N. Noordhuizen-Stassen, and F. J. Grommers. 1998. Undesirable side effects of selection for high production efficiency in farm animals: a review. Livestock Production Science 56:15-33. https://doi.org/https://doi.org/10.1016/S0301-6226(98)00147-X.

Rivas, M. A., M. Pirinen, D. F. Conrad, M. Lek, E. K. Tsang, K. J. Karczewski, J. B. Maller, K. R. Kukurba, D. S. DeLuca, M. Fromer, P. G. Ferreira, K. S. Smith, R. Zhang, F. Zhao, E. Banks, R. Poplin, D. M. Ruderfer, S. M. Purcell, T. Tukiainen, E. V. Minikel, P. D. Stenson, D. N. Cooper, K. H. Huang, T. J. Sullivan, J. Nedzel, G. T. Consortium, C. Geuvadis, C. D. Bustamante, J. B. Li, M. J. Daly, R. Guigo, P. Donnelly, K. Ardlie, M. Sammeth, E. T. Dermitzakis, M. I. McCarthy, S. B. Montgomery, T. Lappalainen, and D. G. MacArthur. 2015. Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science 348:666-669. https://doi.org/10.1126/science.1261877.

Roxström, A., E. Strandberg, B. Berglund, U. Emanuelson, and J. Philipsson. 2001. Genetic and Environmental Correlations Among Female Fertility Traits and Milk Production in Different Parities of Swedish Red and White Dairy Cattle. Acta Agriculturae Scandinavica, Section A — Animal Science 51:7-14. https://doi.org/10.1080/090647001300004745.

Sahana, G., T. Iso-Touru, X. Wu, U. S. Nielsen, D. J. de Koning, M. S. Lund, J. Vilkki, and B. Guldbrandtsen. 2016. A 0.5-Mbp deletion on bovine chromosome 23 is a strong candidate for stillbirth in Nordic Red cattle. Genet Sel Evol 48:35. https://doi.org/10.1186/s12711-016-0215-z.

Sahana, G., U. S. Nielsen, G. P. Aamand, M. S. Lund, and B. Guldbrandtsen. 2013. Novel harmful recessive haplotypes identified for fertility traits in Nordic Holstein cattle. PLoS One 8:e82909. https://doi.org/10.1371/journal.pone.0082909.

Sanchez, M. P., A. Govignon-Gion, P. Croiseau, S. Fritz, C. Hoze, G. Miranda, P. Martin, A. Barbat-Leterrier, R. Letaief, D. Rocha, M. Brochard, M. Boussaha, and D. Boichard. 2017. Within-breed and multi-breed GWAS on imputed whole-genome sequence variants reveal candidate mutations affecting milk protein composition in dairy cattle. Genet Sel Evol 49:68. https://doi.org/10.1186/s12711-017-0344-z.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. https://doi.org/10.1186/1471-2164-15-478.

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

Schulman, N. F., G. Sahana, T. Iso-Touru, S. D. McKay, R. D. Schnabel, M. S. Lund, J. F. Taylor, J. Virta, and J. H. Vilkki. 2011. Mapping of fertility traits in Finnish Ayrshire by genome-wide association analysis. Anim Genet 42:263-269. https://doi.org/10.1111/j.1365-2052.2010.02149.x.

Schwarzenbacher, H., J. Burgstaller, F. R. Seefried, C. Wurmser, M. Hilbe, S. Jung, C. Fuerst, N. Dinhopl, H. Weissenbock, B. Fuerst-Waltl, M. Dolezal, R. Winkler, O. Grueter, U. Bleul, T. Wittek, R. Fries, and H. Pausch. 2016. A missense mutation in TUBD1 is associated with high juvenile mortality in Braunvieh and Fleckvieh cattle. BMC Genomics 17:400. https://doi.org/10.1186/s12864-016-2742-y.

Slatkin, M. 2008. Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. Nat Rev Genet 9:477-485. https://doi.org/10.1038/nrg2361.

Sørensen, P., S. M. Edwards, and P. Jensen. 2014. Genomic Feature Models. in Proc. 10th World Congress of Genetics Applied to Livestock Production, Vancouver, BC, Canada.

van den Berg, I., D. Boichard, B. Guldbrandtsen, and M. S. Lund. 2016. Using Sequence Variants in Linkage Disequilibrium with Causative Mutations to Improve Across-Breed Prediction in Dairy Cattle: A Simulation Study. G3 (Bethesda) 6:2553-2561. https://doi.org/10.1534/g3.116.027730.

van den Berg, I., T. H. E. Meuwissen, I. M. MacLeod, and M. E. Goddard. 2019. Predicting the effect of reference population on the accuracy of within, across, and multibreed genomic prediction. J Dairy Sci. https://doi.org/10.3168/jds.2018-15231.

van Leeuwen, E. M., A. Kanterakis, P. Deelen, M. V. Kattenberg, C. Genome of the Netherlands, P. E. Slagboom, P. I. de Bakker, C. Wijmenga, M. A. Swertz, D. I. Boomsma, C. M. van Duijn, L. C. Karssen, and J. J. Hottenga. 2015. Population-specific genotype imputations using minimac or IMPUTE2. Nat Protoc 10:1285-1296. https://doi.org/10.1038/nprot.2015.077.

VanRaden, P. M., K. M. Olson, D. J. Null, and J. L. Hutchison. 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J Dairy Sci 94:6153-6161. https://doi.org/10.3168/jds.2011-4624.

VanRaden, P. M., M. E. Tooker, J. R. O'Connell, J. B. Cole, and D. M. Bickhart. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol 49:32. https://doi.org/10.1186/s12711-017-0307-4.

Willer, C. J., Y. Li, and G. R. Abecasis. 2010. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26:2190-2191. https://doi.org/10.1093/bioinformatics/btq340.

Wong, G. K. and B. Liu and J. Wang and Y. Zhang and X. Yang and Z. Zhang and Q. Meng and J. Zhou and D. Li and J. Zhang and P. Ni and S. Li and L. Ran and H. Li and J. Zhang and R. Li and S. Li and H. Zheng and W. Lin and G. Li and X. Wang and W. Zhao and J. Li and C. Ye and M. Dai and J. Ruan and Y. Zhou and Y. Li and X. He and Y. Zhang and J. Wang and X. Huang and W. Tong and J. Chen and J. Ye and C. Chen and N. Wei and G. Li and L. Dong and F. Lan and Y. Sun and Z. Zhang and Z. Yang and Y. Yu and Y. Huang and D. He and Y. Xi and D. Wei and Q. Qi and W. Li and J. Shi and M. Wang and F. Xie and J. Wang and X. Zhang and P. Wang and Y. Zhao and N. Li and N. Yang and W. Dong and S. Hu and C. Zeng and W. Zheng and B. Hao and L. W. Hillier and S. P. Yang and W. C. Warren and R. K. Wilson and M. Brandstrom and H. Ellegren and R. P. Crooijmans and J. J. van der Poel and H. Bovenhuis and M. A. Groenen and I. Ovcharenko and L. Gordon and L. Stubbs and S. Lucas and T. Glavina and A. Aerts and P. Kaiser and L. Rothwell and J. R. Young and S. Rogers and B. A. Walker and A. van Hateren and J. Kaufman and N. Bumstead and S. J. Lamont and H. Zhou and P. M. Hocking and D. Morrice and D. J. de Koning and A. Law and N. Bartley and D. W. Burt and H. Hunt and H. H. Cheng and U. Gunnarsson and P. Wahlberg and L. Andersson and E. Kindlund and M. T. Tammi and B. Andersson and C. Webber and C. P. Ponting and I. M. Overton and P. E. Boardman and H. Tang and S. J. Hubbard and S. A. Wilson and J. Yu and J. Wang and H. Yang and C. International Chicken Polymorphism Map. 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432:717-722. https://doi.org/10.1038/nature03156.

Zhang, Z., E. Ersoz, C. Q. Lai, R. J. Todhunter, H. K. Tiwari, M. A. Gore, P. J. Bradbury, J. Yu, D. K. Arnett, J. M. Ordovas, and E. S. Buckler. 2010. Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355-360. https://doi.org/10.1038/ng.546.

Zhao, Q., Q. Feng, H. Lu, Y. Li, A. Wang, Q. Tian, Q. Zhan, Y. Lu, L. Zhang, T. Huang, Y. Wang, D. Fan, Y. Zhao, Z. Wang, C. Zhou, J. Chen, C. Zhu, W. Li, Q. Weng, Q. Xu, Z. X. Wang, X. Wei, B. Han, and X. Huang. 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat Genet 50:278-284. https://doi.org/10.1038/s41588-018-0041-z.

Chapter 2.

A missense mutation (p.Tyr452Cys) in the CAD gene compromises reproductive success in French Normande cattle

Md Mesbah-Uddin,[1,2*] Chris Hoze,[2,3] Pauline Michot,[2,3] Anne Barbat,[2] Rachel Lefebvre,[2] Mekki Boussaha,[2] Goutam Sahana,[1] Sébastien Fritz,[2,3] Didier Boichard,[2] and Aurélien Capitan[2,3*]

[1]Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark

[2]GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[3]Allice, 75595 Paris, France

[*]Corresponding authors: mdmesbah@gmail.com and aurelien.capitan@inra.fr

Journal of Dairy Science, 2019. 102(7): p. 6340-6356.

Supplementary materials are available at https://doi.org/10.3168/jds.2018-16100.

Note: Chapter 2 is identical to the paper except for page layout and journal amended text formatting and editing.

2.1 Abstract

We scanned the genome of 77,815 Normande cattle with different Illumina SNP-chips to map recessive embryonic lethal mutations using homozygous haplotype deficiency (HHD). We detected two novel haplotypes on chromosomes 11 and 24, but did not confirm six previously reported haplotypes. The one on chromosome 11 showed a marked reduction in conception rates and moderate decrease in non-return rate in at-risk versus control mating, supporting late embryonic mortality. After fine-mapping and analyzing whole-genome sequences we prioritized a missense mutation in *CAD* (g.72399397T>C; p.Tyr452Cys)—a gene encoding a protein essential for *de novo* pyrimidine biosynthesis—as a candidate causal variant. This transition mutation replaces a tyrosine residue, which is perfectly conserved among living organisms, by a cysteine residue in the carbamoyl-phosphate synthetase 2 domain of the protein. A single animal was confirmed to be homozygous for the mutation based on Sanger sequencing. However, large-scale genotyping of the candidate variant with Illumina EuroG10k Beadchip revealed an absence of live homozygote in a panel of 33,323 Normande animals and an absence of carriers in 348,593 animals from 19 other cattle breeds. These results support recessive embryonic lethality with nearly complete penetrance, as was previously reported in *CAD* mutants in several Eukaryote species. The only homozygous cow had an extremely poor udder conformation, suggesting a potential role of *CAD* in udder development, but no effect was detected when comparing daughter yield deviations of 250 heterozygous bulls with that of 2,912 homozygotes for the ancestral allele. Together, this study showed the importance of large-scale screening for HHD with hundreds of thousands of animals, validating results with independent dataset, and considering unexpected live homozygotes, to avoid both false positive and false negative discoveries. These discoveries will be used primarily in mating decisions to avoid at-risk mating. In addition, we recommend including *CAD* in the breeding objectives of Normande cattle.

Key words: embryonic lethal, CAD, homozygous haplotype deficiency, dairy cattle, large-scale genotyping

2.2 Introduction

Inbreeding and genetic drift has increased substantially in dairy cattle over the last decades due to the extensive use of a few elite artificial insemination sires (Weigel, 2001, Bjelland et al., 2013). This situation can increase the frequency of some deleterious alleles in the population, and thus unmask recessive genetic defects affecting various fitness traits (Cole et al., 2016).

Homozygosity mapping is an efficient approach to map such recessive deleterious alleles with few inbred live cases (Charlier et al., 2008). However, it requires the prior detection of live animals showing distinctive features, which is not possible for all types of genetic defects (e.g. when homozygotes die very early in the gestation period or when they display symptoms that can be confounded with non-genetic diseases). In 2011, VanRaden et al. proposed a new approach for identifying recessive genetic defects from large-scale screening of homozygous haplotype deficiency (HHD). Embryonic lethality or increased perinatal mortality is suspected when observed homozygotes are significantly less frequent than expected based on population

frequency. This approach enabled the detection of a number of haplotypes with both complete and incomplete penetrance in several breeds (Sahana et al., 2013, Pausch et al., 2015, Schwarzenbacher et al., 2016a), though it required tens or hundreds of thousands of genotyped individuals for rarer disorders.

However, some of these haplotypes could not be confirmed in independent studies, possibly due to false positive results or population structure within a breed. For example, haplotypes HH2 in Holstein and BH1 in Brown Swiss, showed significant HHD in the USA (VanRaden et al., 2011) but not in Europe (Fritz et al., 2013, Segelke et al., 2016, Fritz et al., 2018). For this reason, and since no phenotypic association and candidate mutation have been reported so far (McClure et al., 2014), they have been recently removed from the list of genetic defects considered by the National Association of Animal Breeders (NAAB) of the USA (Olivier Bulot, personal communication).

Previous studies also highlighted the importance of large-scale genotyping to validate candidate causal mutations (Schwarzenbacher et al., 2016a, Schwarzenbacher et al., 2016b, Michot et al., 2017, Fritz et al., 2018). Doing so, our team recently excluded a candidate variant in the *SHBG* gene for MH1 haplotype in Montbeliarde cattle (Fritz et al., 2013) and validated a new causal variant in the *PFAS* gene (Michot et al., 2017).

In 2013, Fritz et al. reported six haplotypes (NH1 to NH6) showing significant HHD in 11,466 Normande cattle, including four also showing some reduction in conception rate in at-risk versus control mating. In this large-scale screen, we were interested in validating these haplotypes in Normande cattle, detecting novel loci, and when possible, identifying causal variants.

2.3 Methods

Study Samples and Genotype Data

We analyzed 77,815 Normande cattle from the French genomic evaluation database, genotyped between the year 2008 and 2018 with different Illumina SNP-chips (Illumina Inc., San Diego, CA, USA), such as, 777K BovineHD BeadChip, 50K BovineSNP50 BeadChip (Matukumalli et al., 2009), BovineLD BeadChip (Boichard et al., 2012a), and EuroG10K BeadChip (Boichard et al., 2018). In the total dataset, 27,632 animals had both parents genotyped, and 49,690 had sire and maternal grandsire (MGS) genotyped. The others, i.e. 353 and 140 animals with no or a single parent genotyped, respectively, were included as sires or MGS only. Since this dataset included all the animals studied by Fritz et al. (2013), we also performed a specific analysis on this subpopulation. This dataset comprised of 2,303 animals with both parents genotyped, and 9,163 animals with sire and MGS (but not dam) genotyped, as well as 111 sires or MGS with no or a single parent genotyped.

The quality control of the genotype data was carried out following the French genomic evaluation pipeline. This includes: (i) filtering of low-quality genotypes, (ii) verifying the pedigree for Mendelian inconsistency,

and (iii) phasing and imputing genotypes (Sargolzaei et al., 2014) from low to medium-density (for details see Boichard et al. (2012b)). After quality control, 43,801 autosomal SNPs from the BovineSNP50 BeadChip were kept for subsequent analysis. The genomic coordinates presented in this study refer to bovine genome assembly UMD3.1 (Zimin et al., 2009).

Screening for Homozygous Haplotype Deficiency

Following the method used by VanRaden et al. (2011) and Fritz et al. (2013), we screened the 29 bovine autosomes, with sliding windows of 20 consecutive markers, for homozygous haplotype deficiency (HHD). We counted occurrences of every haplotypes, and calculated population frequency from the maternal chromosomal phases. In addition, we estimated the expected number of homozygotes using within-family transmission probability, such as, from carrier sires and dams to offspring, or from carrier sires and MGS (with un-genotyped dam) to offspring. For subsequent analyses, we considered haplotypes with population frequency greater than 1.0%, representing on average 17 different haplotypes for each 20-marker window (range 4 to 34 haplotypes). In total, we performed 717,747 tests to identify HHD, and therefore considered haplotypes with $P < 1.39 \times 10^{-8}$ for further analysis, after correcting for multiple testing at 1% level of significance. The expected number of homozygotes (*lambda*) assuming neutrality was computed from sire, dam or MGS genotypes and frequency in the population as in Fritz et al. (2013). We calculated the probability of observing "*q*" homozygotes with expectation of "*lambda*", using Poisson distribution "*ppois(q, lambda)*" function in R software version 3.4.3 (R Core Team, 2017). After multiple testing correction, we considered haplotypes for further analysis when the ratio of observed to expected homozygotes was less than 0.25.

Comparison of Haplotype Status for NH1-NH6 between the 2013 and Present Studies

Haplotype status as computed by Fritz et al. (2013) were available for 2,364 AI bulls out of the 11,466 animals initially studied. We compared carrier status between the two studies for these animals.

Analysis of the Survival of Animals Homozygous for NH2

Information on the date of birth, date of death and cause (slaughter or natural death) was extracted from the French genomic evaluation database for 228 females homozygous for NH2 and 8,871 non-homozygous paternal half-sisters. In both groups, we evaluated the proportion of animals that died of natural causes at one, two and three years of age.

Evaluation of Haplotype Effect on Fertility Traits

We evaluated the haplotype effect on two 'binary' female fertility traits, conception rate (CR) and non-return rate at 56 days (NRR56), separately in heifers and lactating cows, using phenotypic records from the French genomic evaluation database (period January 2000 to May 2018). The CR is a measure of success ('1') or

failure ('0') of each insemination, assessed by a calving after a compatible gestation length. NRR56 measures whether a female is re-inseminated within 56 days after a previous insemination. It is coded '0' if at least one insemination is recorded within this period and '1' otherwise. As both phenotypes are routinely evaluated in the French national evaluation system, non-genetic effects estimates (such as, effects of herd × year, year x month of insemination, parity, days in milk, AI technician, use of sexed semen, etc.) were available and were used to pre-adjust the phenotypes of these two traits. The mating class was coded '1' for at-risk mating (between carrier bulls and daughters of carrier bulls), and '0' for non-risk mating (all other combinations of genotypes regarding the bulls and the sires of the dams). We analyzed the phenotypic effect of the haplotype between at-risk mating vs non-risk mating using the following fixed effect model:

$$Y_{ijk} = \mu + sire_j + mating_class_k + e_{ijk}$$

where, Y_{ijk} represents the phenotype of interest (i.e. adjusted CR or NRR56), μ is the overall phenotypic mean, $sire_j$ is the fixed sire effect, $mating\text{-}class_k$ is the fixed effect of the mating status, and e_{ijk} is the random residual error. The analysis was performed using the GLM procedure of SAS software v9.4 '*Proc GLM*' (SAS Institute Inc., NC, USA). In addition, we calculated the expected decrease in fertility, assuming full lethality under homozygosity, using the formula: $\frac{1}{4}(\frac{1}{2-f_{hap_k}})\mu$, where f_{hap_k} is the frequency of the $haplotype_k$, and μ is the phenotypic mean for CR or NRR56 (as in Michot et al. (2017)).

Fine-Mapping of the Normande Haplotype 7 (NH7) Locus

To fine map the causal mutation associated with the newly detected NH7 haplotype, we followed the approach used by Sonstegard et al. (2013) and Fritz et al. (2018). Briefly, this consisted in identifying phenotypically normal animals presenting a run-of-homozygosity (ROH) on either side of the candidate haplotype among the inbred descendant of the bull ("Nonnic", NORFRAM002977029292), the most ancient and influential carrier of NH7 in our dataset. Such asymptomatic individuals are very unlikely to be homozygous for the causal mutation, and thus regions with long ROH can be ruled-out for causal variant screening.

Analyzing Whole-Genome Sequence (WGS) Data for Candidate Mutations

We analyzed WGS of 2,333 animals (44 Normande cattle and 2,289 animals from other cattle breeds) from the 1000 Bull Genomes Project (*Run6*; 1KBGP) (Daetwyler et al., 2014) for identifying putative causal variant. Among the WGS animals, only two Normande bulls carried NH7 haplotype and the rest did not. We performed Pearson's correlation between haplotype status and allele dosage for all bi-allelic single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) within NH7 (±500 kb), to identify highly associated variants. Both, haplotype status and allele dosage were coded as '0', '1', or '2', when the animal carried either zero, one or two copies of alternative allele/candidate haplotype, respectively. We presented squared values of Pearson's correlation (r^2) in a scatterplot.

We also analyzed structural variant (SV) calls on 389 whole-genome sequenced French cattle, which included 44 WGS Normande cattle and 345 animals from other breeds. The SVs were identified using three different software, namely BreakDancer v1.3.6 (Chen et al., 2009), Delly v0.6.1 (Rausch et al., 2012), and Pindel v2.5 (Ye et al., 2009), and consensus SV calls from these tools were included for the analysis (for details see Letaief et al. (2017)). For candidate SVs, we focused our search within NH7 and ±500 kb surrounding regions. We checked for the concordance between the carrier status of NH7 and SVs.

Integrative Annotation of Candidate Mutations and Nearby Genes

We annotated all strongly correlated variants ($r > 0.8$) using Variant Effect Predictor (VEP v87) software (McLaren et al., 2016), which includes annotation of sequence ontology, SIFT score (Kumar et al., 2009) for missense variants, overlapping gene name, etc. We retrieved three gene-level constraint scores: (i) the ratio of non-synonymous (d_N) to synonymous (d_S) substitution rate (*dN/dS* scores) for cow-human 1-to-1 orthologs, using BioMart (Kinsella et al., 2011, Zerbino et al., 2018), (ii) Residual Variation Intolerance Score (RVIS) (Petrovski et al., 2013) and (iii) missense *Z-score* (Lek et al., 2016) for human orthologs of cattle genes. We converted genomic coordinates of variants from "UMD3.1/bosTau6" to "GRCh37/hg19", for retrieving variant-level conservation scores using LiftOver software (http://genome.ucsc.edu/cgi-bin/hgLiftOver) (Casper et al., 2018). We retrieved evolutionary conservation statistics, such as, Genomic Evolutionary Rate Profiling (GERP++) "rejected substitutions" scores for 35-mammalian alignments (Davydov et al., 2010), PhyloP and PhastCons scores for 100-vertebrate alignments (Pollard et al., 2010), using UCSC Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) (Karolchik et al., 2004).

Multiple-Sequence Alignment

We performed multiple-sequence alignment among 24 different species, representing six groups, e.g. vertebrates, metazoan, plants, fungi, protists, and bacteria. We extracted protein sequences from either Ensembl Genomes (Kersey et al., 2018) or UniProt (The UniProt, 2017). We used CLUSTAL Omega version-1.2.4 (Sievers et al., 2011) for multiple-sequence alignment. For the candidate gene, we also retrieved functional annotation of protein domains from NCBI's Conserved Domain Database (CDD) (Marchler-Bauer et al., 2017) [accessible from "https://www.ncbi.nlm.nih.gov/cdd"].

Large-Scale Genotyping of the CAD Candidate Mutation in French Cattle Breeds

A molecular test was designed for genotyping a candidate missense mutation in the *CAD* gene (g.72399397A>G on Chromosome 11, *CAD* p.Tyr452Cys) with the EuroG10K BeadChip (Boichard et al., 2018) routinely used for genomic evaluation. Genotypes of 381,916 animals, from 20 French cattle breeds, were available for this variant.

Genotyping of the CAD Mutation by PCR and Sanger Sequencing

We used PCR and Sanger sequencing to genotype the same mutation in a single individual found to be homozygous for the NH7 haplotype based on 50K genotypes, and a non-carrier animal. DNA was obtained from LABOGENA DNA (Jouy-en-Josas, France). We designed primers using Primer3 software (Untergasser et al., 2012). The forward and reverse primers are 5′-AGGGGATCACAGAATTGGCA-3′ and 5′-GCCTCTGTCTCCTTGGGATT-3′, respectively. We performed PCR using "GoTaq Flexi DNA Polymerase" (Promega, Madison, WI, USA) in Mastercycler Pro (Eppendorf, Montesson, France) thermocycler, following manufacturer's instructions and sent amplicons to Eurofins MWG (Hilden, Germany) for sequencing. We next used novoSNP software (Weckx et al., 2005) for variant calling and visualization.

Phenotypic Examination of the Homozygous Cow and Paternal Half-Sisters

Clinical examination of the NH7 homozygote cow, named "Genetique", was not possible since this animal was slaughtered at the beginning of her first lactation due to poor udder conformation (personal communication from the breeder). However, own performance records for 34 morphological traits were available for this cow and its 2,239 paternal half-sisters from the French National databases (see Supplementary Figure S1). Individual performances were ranked for each trait and the position of "Genetique" was expressed in term of percentile.

Effect of the CAD Mutation on Udder Conformation Traits

Since "Genetique" displayed extremely poor udder conformation, we evaluated the effect of the *CAD* missense mutation, on four udder traits, namely, teat direction (TD), teat placement (TP), udder cleft (UC), and udder depth (UD). These traits are initially recorded in a scale of 1 (low) to 9 (high) (see ICAR (2018) for details), before being pre-adjusted for non-genetic factors and mean-centered in the French national evaluation system. Daughter yield deviations (DYD) were available for 250 heterozygous carriers and 2,912 non-carriers of the *CAD* mutation. The phenotypic effect of the mutation was estimated using the following fixed effect model:

$$Y_{ijk} = \mu + sire_j + CAD_genotype_k + e_{ijk}$$

Here, Y_{ijk} represents the phenotype of interest (i.e. DYD for TD, TP, UC, or UD), μ is the overall phenotypic mean, $sire_j$ is the fixed effect of the sire of the bull, CAD_genotype$_k$ is the fixed effect of the genotype (k = "AG" or "AA"), and e_{ijk} is the random residual error. The analysis was performed using the GLM procedure of SAS software v9.4 (*Proc GLM)*' (SAS Institute Inc., NC, USA).

2.4 Results and Discussion

Homozygous Haplotype Deficiency in Normande Cattle

In a sliding window approach, we analyzed 77,815 Normande cattle phased and imputed for 43,801 autosomal markers, to identify homozygous haplotype deficiency (HHD). Surprisingly, none of the six Normande haplotypes (NH1 to NH6) previously reported by Fritz et al. (2013) showed significant HHD in this follow-up study after correction for multiple testing (Table 2.1).

To elucidate the causes of these discrepancies, we reanalyzed the mapping population of 11,466 Normande cattle from Fritz et al. (2013) using the current French phasing and imputation dataset. Contrary to Fritz et al. (2013), the current dataset do not comprise the markers of the Illumina BovineSNP50v1 BeadChip, which have been removed from the subsequent versions of the chip for technical reasons. This concerns markers ARS-BFGL-NGS-97976 and BTA-58084-no-rs within NH1, Hapmap59512-rs29027076 within NH3, BTB-01218649 within NH5 and BTA-37438-no-rs and ARS-BFGL-NGS-29141 within NH6 haplotype. The removal of these problematic markers, which likely caused local phasing and imputation errors, explains the substantial changes in haplotype frequencies (Table 2.1) and carrier status (Table S2) observed between the two studies. Insufficient correction for multiple testing (arbitrary threshold of $P < 1 \times 10^{-4}$ instead of Bonferroni correction) was also another reason for the false discovery. Among the six haplotypes, only NH2 was significant for HHD after Bonferroni correction (Table S1).

Table 2.1. List of Normande haplotypes with deficit of homozygotes (haplotypes NH1-NH6 from Fritz et al. (2013); NH7-NH8 from the present study)

Haplotype ID	Chromosome	Start Position[1]	End Position	Frequency (%)	No. of Homozygotes		
					Expected	Observed	*P-value*[†]
NH1[*]	24	38,086,180	39,153,166	1.9 (1.8)[2]	9 (12)	6 (0)	0.21 (5.3×10^{-4})
NH2	1	145,682,206	146,833,973	6.3 (3.8)	394 (49)	314 (14)	1.7×10^{-5} (5.7×10^{-7})
NH3[*]	4	92,261,682	93,828,702	4.6 (5.9)	189 (41)	159 (10)	0.01 (1.3×10^{-6})
NH4	6	37,704,254	38,869,785	4.0 (5.2)	172 (38)	108 (12)	1.08×10^{-7} (2.5×10^{-5})
NH5[*]	7	3,661,458	4,592,227	18.0 (1.9)	2,764 (58)	2,735 (20)	0.29 (6×10^{-7})
NH6[*]	15	59,789,154	61,044,897	12.7 (1.9)	1,525 (45)	1,503 (17)	0.29 (3×10^{-5})
NH7	11	70,750,209	72,476,622	3.7	57	1	1.0×10^{-23}
NH8	24	52,606,336	54,542,635	1.9	40	7	1.7×10^{-10}

[1]Start and End positions represent first and last 50K-marker of the haplotype; positions correspond to bovine genome assembly UMD3.1

[2]Information within the parenthesis and Normande haplotype NH1 to NH6 are from Fritz et al. (2013)

[†]*P-values* are from Poisson distribution, calculated using R software's "ppois(q=Observed, lambda=Expected)" function

[*]Genotypes for markers ARS-BFGL-NGS-97976 and BTA-58084-no-rs within NH1, Hapmap59512-rs29027076 within NH3, BTB-01218649 within NH5 and BTA-37438-no-rs and ARS-BFGL-NGS-29141 within NH6 haplotype were considered in Fritz et al. (2013) but not in the present study.

Finally, it is worth mentioning that the population analyzed by Fritz et al. (2013) consisted in animals aged one year or more whereas the present dataset includes numerous young animals genotyped in their first months of life, which is expected to reduce the power of detection when mapping loci responsible for juvenile

mortality. To verify if NH2 is associated with postnatal mortality, and potentially explain why it is not detected in the present study, we analyzed the survival of 228 homozygous females and 8,871 non-homozygous paternal half-sisters. We found no difference between the two groups at one, two and three years of age (Supplementary Table S3). In this context, we assume that NH2 was a false positive due to a lower power of detection in the analysis of Fritz et al. (2013) as compared with that of the current study. Indeed the present dataset is sevenfold larger than the previous one, and the frequency of NH2 more than doubled in the population, which increased the probability of observing homozygotes.

Importantly, we detected two novel haplotypes with significant deficit of homozygotes ($P < 1.39\times10^{-8}$). One, named NH7, is located on chromosome 11 between positions 70,750,209 and 72,476,622 bp, and the other, named NH8, on chromosome 24 between positions 52,606,336 and 54,542,635 bp. Haplotype frequencies in Normande breed were 3.7% and 1.9% for NH7 and NH8, respectively. In our dataset, we observed one homozygote for NH7 and seven homozygotes for NH8, while the expectations under neutrality were 57 and 40, respectively (Table 2.1). The oldest carriers in our dataset were born in 1977 for NH7 ("Newgate" NORFRAM002277007498 and "Nonnic" NORFRAM002977029292, two sons of "Valhalla" NORFRAM007668031767 who was not genotyped) and in 2003 for NH8 ("Uvray" NORFRAM002951401444). NH8 is a relatively new haplotype in Normande cattle, created by a recombination in the maternal gamete that was transmitted to "Uvray", with fewer opportunities to be observed in the homozygous state in the population. However, we considered both haplotypes for analyzing their effect on CR and NRR56.

Table 2.2. Estimated effect of NH7 and NH8 on conception rate (CR) and non-return rate at 56 days (NRR56) in at-risk mating[1]

Haplotype	Trait	Category	Number of mating			Mean (μ %)	Effect on fertility (%)	
			Non-risk	At-risk	Total		Expected[2]	Observed
NH7	CR	Heifer	2,981,552	14,474	2,996,026	57.37	-7.30	-5.53
		Cow	7,008,668	40,618	7,049,286	47.20	-6.01	-5.56
	NRR56	Heifer	1,870,062	7,513	1,877,575	76.62	-9.75	-1.76
		Cow	3,999,038	22,494	4,021,532	67.13	-8.55	-1.52
NH8	CR	Heifer	3,014,200	2,497	3,016,697	57.44	-7.25	0.64
		Cow	7,095,799	3,962	7,099,761	47.21	-5.95	0.38
	NRR56	Heifer	1,893,593	1,520	1,892,073	76.60	-9.66	0.47
		Cow	4,047,881	2,261	4,050,142	67.16	-8.47	0.55

[1]At risk mating are defined by mating between a carrier AI bull and a daughter of a carrier sire. Non-risk mating include all other combinations (carrier x non-carrier, non-carrier x carrier, and non-carrier x non-carrier).

[2]Expected effects on CR and NRR56 were calculated using the formula: $\frac{1}{4}\left(\frac{1}{2-f_{hap_k}}\right)\mu$; here f_{hap_k} is the frequency of NH7 or NH8 in the population, and μ is the phenotypic mean of CR or NRR56.

The NH7 but not the NH8 Haplotype Compromises Conception Rate in Normande Cattle

To elucidate whether NH7 and NH8 have negative effect on CR and NRR56, we analyzed ~10 million and ~6 million mating records, respectively, between genotyped bulls and daughters of genotyped bulls (Table 2.2).

For NH7, we observed a marked decrease in CR of 5.53% in heifers and 5.56% in cows, which is close to the expectation (7.30% and 6.01%, respectively). We also observed a moderate drop in NRR56 of 1.76% in

heifers and 1.52% in cows, much lower than the 9.75 and 8.55% expected values in both groups assuming complete lethality in homozygotes. The re-insemination within 56 days (NRR56) is an indirect measure of embryonic losses within 35 days of gestation, since the average estrus cycle in cow is 21 days (Reece et al., 2015). Our analysis shows that on average only 18% of the expected embryonic losses occurred within 35 days of gestation, while the majority of the failure may occurred later in the gestation. Here, we suspect a gradual embryonic loss over the length of gestation, rather than abrupt mass loss in the first month after fertilization.

We also analyzed the effect of NH8 on CR and NRR56, and found that this haplotype does not affect these two traits in Normande cattle (Table 2.2). It is worth mentioning that detection power was limited by the reduced number of at risk mating (0.06% vs. >0.5% of total mating for NH7) in our dataset because of the low frequency (1.9%) and the recent spreading of this haplotype in the population. Another possibility is that NH8 cause postnatal mortality, possibly with incomplete penetrance. Seven homozygotes, all inbred descendants of "Uvray", have been observed for this haplotype. These were all alive and aged between 11 months and 3 years at the time of this study. Despite NH8's significant HHD, it is necessary to accumulate more carrier-to-carrier mating, and inheritance in more generations, to be confident about the effect and HHD. Considering this and the fact that juvenile mortality is out of the scope of our study, we focused the subsequent analyses on NH7.

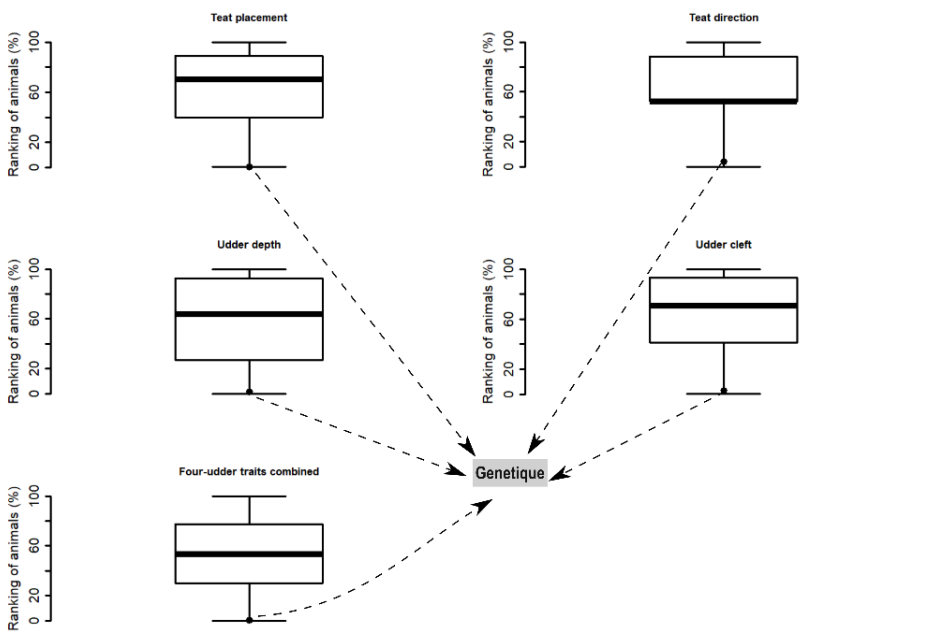


Figure 2.1. Ranking of animals (n=2,240) for four udder traits (%). *Black dots indicate the relative percentile of "Genetique" (the cow homozygous for NH7 haplotype) amongst the 2,239 paternal half-sisters. The percentile was calculated as:* $\frac{\#\ of\ animals\ with\ (\leq)\ of\ a\ given\ score}{total\ animals} \times 100$. *Combined scores (per animal): summation of scores for four-udder traits, followed by calculation of relative percentile for the given animal.*

The Homozygote for NH7 has Particularly Poor Udder Morphology

We found a single cow, named "Genetique", homozygous for NH7 haplotype and for a 20-Mb surrounding segment (~4.2 Mb upstream and 14.8 Mb downstream of the source haplotype) due to recent inbreeding. This individual deserves special attention. If it is homozygous at the causal mutation, penetrance is not complete but the animal may be affected by severe abnormality. Alternatively, it could be homozygous for the haplotype but not for the mutation, by carrying an ancestral version of the haplotype before the mutation event along with the lethal haplotype. The breeder informed us that this animal had normal health but was slaughtered during its first lactation due to very poor udder morphology. To verify this, we analyzed own performance records for 34 morphological traits on "Genetique" and its 2,239 paternal half-sisters. We confirmed that it was in the normal range, except for four udder traits, e.g. teat placement, teat direction, udder depth, and udder cleft, for which it ranked amongst the lowest 1, 5, 2, and 4% of its siblings, respectively (Figure 2.1, and Supplementary Figure S1). Furthermore, "Genetique" ranked amongst the lowest 1% on a linear combination of the scores of the four traits (Figure 2.1).

Fine-Mapping of the NH7 Locus in a ~1.01Mb Genomic Region

To fine-map the NH7 locus, we analyzed pedigree and found that among genotyped animals, half-brothers "Newgate" (NORFRAM002277007498) and "Nonnic" (NORFRAM002977029292) were the two most ancient carriers of the haplotype. Among them, however, "Nonnic" is the ancestor of many bulls with high breeding value. For example, "Elixir" (NORFRAM 005389014161), a descendent of "Nonnic", is the sixth most influential bull of the Normande breed with 4.7% contribution in the population (IDELE, 2017). Therefore, we considered only phenotypically normal descendants of "Nonnic" (and "Elixir") to fine-map the NH7 locus.

We identified two cows, "0520" (NORFRAF003708210520) and "Fanion" (NORFRAF001447822444), presenting a run-of-homozygosity on either side of the candidate haplotype (Figure 2.2). Both inherited NH7 paternally from "Nonnic" (through "Elixir") and NH7 recombinant haplotypes maternally. In dairy cattle, females were rarely genotyped and therefore, it is difficult to track maternally inherited recombinant haplotypes (Adams et al., 2016). Nevertheless, it is likely, from pedigree and sire-haplotype analyses, that "Nimbus" (Figure 2.2a) received the recombinant haplotype from "Elixir" and passed it to "0520". On the other hand, "Fanion" received a portion of the recombinant haplotype from "Nonnic", though the origin and source of the remaining portion is inconclusive due to un-genotyped female ancestors (Figure 2.2a).

It is noteworthy that both cows have normal health and udder morphology. In addition, during this study, "0520" and "Fanion" were in second and fourth lactation, respectively. Such phenotypically normal animals are very unlikely to be homozygous for genomic region harboring the causal mutation, unless the mutation is not fully penetrant. Under this assumption, the critical interval is defined by the two recombination points located: (i) between marker rs29022293 (Chr11:71,465,910) and rs29010798 (Chr11:71,520,448) in "Fanion", and (ii) between marker rs110176879 (Chr11:72,159,929) and rs109299742 (Chr11:72,476,622), in "0520". This enabled us to fine-map the locus from a ~1.73Mb initial region to a ~1.01Mb mutation-critical region (Figure 2.2b).

Identification of Causal Variant from WGS Data

To identify putative causal variants, we considered bi-allelic WGS variants (SNPs and indels), and large SVs.

We used SV calls on 389 WGS French cattle (2 NH7 carriers and 387 non-carriers) to identify SVs overlapping the NH7 haplotype and/or ±500 kb surrounding regions. We identified two deletions and one inversion within this interval, but none of them was present in the genomes of the NH7 carriers (Supplementary Table S4).

Then, we performed Pearson's correlation between NH7 carrier status of 2,333 animals from the 1KBGP (2 carriers and 2,331 non-carriers) and allele dosages for 33,640 bi-allelic WGS variants within the haplotype (±500 kb). Here, we identified eight strongly correlated (Pearson's correlation, $r > 0.8$) variants within the given interval (Figure 2.3). All these variants have high confidence call (avg. phred-scaled quality=162; min. 97), with sufficient mapping quality (avg. MQ=57; min. MQ=33) and read coverage (avg. coverage = 11.75x; min. 8.95x). Interestingly, five out of the eight variants were within our fine-mapped critical region, which could include the causal variant. However, it is known from studies that the best candidates do not always have perfect correlation with the causal haplotype (Michot et al., 2017, Fritz et al., 2018), which could be due to various artifacts (e.g. errors in phasing and imputation, errors in sequencing and WGS genotype calls, incomplete linkage disequilibrium with causal variant, etc.). We therefore considered all the eight variants for subsequent gene- and variant-level prioritization.

Identification of Causal Variant from WGS Data

To identify putative causal variants, we considered bi-allelic WGS variants (SNPs and indels), and large SVs. We used SV calls on 389 WGS French cattle (2 NH7 carriers and 387 non-carriers) to identify SVs overlapping the NH7 haplotype and/or ±500 kb surrounding regions. We identified two deletions and one inversion within this interval, but none of them was present in the genomes of the NH7 carriers (Supplementary Table S4).

Then, we performed Pearson's correlation between NH7 carrier status of 2,333 animals from the 1KBGP (2 carriers and 2,331 non-carriers) and allele dosages for 33,640 bi-allelic WGS variants within the haplotype (±500 kb). Here, we identified eight strongly correlated (Pearson's correlation, $r > 0.8$) variants within the given interval (Figure 2.3). All these variants have high confidence call (avg. phred-scaled quality=162; min. 97), with sufficient mapping quality (avg. MQ=57; min. MQ=33) and read coverage (avg. coverage = 11.75x; min. 8.95x). Interestingly, five out of the eight variants were within our fine-mapped critical region, which could include the causal variant. However, it is known from studies that the best candidates do not always have perfect correlation with the causal haplotype (Michot et al., 2017, Fritz et al., 2018), which could be due to various artifacts (e.g. errors in phasing and imputation, errors in sequencing and WGS genotype calls, incomplete linkage disequilibrium with causal variant, etc.). We therefore considered all the eight variants for subsequent gene- and variant-level prioritization.

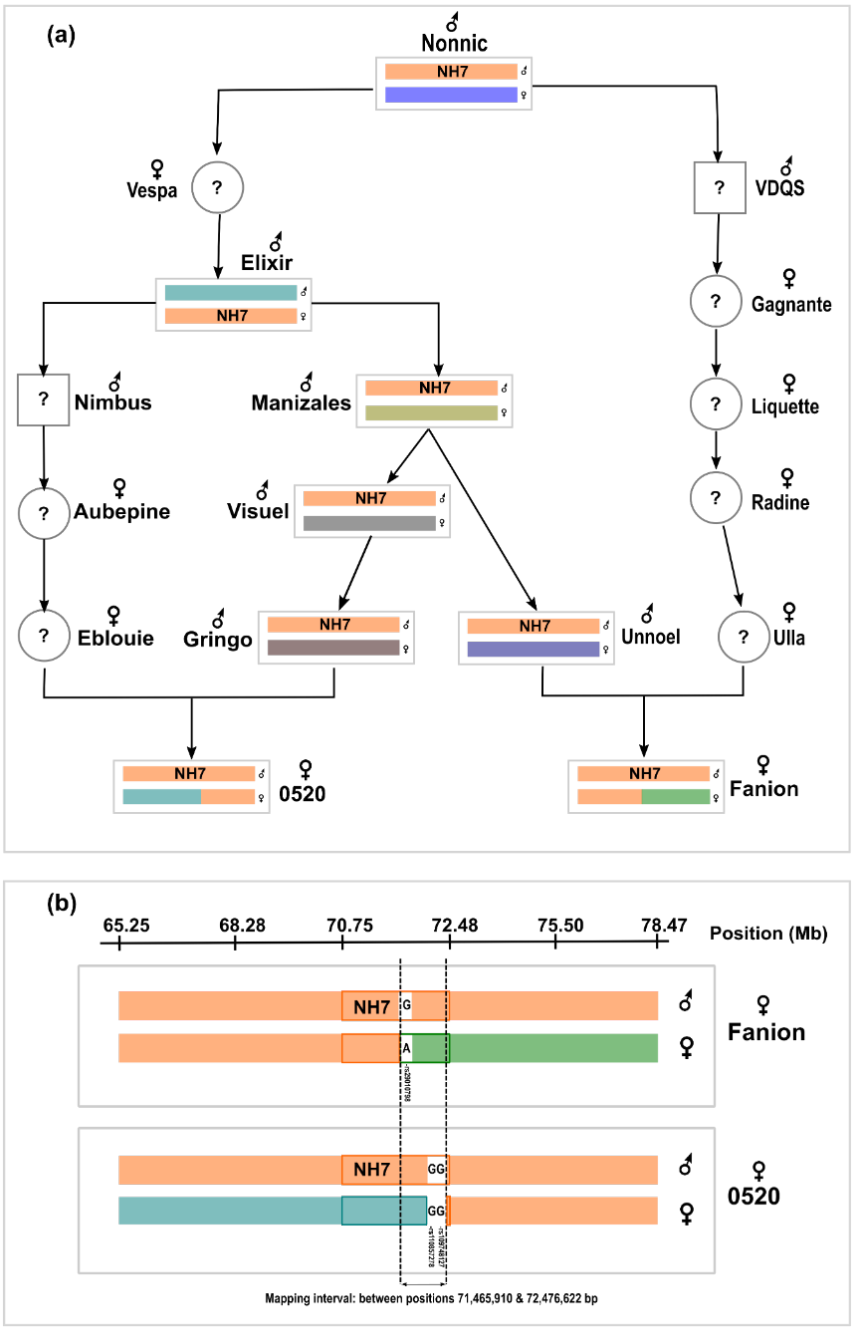


Figure 2.2. Fine-mapping at NH7 locus. *(a) Pedigrees of the two inbred descendants of "Nonnic" (NORFRAM002977029292), "0520" (NORFRAF003708210520) and "Fanion" (NORFRAF001447822444), both carry a copy of the 26-marker long original NH7 haplotype and a recombinant NH7 haplotype facilitating fine-mapping at the locus. (b) Diagram showing original NH7 and recombinant haplotype located between the marker rs41610536 (Chr11:70,750,209) and rs109299742 (Chr11:72,476,622), along with 100 markers on both sides of the haplotype: upstream markers are from rs41569373 (Chr11:65,246,700) to rs41572406 (Chr11:70,703,633), and downstream markers are from rs109666167 (Chr11:72,524,223) to rs110443631 (Chr11:78,468,627). The recombination occurred between marker rs29022293 (Chr11:71,465,910) and rs29010798 (Chr11:71,520,448) in "Fanion", and between marker rs110176879 (Chr11:72,159,929) and rs109299742 (Chr11:72,476,622), in "0520". The resulting critical region is located between marker rs29022293 and rs109299742, i.e. between chromosomal interval of 71,465,910 and 72,476,622 bp. Here, "NH7" denotes "Normande haplotype 7"; "?" denotes un-genotyped animal.*

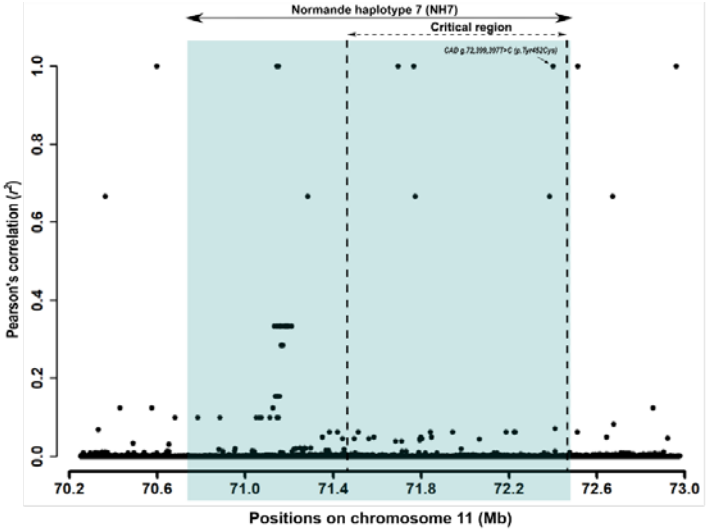


Figure 2.3. Identification of putative causal variants within NH7 locus (±500kb) using Pearson's correlation between haplotype status vs allele dosage of WGS variants. *Each dot in the scatterplot indicates a WGS variant. Perfectly correlated (r^2=1) CAD missense mutation is indicated by an arrow (with annotation).*

We annotated these eight WGS variants using the Variant Effect Predictor (VEP) software tool (version 87) from Ensembl (McLaren et al., 2016). There were one intergenic variant, two intronic variants in *PLB1*, three intronic variants in *BRE*, and one intronic and one missense variants in *CAD* gene (Table 2.3). Next, to prioritize the causal variant from the list, we assessed functional impact of these variants using three gene-level-, e.g. d_N/d_S, RVIS, and Z-score, and four variant-level-, e.g. SIFT, GERP, PhyloP and PhastCons, conservation scores.

To identify selective constraint on genes, we first analyzed the d_N/d_S ratio of cow-human 1-to-1 orthologs. Though all the three genes are somewhat constrained ($d_N/d_S < 1$), the selective constraint is stronger on *CAD* and *BRE* ($d_N/d_S < 0.1$). We later focused on RVIS and Z-score for human homologs of those genes, and found that *CAD* is the most constraint gene among the three, in terms of tolerance to functional variants (Table 2.3).

We next focused on variant-level conservation scores, e.g. SIFT, GERP, PhyloP, and PhastCons, to elucidate the strength of evolutionary constraint on those positions. SIFT score represents conservation at protein-level and predict whether an amino acid substitution is deleterious (SIFT < 0.05) or tolerated (SIFT > 0.05) for protein function. SIFT predicted the missense mutation *CAD* g.72399397T>C (p.Tyr452Cys) as "deleterious" for the protein function (Table 2.3). In contrast, GERP, PhyloP, and PhastCons represent conservation at nucleotide-level, generated from multi-species whole-genome sequence alignments, where higher positive scores indicate stronger evolutionary constraint. We were able to retrieve conservation scores for seven positions, using "LiftOver" and "UCSC Table Browser", while the remaining one did not have any corresponding position from UMD3.1/bosTau6 to GRCh37/hg19. Interestingly, all three classifiers predicted *CAD* g.72399397T>C as the most constrained position, where damaging mutations are purged from the population by strong purifying selection (Table 2.3).

Table 2.3. Integrative annotation of whole-genome sequence (WGS) variants strongly correlated with the NH7 haplotype

CHROM:POS[1]	REF/ALT	Sequence Ontology	Pearson's correlation (*r*)	Gene-level scores			Position-level scores[5]			
				d_N/d_S[2]	RVIS (%)[3]	Z-score[4]	SIFT	GERP++	PhyloP	PhastCons
Chr11:71144269	C/T	Intron variant of *PLB1* (ENSBTAG00000018669)	1	0.318	2.53 (98.5%)	-2.33	—[6]	1.14	0.18	0
Chr11:71151125	C/T	Intron variant of *PLB1* (ENSBTAG00000018669)	1	0.318	2.53 (98.5%)	-2.33	—	-3.37	-2.32	0
Chr11:71282689	C/T	Intergenic variant	0.82	—	—	—	—	—	—	—
Chr11:71695719	T/A	Intron variant of *BRE* (ENSBTAG00000031335)	1	0.095	-0.38 (32.33%)	1.13	—	-0.72	0.90	0
Chr11:71766806	A/C	Intron variant of *BRE* (ENSBTAG00000031335)	1	0.095	-0.38 (32.33%)	1.13	—	3.45	1.32	0.016
Chr11:71772162	G/A	Intron variant of *BRE* (ENSBTAG00000031335)	0.82	0.095	-0.38 (32.33%)	1.13	—	-4.32	-0.48	0
Chr11:72384541	C/T	Intron variant of *CAD* (ENSBTAG00000017894)	0.82	0.051	-4.84 (0.18%)	4.85	—	1.69	0.70	0
Chr11:72399397	T/C	Missense variant of *CAD* (ENSBTAG00000017894)	1	0.051	-4.84 (0.18%)	4.85	0.01	5.71	6.85	1

[1]Chromosomal positions correspond to bovine genome assembly UMD3.1

[2]*dN/dS* scores for human-cow 1-to-1 orthologs using BioMart (Kinsella et al., 2011) from Ensembl (Zerbino et al., 2018)

[3]Residual Variation Intolerance Score (RVIS) is retrieved from RVIS v4 (based on ExAC release 2.0; last accessed on 10 September 2018 from: http://genic-intolerance.org/) (Petrovski et al., 2013). Positive and negative RVIS scores represent tolerant and intolerant genes, respectively. RVIS "x%": it represents that the respective gene is amongst the "x" percentile of most intolerant human genes.

[4]Missense Z-scores were retrieved from http://exac.broadinstitute.org/ (Lek et al., 2016). Positive Z-scores indicate fewer variants than expected and thus stronger constraint on gene, vice versa.

[5]SIFT and sequence ontology annotations are from *Variant Effect Predictor* (VEP v87) software (McLaren et al., 2016); GERP, PhyloP, and PhastCons scores are from the *UCSC Table Browser* (http://genome.ucsc.edu/cgi-bin/hgTables) (Karolchik et al., 2004).

[6]"—" denotes unavailable information

Integrating gene- and variant-level conservation scores is an effective approach to prioritize causal variants from human genomes (Petrovski et al., 2013). Although bovine genome annotations are somewhat limited to sequence ontology and SIFT scores, incorporating cross-species annotations from well-studied genes or sequence homologs in humans improved our prioritization of the causal variant. Finally, the perfect correlation with NH7, localization within the critical interval, along with strong gene- and variant-level evolutionary conservation, taken together, strongly support *CAD* g.72399397T>C (p.Tyr452Cys) as the causal variant tagging the (nearly) lethal haplotype in Normande cattle.

The Tyrosine Residue at CAD p.Tyr452Cys is Fully Conserved from Humans to Bacteria

Pyrimidines are important building blocks of DNA and RNA, as well as precursors for uridine-diphosphate (UDP) sugars that play a vital role in post-translational modification, such as, UDP-dependent glycosylation of proteins (Jones, 1980, Fairbanks et al., 1995). Pyrimidines can either be synthesized *de novo* from small precursors (e.g. glutamine, ATP, and HCO_3^-), such as, in S-phage of proliferating cells (Mitchell and Hoogenraad, 1975), or through salvaging of nucleosides (e.g. uridine), such as, in resting or matured cells (Jones, 1980, Fairbanks et al., 1995). In mammals, the first three reactions in the *de novo* biosynthesis of pyrimidine are catalyzed by *CAD*—a tri-functional cytosolic protein composed of carbamoyl-phosphate synthetase 2 (CPSase-2), aspartate transcarbamylase (ATCase), and dihydroorotase (DHOase) domains (Jones, 1980, Moreno-Morcillo et al., 2017). An illustration of bovine *CAD* protein domains, from NCBI's Conserved Domain Database (CDD), is presented in Figure 2.4a.

Our candidate *CAD* mutation is located in CPSase-2 domain (Figure 2.4). We suspect that the wild-type allele (p.Tyr452 residue) is vital for CPSase-2 domain's activity, and therefore should be under strong evolutionary constraint. To assess the selective constraint on *CAD*, we performed multi-species sequence alignment among 24-species using Clustal Omega software (Sievers et al., 2011). We found that this T-to-C transition mutation (*CAD* g.72399397T>C) produces a tyrosine-to-cysteine substitution (p.Tyr452Cys) on CPSase-2 domain, replacing a fully conserved amino acid residue among the 24-species analyzed (Figure 2.4b) distributed in all life kingdoms.

Large-Scale Genotyping Confirmed Recessive Inheritance of CAD p.Tyr452Cys in Normande Cattle

We genotyped 33,323 Normande cattle for Chr11 g.72399397T>C (CAD p.Tyr452Cys) mutation (corresponding test in the Chip: g.72399397A>G in the TOP format) using Illumina EuroG10K BeadChip (Boichard et al., 2018) (Table 2.4). In this dataset, the NH7 haplotype and g.72399397G allele frequencies were 2.58 and 2.92%, respectively (Table 2.5). We did not observe homozygotes for g.72399397G allele, though 28 were expected under neutrality ($P < 7\times10^{-13}$), which largely corroborate with the observed recessive-lethal inheritance of NH7 (Table 2.1). This *CAD* mutation was also strongly associated with NH7 haplotype status; Pearson's correlation between g.72399397G allele dosages and haplotype status is, $r = 0.94$.

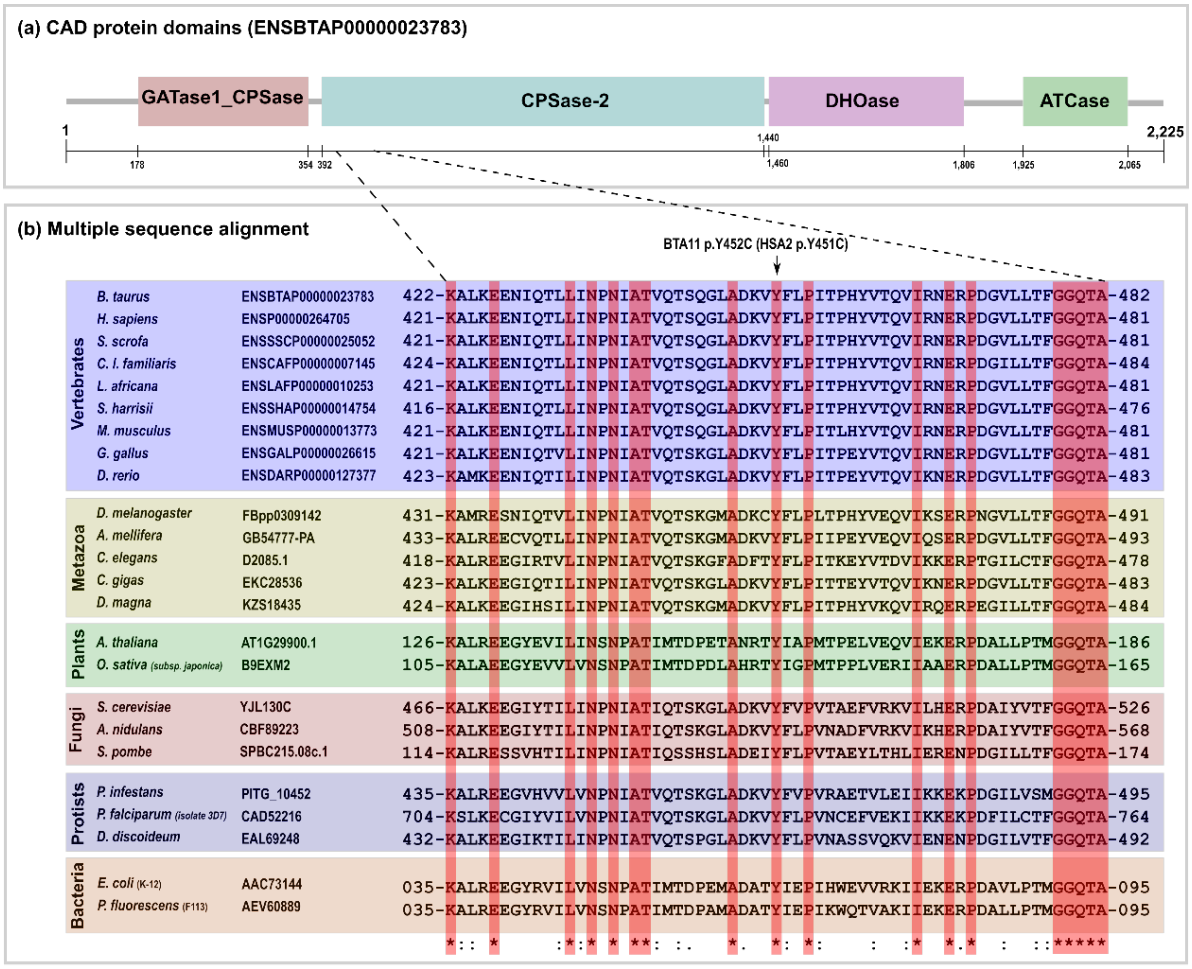


Figure 2.4. Protein domains and conserved sequences in CAD. *(a) Schematic representation of Bovine CAD protein domains from NCBI's Conserved Domain Database (CDD) (domain architecture ID 11486011; accessed from webserver "https://www.ncbi.nlm.nih.gov/cdd"). The diagram represents four functional protein domains, such as: (i) a small subunit of the glutamine-dependent form of carbamoyl-phosphate synthase II (GATase1_CPSase from 178-354; accession: cd01744), (ii) a large subunit of carbamoyl-phosphate synthase II (CPSase-II from 392-1440; accession: TIGR01369)—which catalyzes the first step in de novo pyrimidine biosynthesis, (iii) a subunit of dihydroorotase (DHOase from 1460-1806; accession: cd01316)—which catalyzes the third step in de novo pyrimidine biosynthesis, and (iv) a subunit of aspartate/ornithine carbamoyltransferase (OTCase_N from 1925-2065; accession: pfam02729)—which catalyzes the second step in de novo pyrimidine biosynthesis. (b) Multi-species sequence alignment representing a 61-residue long region in CPSase-II domain of CAD protein (centering at BTA11 p.Y452C) among nine vertebrates, five metazoans, two plants, three fungi, three protists, and two bacterial species. An arrow highlights the position of the CAD mutation, in bovine and orthologus mutation in humans. An asterisk (*) at the bottom of the alignment represents residues conserved in all 24-species (shaded vertical boxes); a colon (:) represents conservation among residues with strong similarity; a period (.) represents conservation among residues with weak similarity. BTA: Bos taurus autosome; HSA: Homo sapiens autosome.*

However, the association was not perfect: 237 heterozygotes for the *CAD* mutation do not carry NH7 haplotype and eight heterozygous for NH7 do not carry the *CAD* mutation (Table 2.5). Such imperfect association can arise from recombination between NH7 and other haplotypes, along with other reasons, such as, genotyping, phasing or imputation errors. Further investigations revealed that approximately 67% (158/237) of the heterozygotes for the *CAD* mutation but not for NH7 carry the same recombinant NH7

haplotype with a long stretch of source haplotype downstream. In addition, the eight animals heterozygous for *CAD* mutation do not carry the original NH7 but a recombinant identical-by-state (IBS) portion of the haplotype.

Similarly, previous studies showed that haplotypes are not a perfect proxy for causal mutations (Michot et al., 2017, Fritz et al., 2018), and there were instances where two different versions of the same haplotype segregate in the population, i.e. one with- and other without-causal mutation (Kipp et al., 2016, Menzi et al., 2016). Yet, haplotype is a good substitute to avoid carrier-to-carrier mating, when causal mutation is unknown (VanRaden et al., 2011, Sahana et al., 2013), or when test results are confidential due to patents (Cole et al., 2016).

We next analyzed 348,593 additional animals (from 19 other French cattle breeds), genotyped for genomic selection, to elucidate whether this mutation is also observed in other populations. We found that the *CAD* mutation is only segregating in Normande cattle (Table 2.4).

Table 2.4. Genotypes at CAD g.72399397A>G mutation in 20 French cattle breeds[1]

Breed	**Genotypes[2]**		
	AA	AG	GG
Abondance	**5,161**	**0**	**0**
Aubrac	**66**	**0**	**0**
Blonde d'Aquitaine	**2,705**	**0**	**0**
Bretonne pie noir	**7**	**0**	**0**
Brown Swiss	**2,742**	**0**	**0**
Charolaise	**10,213**	**0**	**0**
Créole	**32**	**0**	**0**
Gasconne	**1**	**0**	**0**
Holstein	**180,565**	**0**	**0**
INRA95	**20**	**0**	**0**
Jersey	**785**	**0**	**0**
Limousine	**887**	**0**	**0**
Montbéliarde	**138,851**	**0**	**0**
Normande	**31,376**	**1,947**	**0**
Parthenaise	**1,439**	**0**	**0**
Rouge des Prés	**4**	**0**	**0**
Salers	**15**	**0**	**0**
Simmental	**268**	**0**	**0**
Tarentaise	**2,750**	**0**	**0**
Vosgienne	**2,082**	**0**	**0**

[1]In total, 381,916 animals were genotyped for the candidate-SNP with EuroG10K custom SNP chip (Boichard et al., 2018).
[2]Genotype: AA= "*CAD* p.452Tyr/Tyr"; AG= "*CAD* p. 452Tyr/Cys"; GG= "*CAD* p.452Cys/Cys".

Table 2.5. Concordance between NH7 status and genotypes for CAD g.72399397A>G in 33,323 Normande cattle

Genotype	**+/+[1]**	**NH7/+**	**NH7/NH7**	**Total**
g.72399397 A/A	31,368	8	0	31,376
g.72399397 A/G	237	1,710	0	1,947
g.72399397 G/G	0	0	0	0
Total	31,605	1,718	0	33,323

[1]"+": Non-NH7 haplotype

Surviving CAD Homozygotes Exhibit Broad Range of Symptoms Depending on the Species and Nature of Mutation

For verification, we genotyped "Genetique", the cow homozygous for a 20-Mb region including NH7, through PCR and Sanger sequencing and found that it was also homozygous for our candidate mutation (Figure 2.5).

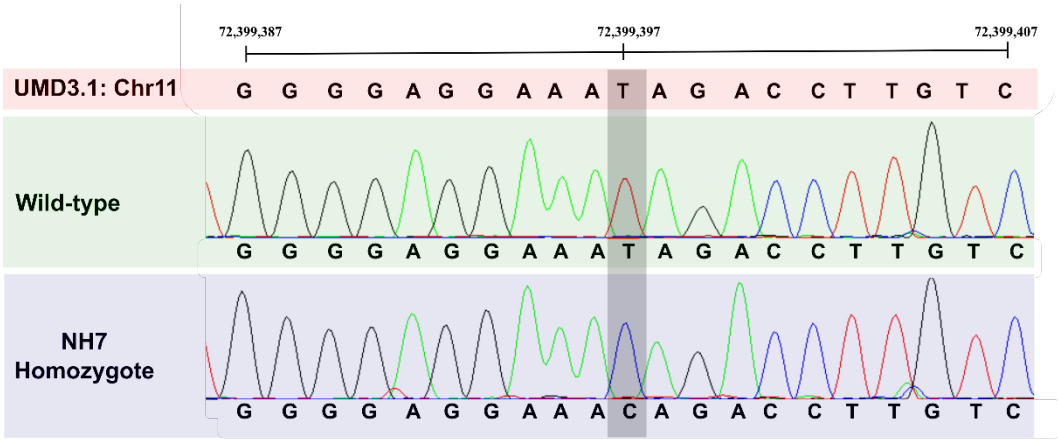


Figure 2.5. Genotype of NH7 homozygote at Chr11 g.72399397T>C mutation. *The CAD mutation is indicated by the shaded vertical box.*

Despite being alive and apparently healthy and fertile, this animal had very poor udder morphology (scored amongst the lowest 1% of animals with same sire as discussed earlier; Figure 2.1). This is consistent with previous observation in model organisms and humans where deleterious mutation in *CAD* lead to embryonic lethality with incomplete penetrance (Norby, 1973, Falk and Nash, 1974, Willer et al., 2005, Franks et al., 2006, Cox et al., 2014, Ng et al., 2015). Noteworthy, studies in human and Drosophila demonstrated that uridine supplementation reduces the severity of the symptoms by salvage pathway of pyrimidine biosynthesis (Falk and Nash, 1974, Ng et al., 2015, Koch et al., 2017). Depending on the species and nature of the mutation, surviving homozygotes display a broad range of clinical manifestations including defective development of embryonic ectoderm and ectodermal derivatives. Defective pectoral fins in zebrafish (Willer et al., 2005), nerve cells in humans (Ng et al., 2015, Koch et al., 2017), and wings in fruit flies (Norby, 1970, 1973, Falk and Nash, 1974), are among such examples. The poor udder morphology observed in the homozygote cow could be added to this list since the mammary gland is also of ectodermal origin (Hens and Wysolmerski, 2005, Cowin and Wysolmerski, 2010, Macias and Hinck, 2012).

To test whether the *CAD* mutation had an effect on the development of the mammary gland, we analyzed the performances of the daughters of 250 heterozygous carriers and 2,912 non-carriers for four traits (teat direction, teat placement, udder cleft, and udder depth). We did not find significant difference between DYD of heterozygous vs wild homozygous bulls (Figure 2.6). In addition, there was no reported udder related QTL in Normande cattle, either within NH7 interval or on chromosome 11 (Boichard et al., 2003, Marete et al., 2018). This is in line with a possible recessive effect of the mutation on udder morphology in surviving homozygotes, although more live homozygous cows are needed to confirm this assumption.

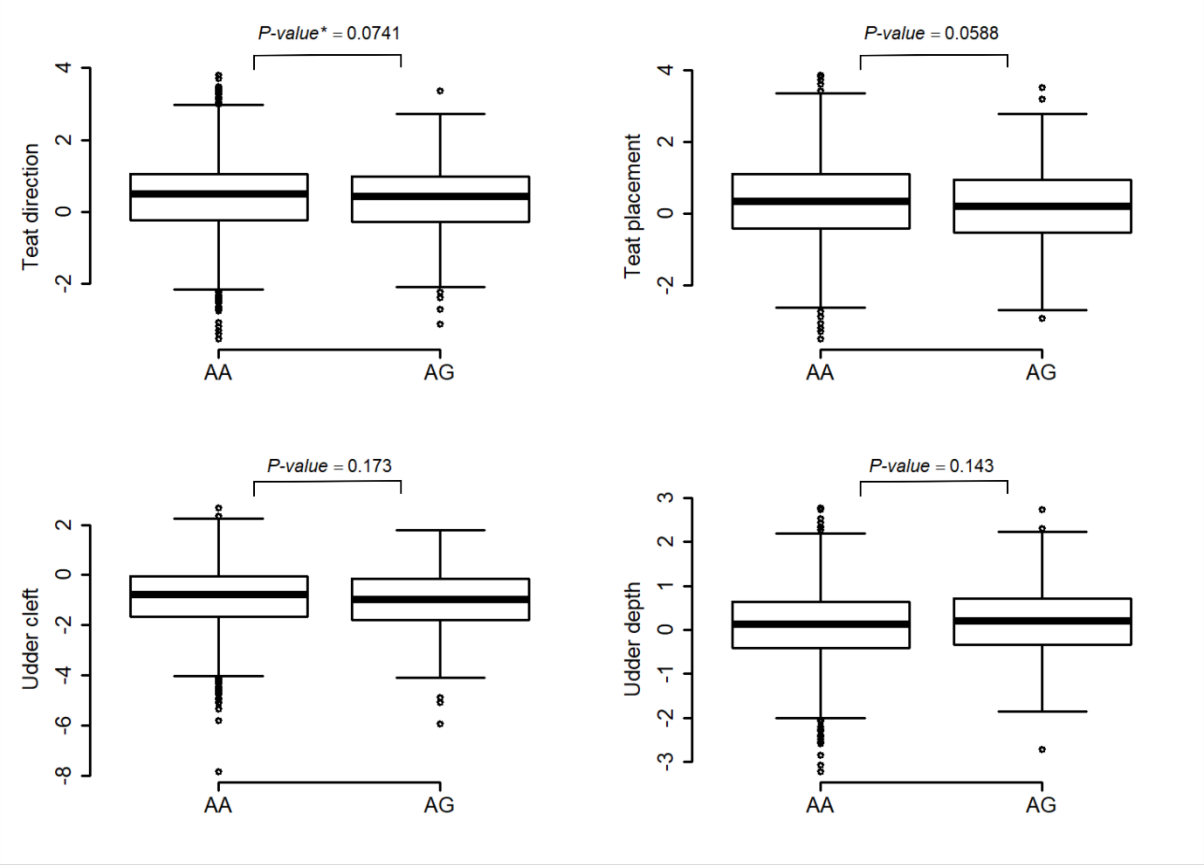


Figure 2.6. Effect of CAD mutation on four-udder traits. ** Here, P-values are from Student's t-test on 250-animals with AG-genotype (carriers) vs 2,912 animals with AA-genotypes (non-carriers).*

2.5 Conclusion

Here, we perform a systematic screening for recessive lethal mutations from large-scale haplotype scan to single-variant resolution mapping in Normande cattle. We report the identification of a deleterious substitution in *CAD* (p.Tyr452Cys) responsible for embryonic lethality with incomplete penetrance. We illustrate the influence of sample size, age of haplotype, and correction for multiple testing, on discovery and replicability of HHD results. Furthermore, we highlight the importance of large-scale genotyping of candidate variants, and phenotyping of homozygotes, to validate or invalidate them. Though such approaches require collection of genomic data on hundreds of thousands of animals over several years, thanks to genomic selection, this is a routine practice in dairy cattle breeding. Finally, we added a test for *CAD* (p.Tyr452Cys) to the EuroG10K BeadChip, which is routinely used for genomic evaluation in France. We recommend using *CAD* information to avoid at-risk mating. Because the economic weight of embryonic lethal defects is limited, we also suggest including it in the breeding objective as proposed by others, instead of eliminating all carriers.

Author Contributions: MMU, DB and AC conceived the study. MMU, CH, AB, PM, RL, MB, SF, DB and AC contributed in collection and/or analysis of the data. MMU drafted the manuscript. GS, DB and AC critically revised the manuscript. AC coordinated the study. All authors read and approved the final manuscript.

Acknowledgements: This study is part of the BOVANO project (ANR-14- CE19-0011) funded by the French Agence Nationale de la Recherche (Paris, France) and APIS-GENE (Paris, France). Md Mesbah-Uddin (MMU) benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate "EGS-ABG". MMU's PhD project is also supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B). The authors are grateful to the partners of the 1000 bull genomes consortium for their excellent collaboration. We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources.

Supplementary Material

2.6 References

Adams, H. A., T. S. Sonstegard, P. M. VanRaden, D. J. Null, C. P. Van Tassell, D. M. Larkin, and H. A. Lewin. 2016. Identification of a nonsense mutation in APAF1 that is likely causal for a decrease in reproductive efficiency in Holstein dairy cattle. J Dairy Sci 99:6693-6701. https://doi.org/10.3168/jds.2015-10517.

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

Boichard, D., M. Boussaha, A. Capitan, D. Rocha, C. Hozé, M. P. Sanchez, T. Tribout, R. Letaief, P. Croiseau, C. Grohs, W. Li, C. Harland, C. Charlier, M. S. Lund, G. Sahana, M. Georges, S. Barbier, W. Coppieters, S. Fritz, and B. Guldbrandtsen. 2018. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. Page 675 in Proc. Proc. World Congr. Genet. Appl. Livest. Prod., Auckland, New Zealand. AL Rae Centre for Genetics and Breeding,

Massey University, Palmerston North, New Zealand. http://www.wcgalp.org/system/files/proceedings/2018/experience-large-scale-use-eurogenomics-custom-snp-chip-cattle.pdf.

Boichard, D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, K. J. Gietzen, B. J. Hayes, C. T. Lawley, T. S. Sonstegard, C. P. Van Tassell, P. M. VanRaden, K. A. Viaud-Martinez, G. R. Wiggans, and L. D. C. Bovine. 2012a. Design of a bovine low-density SNP array optimized for imputation. PLoS One 7:e34130. https://doi.org/10.1371/journal.pone.0034130.

Boichard, D., C. Grohs, F. Bourgeois, F. Cerqueira, R. Faugeras, A. Neau, R. Rupp, Y. Amigues, M. Y. Boscher, and H. Leveziel. 2003. Detection of genes influencing economic traits in three French dairy cattle breeds. Genet Sel Evol 35:77-101. https://doi.org/10.1051/gse:2002037.

Boichard, D., C. Grohs, P. Michot, C. Danchin-Burge, A. Capitan, L. Genestout, S. Barbier, and S. Fritz. 2016. Prise en compte des anomalies génétiques en sélection : le cas des bovins. INRA Prod. Anim. 29:351-358.

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M. N. Rossignol, M. Y. Boscher, T. Druet, L. Genestout, J. J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012b. Genomic selection in French dairy cattle. Animal Production Science 52:115-120. https://doi.org/https://doi.org/10.1071/AN11119.

Casper, J., A. S. Zweig, C. Villarreal, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, D. Karolchik, A. S. Hinrichs, M. Haeussler, L. Guruvadoo, J. Navarro Gonzalez, D. Gibson, I. T. Fiddes, C. Eisenhart, M. Diekhans, H. Clawson, G. P. Barber, J. Armstrong, D. Haussler, R. M. Kuhn, and W. J. Kent. 2018. The UCSC Genome Browser database: 2018 update. Nucleic Acids Res 46:D762-D769. https://doi.org/10.1093/nar/gkx1020.

Charlier, C., W. Coppieters, F. Rollin, D. Desmecht, J. S. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, J. C. Frennet, R. Hanset, X. Hubin, C. Jorgensen, L. Karim, M. Kent, K. Harvey, B. R. Pearce, P. Simon, N. Tama, H. Nie, S. Vandeputte, S. Lien, M. Longeri, M. Fredholm, R. J. Harvey, and M. Georges. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nat Genet 40:449-454. https://doi.org/10.1038/ng.96.

Chen, K., J. W. Wallis, M. D. McLellan, D. E. Larson, J. M. Kalicki, C. S. Pohl, S. D. McGrath, M. C. Wendl, Q. Zhang, D. P. Locke, X. Shi, R. S. Fulton, T. J. Ley, R. K. Wilson, L. Ding, and E. R. Mardis. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6:677-681. https://doi.org/10.1038/nmeth.1363.

Cole, J. B., D. J. Null, and P. M. VanRaden. 2016. Phenotypic and genetic effects of recessive haplotypes on yield, longevity, and fertility. J Dairy Sci 99:7274-7288. https://doi.org/10.3168/jds.2015-10777.

Cowin, P. and J. Wysolmerski. 2010. Molecular mechanisms guiding embryonic mammary gland development. Cold Spring Harb Perspect Biol 2:a003251. https://doi.org/10.1101/cshperspect.a003251.

Cox, J. A., A. LaMora, S. L. Johnson, and M. M. Voigt. 2014. Novel role for carbamoyl phosphate synthetase 2 in cranial sensory circuit formation. Int J Dev Neurosci 33:41-48. https://doi.org/10.1016/j.ijdevneu.2013.11.003.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46:858-865. https://doi.org/10.1038/ng.3034.

Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025. https://doi.org/10.1371/journal.pcbi.1001025.

Fairbanks, L. D., M. Bofill, K. Ruckemann, and H. A. Simmonds. 1995. Importance of ribonucleotide availability to proliferating T-lymphocytes from healthy humans. Disproportionate expansion of pyrimidine pools and contrasting effects of de novo synthesis inhibitors. J Biol Chem 270:29682-29689.

Falk, D. R. and D. Nash. 1974. Sex-linked auxotrophic and putative auxotrophic mutants of Drosophila melanogaster. Genetics 76:755-766.

Franks, D. M., T. Izumikawa, H. Kitagawa, K. Sugahara, and P. G. Okkema. 2006. C. elegans pharyngeal morphogenesis requires both de novo synthesis of pyrimidines and synthesis of heparan sulfate proteoglycans. Dev Biol 296:409-420. https://doi.org/10.1016/j.ydbio.2006.06.008.

Fritz, S., A. Capitan, A. Djari, S. C. Rodriguez, A. Barbat, A. Baur, C. Grohs, B. Weiss, M. Boussaha, D. Esquerre, C. Klopp, D. Rocha, and D. Boichard. 2013. Detection of haplotypes associated with prenatal death in dairy cattle and identification of deleterious mutations in GART, SHBG and SLC37A2. PLoS One 8:e65550. https://doi.org/10.1371/journal.pone.0065550.

Fritz, S., C. Hoze, E. Rebours, A. Barbat, M. Bizard, A. Chamberlain, C. Escouflaire, C. Vander Jagt, M. Boussaha, C. Grohs, A. Allais-Bonnet, M. Philippe, A. Vallee, Y. Amigues, B. J. Hayes, D. Boichard, and A. Capitan. 2018. An initiator codon mutation in SDE2 causes recessive embryonic lethality in Holstein cattle. J Dairy Sci. https://doi.org/10.3168/jds.2017-14119.

Hens, J. R. and J. J. Wysolmerski. 2005. Key stages of mammary gland development: molecular mechanisms involved in the formation of the embryonic mammary gland. Breast Cancer Res 7:220-224. https://doi.org/10.1186/bcr1306.

ICAR. 2018. Guidelines for Conformation Recording of Dairy Cattle, Beef Cattle and Dairy Goats. date accessed (10/25/2018). url:https://www.icar.org/Guidelines/05-Conformation-Recording.pdf.

IDELE. 2017. Indicators of genetic variability - Normande. date accessed (10/25/2018). url:http://idele.fr/fileadmin/medias/Documents/SIG_56_2017.pdf.

Jones, M. E. 1980. Pyrimidine nucleotide biosynthesis in animals: genes, enzymes, and regulation of UMP biosynthesis. Annu Rev Biochem 49:253-279. https://doi.org/10.1146/annurev.bi.49.070180.001345.

Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32:D493-496. https://doi.org/10.1093/nar/gkh103.

Kersey, P. J., J. E. Allen, A. Allot, M. Barba, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, C. Grabmueller, N. Kumar, Z. Liu, T. Maurel, B. Moore, M. D. McDowall, U. Maheswari, G. Naamati, V. Newman, C. K. Ong, M. Paulini, H. Pedro, E. Perry, M. Russell, H. Sparrow, E. Tapanari, K. Taylor, A. Vullo, G. Williams, A. Zadissia, A. Olson, J. Stein, S. Wei, M. Tello-Ruiz, D. Ware, A. Luciani, S. Potter, R. D. Finn, M. Urban, K. E. Hammond-Kosack, D. M. Bolser, N. De Silva, K. L. Howe, N. Langridge, G. Maslen, D. M. Staines, and A. Yates. 2018. Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res 46:D802-D808. https://doi.org/10.1093/nar/gkx1011.

Kinsella, R. J., A. Kahari, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, P. Kersey, and P. Flicek. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030. https://doi.org/10.1093/database/bar030.

Kipp, S., D. Segelke, S. Schierenbeck, F. Reinhardt, R. Reents, C. Wurmser, H. Pausch, R. Fries, G. Thaller, J. Tetens, J. Pott, D. Haas, B. B. Raddatz, M. Hewicker-Trautwein, I. Proios, M. Schmicke, and W. Grunberg. 2016. Identification of a haplotype associated with cholesterol deficiency and increased juvenile mortality in Holstein cattle. J Dairy Sci 99:8915-8931. https://doi.org/10.3168/jds.2016-11118.

Koch, J., J. A. Mayr, B. Alhaddad, C. Rauscher, J. Bierau, R. Kovacs-Nagy, K. L. Coene, I. Bader, M. Holzhacker, H. Prokisch, H. Venselaar, R. A. Wevers, F. Distelmaier, T. Polster, S. Leiz, C. Betzler, T. M. Strom, W. Sperl, T. Meitinger, S. B. Wortmann, and T. B. Haack. 2017. CAD mutations and uridine-responsive epileptic encephalopathy. Brain 140:279-286. https://doi.org/10.1093/brain/aww300.

Kumar, P., S. Henikoff, and P. C. Ng. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073-1081. https://doi.org/10.1038/nprot.2009.86.

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O'Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H. H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M. J. Daly, D. G. MacArthur, and C. Exome Aggregation. 2016. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536:285-291. https://doi.org/10.1038/nature19057.

Letaief, R., E. Rebours, C. Grohs, C. Meersseman, S. Fritz, L. Trouilh, D. Esquerre, J. Barbieri, C. Klopp, R. Philippe, V. Blanquet, D. Boichard, D. Rocha, and M. Boussaha. 2017. Identification of copy number variation in French dairy and beef breeds using next-generation sequencing. Genet Sel Evol 49:77. https://doi.org/10.1186/s12711-017-0352-z.

Macias, H. and L. Hinck. 2012. Mammary gland development. Wiley Interdiscip Rev Dev Biol 1:533-557. https://doi.org/10.1002/wdev.35.

Marchler-Bauer, A., Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, L. Y. Geer, and S. H. Bryant. 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res 45:D200-D203. https://doi.org/10.1093/nar/gkw1129.

Marete, A., M. S. Lund, D. Boichard, and Y. Ramayo-Caldas. 2018. A system-based analysis of the genetic determinism of udder conformation and health phenotypes across three French dairy cattle breeds. PLoS One 13:e0199931. https://doi.org/10.1371/journal.pone.0199931.

Matukumalli, L. K., C. T. Lawley, R. D. Schnabel, J. F. Taylor, M. F. Allan, M. P. Heaton, J. O'Connell, S. S. Moore, T. P. Smith, T. S. Sonstegard, and C. P. Van Tassell. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One 4:e5350. https://doi.org/10.1371/journal.pone.0005350.

McClure, M. C., D. Bickhart, D. Null, P. Vanraden, L. Xu, G. Wiggans, G. Liu, S. Schroeder, J. Glasscock, J. Armstrong, J. B. Cole, C. P. Van Tassell, and T. S. Sonstegard. 2014. Bovine exome sequence analysis and targeted SNP genotyping of recessive fertility defects BH1, HH2, and HH3 reveal a putative causative mutation in SMC2 for HH3. PLoS One 9:e92769. https://doi.org/10.1371/journal.pone.0092769.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. The Ensembl Variant Effect Predictor. Genome Biol 17:122. https://doi.org/10.1186/s13059-016-0974-4.

Menzi, F., N. Besuchet-Schmutz, M. Fragniere, S. Hofstetter, V. Jagannathan, T. Mock, A. Raemy, E. Studer, K. Mehinagic, N. Regenscheit, M. Meylan, F. Schmitz-Hsu, and C. Drogemuller. 2016. A transposable element insertion in APOB causes cholesterol deficiency in Holstein cattle. Anim Genet 47:253-257. https://doi.org/10.1111/age.12410.

Michot, P., S. Fritz, A. Barbat, M. Boussaha, M. C. Deloche, C. Grohs, C. Hoze, L. Le Berre, D. Le Bourhis, O. Desnoes, P. Salvetti, L. Schibler, D. Boichard, and A. Capitan. 2017. A missense mutation in PFAS (phosphoribosylformylglycinamidine synthase) is likely causal for embryonic lethality associated with the MH1 haplotype in Montbeliarde dairy cattle. J Dairy Sci 100:8176-8187. https://doi.org/10.3168/jds.2017-12579.

Mitchell, A. D. and N. J. Hoogenraad. 1975. De novo pyrimidine nucleotide biosynthesis in synchronized rat hepatoma (HTC) cells and mouse embryo fibroblast (3T3) cells. Exp Cell Res 93:105-110.

Moreno-Morcillo, M., A. Grande-Garcia, A. Ruiz-Ramos, F. Del Cano-Ochoa, J. Boskovic, and S. Ramon-Maiques. 2017. Structural Insight into the Core of CAD, the Multifunctional Protein Leading De Novo Pyrimidine Biosynthesis. Structure 25:912-923 e915. https://doi.org/10.1016/j.str.2017.04.012.

Ng, B. G., L. A. Wolfe, M. Ichikawa, T. Markello, M. He, C. J. Tifft, W. A. Gahl, and H. H. Freeze. 2015. Biallelic mutations in CAD, impair de novo pyrimidine biosynthesis and decrease glycosylation precursors. Hum Mol Genet 24:3050-3057. https://doi.org/10.1093/hmg/ddv057.

Norby, S. 1970. A specific nutritional requirement for pyrimidines in rudimentary mutants of Drosophila melanogaster. Hereditas 66:205-214.

Norby, S. 1973. The biochemical genetics of rudimentary mutants of Drosophila melanogaster. I. Aspartate carbamoyltransferase levels in complementing and non-complementing strains. Hereditas 73:11-16.

Pausch, H., H. Schwarzenbacher, J. Burgstaller, K. Flisikowski, C. Wurmser, S. Jansen, S. Jung, A. Schnieke, T. Wittek, and R. Fries. 2015. Homozygous haplotype deficiency reveals deleterious mutations compromising reproductive and rearing success in cattle. BMC Genomics 16:312. https://doi.org/10.1186/s12864-015-1483-7.

Petrovski, S., Q. Wang, E. L. Heinzen, A. S. Allen, and D. B. Goldstein. 2013. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 9:e1003709. https://doi.org/10.1371/journal.pgen.1003709.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom, and A. Siepel. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20:110-121. https://doi.org/10.1101/gr.097857.109.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stutz, V. Benes, and J. O. Korbel. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28:i333-i339. https://doi.org/10.1093/bioinformatics/bts378.

Reece, W. O., H. H. Erickson, J. P. Goff, and E. E. Uemura. 2015. Dukes' physiology of domestic animals. 13th edition / ed. Wiley Blackwell, Ames, Iowa, USA.

Sahana, G., U. S. Nielsen, G. P. Aamand, M. S. Lund, and B. Guldbrandtsen. 2013. Novel harmful recessive haplotypes identified for fertility traits in Nordic Holstein cattle. PLoS One 8:e82909. https://doi.org/10.1371/journal.pone.0082909.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. https://doi.org/10.1186/1471-2164-15-478.

Schwarzenbacher, H., J. Burgstaller, F. R. Seefried, C. Wurmser, M. Hilbe, S. Jung, C. Fuerst, N. Dinhopl, H. Weissenbock, B. Fuerst-Waltl, M. Dolezal, R. Winkler, O. Grueter, U. Bleul, T. Wittek, R. Fries, and H. Pausch. 2016a. A missense mutation in TUBD1 is associated with high juvenile mortality in Braunvieh and Fleckvieh cattle. BMC Genomics 17:400. https://doi.org/10.1186/s12864-016-2742-y.

Schwarzenbacher, H., C. Wurmser, K. Flisikowski, L. Misurova, S. Jung, M. C. Langenmayer, A. Schnieke, G. Knubben-Schweizer, R. Fries, and H. Pausch. 2016b. A frameshift mutation in GON4L is associated with proportionate dwarfism in Fleckvieh cattle. Genet Sel Evol 48:25. https://doi.org/10.1186/s12711-016-0207-z.

Segelke, D., H. Taubert, F. Reinhardt, and G. Thaller. 2016. Considering genetic characteristics in German Holstein breeding programs. J Dairy Sci 99:458-467. https://doi.org/10.3168/jds.2015-9764.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75.

Sonstegard, T. S., J. B. Cole, P. M. VanRaden, C. P. Van Tassell, D. J. Null, S. G. Schroeder, D. Bickhart, and M. C. McClure. 2013. Identification of a nonsense mutation in CWC15 associated with decreased reproductive efficiency in Jersey cattle. PLoS One 8:e54872. https://doi.org/10.1371/journal.pone.0054872.

The UniProt, C. 2017. UniProt: the universal protein knowledgebase. Nucleic Acids Res 45:D158-D169. https://doi.org/10.1093/nar/gkw1099.

Untergasser, A., I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen. 2012. Primer3--new capabilities and interfaces. Nucleic Acids Res 40:e115. https://doi.org/10.1093/nar/gks596.

VanRaden, P. M., K. M. Olson, D. J. Null, and J. L. Hutchison. 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J Dairy Sci 94:6153-6161. https://doi.org/10.3168/jds.2011-4624.

Weckx, S., J. Del-Favero, R. Rademakers, L. Claes, M. Cruts, P. De Jonghe, C. Van Broeckhoven, and P. De Rijk. 2005. novoSNP, a novel computational tool for sequence variation discovery. Genome Res 15:436-442. https://doi.org/10.1101/gr.2754005.

Weigel, K. A. 2001. Controlling Inbreeding in Modern Breeding Programs. Journal of Dairy Science 84:E177-E184. https://doi.org/https://doi.org/10.3168/jds.S0022-0302(01)70213-5.

Willer, G. B., V. M. Lee, R. G. Gregg, and B. A. Link. 2005. Analysis of the Zebrafish perplexed mutation reveals tissue-specific roles for de novo pyrimidine synthesis during development. Genetics 170:1827-1837. https://doi.org/10.1534/genetics.105.041608.

Ye, K., M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25:2865-2871. https://doi.org/10.1093/bioinformatics/btp394.

Zerbino, D. R., P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Giron, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier,

D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek. 2018. Ensembl 2018. Nucleic Acids Res 46:D754-D761. https://doi.org/10.1093/nar/gkx1098.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol 10:R42. https://doi.org/10.1186/gb-2009-10-4-r42.

Chapter 3.
Genome-wide mapping of large deletions and their population-genetic properties in dairy cattle

Md Mesbah-Uddin[1,2,*], Bernt Guldbrandtsen[1], Terhi Iso-Touru[3], Johanna Vilkki[3], Dirk-Jan De Koning[4], Didier Boichard[2], Mogens Sandø Lund[1], and Goutam Sahana[1,*]

[1]Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark

[2]Animal Genetics and Integrative Biology, UMR 1313 GABI, INRA, AgroParisTech, Universite´ Paris-Saclay, 78350 Jouy-en-Josas, France

[3]Green Technology, Natural Resources Institute Finland, FI-31600 Jokioinen, Finland

[4]Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, SE-750 07 Uppsala, Sweden

[*]Corresponding Authors: mdmesbah@gmail.com (M.M.-U.) and goutam.sahana@mbg.au.dk (G.S.)

DNA Research, 2018. 25(1): p. 49-59.

Supplementary materials are available at https://doi.org/10.1093/dnares/dsx037

Note: Chapter 3 is identical to the paper except for page layout and journal amended text formatting and editing, and "Author Contributions" section.

3.1 Abstract

Large genomic deletions are potential candidate for loss-of-function, which could be lethal as homozygote. Analyzing whole genome data of 175 cattle, we report 8,480 large deletions (199bp to 773KB) with an overall false discovery rate of 8.8%; 82% of which are novel compared to deletions in the dbVar database. Breakpoint sequence analyses revealed that majority (24 of 29 tested) of the deletions contain microhomology/homology at breakpoint, and therefore, most likely generated by microhomology-mediated end joining. We observed higher differentiation among breeds for deletions in some genic-regions, such as ABCA12, TTC1, VWA3B, TSHR, DST/BPAG1, and CD1D. The genes overlapping deletions are on average evolutionarily less conserved compared to known mouse lethal genes (p-value=2.3×10^{-6}). We report 167 natural gene knockouts in cattle that are apparently nonessential as live homozygote individuals are observed. These genes are functionally enriched for immunoglobulin domains, olfactory receptors, and MHC classes (FDR=2.06×10^{-22}, 2.06×10^{-22}, 7.01×10^{-6}, respectively). We also demonstrate that deletions are enriched for health and fertility related QTL (2 and 1.5 fold enrichment, Fisher's p-value=8.91×10^{-10} and 7.4×10^{-11}, respectively). Finally, we identified and confirmed the breakpoint of a ~525 KB deletion on Chr23:12,291,761-12,817,087 (overlapping BTBD9, GLO1 and DNAH8), causing stillbirth in Nordic Red Cattle.

Key words: dairy cattle, structural variants, whole genome sequence, population genetics, loss-of-function

3.2 Introduction

Embryonic lethality has become a challenge to cattle breeders, especially for dairy cattle where a limited number of bulls were extensively used in breeding for fast genetic progress in economic traits like milk and protein yield (Charlier et al., 2016). An estimated yearly loss of ~$10.74 million is attributed to known recessive lethals in four dairy cattle breeds from USA only, where Holstein accounts for ~70% of the total losses, followed by Jersey, Brown Swiss, and Ayrshire (Cole et al., 2016). Hence, understanding the genomic architecture of cattle populations is important, now more than ever, for optimizing genetic gain while constraining negative impact of deleterious mutations responsible for genetic defects and inbreeding depression.

Unlike single nucleotide polymorphism (SNP) and small insertion or deletion (indel), structural variants (SVs), i.e. DNA alterations larger than 50 base pairs (bp) that include insertions, deletions, duplications, inversions, and translocations (Weischenfeldt et al., 2013), are the least explored polymorphisms in cattle. SVs contribute substantially to phenotypic variations and have a wide-spectrum of impact ranging from beneficial to lethal in both humans (Weischenfeldt et al., 2013, Zarrei et al., 2015) and animals (Bickhart and Liu, 2014). The phenotypic impact of SVs in cattle is well evident from numerous studies. For example, Xu et al. (Xu et al., 2014) showed that a combination of SNPs with SVs could explain additional genetic variance underlying milk production traits, while Charlier et al. (Charlier et al., 2012), Schutz et al. (Schutz et al., 2016), and Kadri et al. (Kadri et al., 2014) showed the lethal effect of large deletions in dairy cattle.

Furthermore, a ~525 KB deletion on chromosome 23 is reported to be associated with stillbirth in Nordic Red Cattle (Sahana et al., 2016).

Earlier SV studies on cattle were mostly SNP-array based, such as array-comparative genomic hybridization (Liu et al., 2010), 50K BovineSNP50 BeadChip (Hou et al., 2011) or 777K BovineHD BeadChip (BovineHD chip) (Xu et al., 2016) based. But, using these approaches a substantial portion of the genome could not be explored and breakpoint resolution is still an issue (Alkan et al., 2011). However, whole-genome sequence (WGS) based techniques could improve resolution as well as power to capture SVs in a wide size and frequency spectrum (Alkan et al., 2011). For example, majority of the novel SVs in humans (Mills et al., 2011, Sudmant et al., 2015) and mouse (Yalcin et al., 2011) were detected using WGS approaches. Besides, breakpoint sequences could also be assembled with high accuracy from sequencing reads (Chen et al., 2014), which are necessary for elucidating the mechanisms underlying SV formation (Carvalho and Lupski, 2016).

In the advent of next-generation sequencing (NGS) techniques, hundreds of cattle (bull or cow) genomes were re-sequenced in collaborative initiative such as 1000 Bull Genomes Project (1KBGP) (Daetwyler et al., 2014) (and other independent projects, e.g. (Jansen et al., 2013, Brondum et al., 2014)) to build a comprehensive database of sequence variants, mainly SNPs and indels. This NGS data provides a unique opportunity to study SVs in cattle. However, few studies (Boussaha et al., 2015, Chen et al., 2017) utilized these (and/or other) NGS resources so far for studying SVs in cattle.

Therefore, in this study we scanned the whole genome sequences of 175 cattle from three dairy breeds, namely Holstein, Jersey, and Nordic Red Cattle, to discover large deletions segregating in the population, and analyze their population-genetic properties. In particular, we focused on understating the population diversity, stratification, and plausible functional effects. We also explored the probable mechanisms of SV formation for a set of breakpoint-resolved deletions.

3.3 Materials and methods

Animal samples and ethics

This study was performed on whole-genome sequences (WGS) of 175 dairy cattle from three breeds, e.g. 67 Holstein, 27 Jersey, and 81 Nordic Red Cattle. The sample included 7 Holstein cows and 168 bulls from these three breeds − 144 animals from Run 5 of 1K bull genomes project (1KBGP) (Daetwyler et al., 2014) and 31 animals from Nordic sequence data (Brondum et al., 2014). Genome sequences were generated using Illumina paired-end sequencing to an average coverage of 10-fold.

Here, we did not include any experimentation on animals and only dealt with analysis-ready WGS data; hence, no ethical approval was required.

Sequence alignment to reference genome and SNPs/indels calling

Raw sequencing reads were filtered and *FASTQ* files were aligned to bovine reference genome assembly *UMD3.1* using *BWA* software (Li and Durbin, 2009) to produce BAM files for subsequent variant calling. In 1KBGP, SNPs and indels were called using *SAMtools 0.1.18 mpileup* software (Li et al., 2009), while *GATK v1.6* software (McKenna et al., 2010) was used for Nordic WGS data (detailed method in (Daetwyler et al., 2014) and (Brondum et al., 2014), respectively). For all the analysis, bovine genome assembly *UMD3.1* was used as the reference genome.

Discovery and genotyping of deletions

Structural variants can be detected from NGS data based on sequence signatures such as (discordant) read-pair (RP), split-read (SP) and read-depth (RD), as well as *de novo* assembly of reads (Alkan et al., 2011). However, approaches based on only one sequence signature could be constrained by high false discovery rate (FDR) (Handsaker et al., 2011), hence we employed a population scale SV detection method called "*Genome STRucture in Populations* (*Genome STRiP)"* (Handsaker et al., 2011) – which leverages technical (e.g. RP and RD signals) and population-level sequence features (e.g. coherence around shared alleles, and heterogeneity of evidentiary sequences in different genomes) for accurate discovery of deletions, and determines genotype (allelic state) of each locus from read-depth using a Gaussian mixture model.

Genome STRiP. For deletion discovery and genotyping *Genome STRiP* software version 2.00.1678 (Handsaker et al., 2011) was used. Following the documentation, we built a custom reference metadata bundle for cattle samples that includes alignability mask, copy-number mask (CN2 mask), ploidy map, gender map. Alignability mask represents sites on the reference genome that are uniquely alignable by sequence read of a certain length (readLength). Our WGS data was a mixture of different *Illumina* paired-end reads ranging from 90 to 101 bp (Q1 = 90, median = 100, and Q3 = 100), hence genome alignability mask was prepared with readLength value of 90 using '*ComputeGenomeMask*' utility from *Genome STRiP*. Copy-number mask (CN2 mask), i.e. regions on the reference genome unlikely to be copy-number variable in most individuals, was produced for the bovine assembly *UMD3.1* excluding sex chromosome X, unplaced contigs, and repeat sequences (retrieved from *RepeatMasker* track of *UCSC Table Browser* (Karolchik et al., 2004), accessed on July 4, 2016).

Preprocessing, deletion discovery and genotyping. We ran the preprocessing Queue script (dry run) to emit all the commands, prepared bash scripts to run in *Portable Batch System* job scheduler, and executed these commands proving 175 BAM files (one for each sample) as input.

Large deletions (100 bp ≤ size ≤ 1 MB) were discovered and filtered using *SVDiscovery Queue* script. Discovered sites were filtered (default filters) if (i) the site contained too high or too low read pileup, (ii) read-pairs spacing was inconsistent with a single segregating deletion, (iii) read-depth and read-pair evidences were inconsistent across samples, (iv) read depth differences were not significant, and (v) read

pair evidence was thinly distributed across samples (Genome STRiP Tutorial – GATK Workshop 2013, http://software.broadinstitute.org/software/genomestrip/workshop-presentations, accessed on August 26, 2016).

All passed sites were genotyped by *SVGenotyper* with default parameters. Genotyped deletion calls were then filtered based on following criteria, e.g. (i) sites with excess number of heterozygote calls (inbreeding coefficient ≤ -0.15), (ii) non-variant site based on genotype likelihood (parameter: non-variance score $\geq$ 13.0), (iii) sites with too low or too high read depth (parameter: $0.5 \geq$ GSM1 ≥ 2.0), (iv) sites with less than 30% uniquely alignable bases, (v) potential duplicate of another site (parameter: duplicateOverlapThreshold 0.5 & duplicateScoreThreshold ≥ 0.0), (vi) start/end position of a deletion call within 150 bp of assembly gap, (vii) all samples homozygous for reference allele (95% confidence), and (viii) sites with $\geq 10\%$ missing genotype.

Validation of deletions

Validation using 777K BovineHD BeadChip intensity data. We validated deletion calls using 777K BovineHD BeadChip (Illumina, San Diego, CA, USA) intensity data on 26 Holstein samples that were both whole-genome sequenced and 777K chip typed. We calculated false discovery rate (FDR) for the deletion call-set using *IntensityRankSum (IRS)* test implemented in *Genome STRiP*. Intensity file was prepared from raw chip intensity data following the guideline for IRS test. Overall FDR for the call-set was calculated as two times the fraction of sites with IRS *p-value* $\geq$0.5 (i.e. sites with IRS *p-value* $\geq$0.5 to the sites with valid *p-value*). Details of IRS test could be found in (Handsaker et al., 2011, Sudmant et al., 2015).

Validation by targeted assembly of breakpoint. Targeted iterative graph routing assembler (*TIGRA-0.4.3*) software (Chen et al., 2014) was used, with default parameters, for assembling deletion breakpoint sequences from a set of randomly selected deletions along with three previously known deletions segregating in the study populations. *TIGRA* extracted all reads mapped to 500 bp upstream and 50 bp downstream of start coordinate, and 50 bp upstream and 500 bp downstream of end coordinate of a given deletion; and reads were then assembled iteratively using *de Bruijn* graph assembler with multiple k-mers (e.g. 15 bp followed by 25 bp). We aligned the assembled contigs to *UMD3.1* using *Cow BLAT Search* (Kent, 2002) from *UCSC Genome Browser* (https://genome.ucsc.edu/cgi-bin/hgBlat) to visualize and infer breakpoints from the alignments.

Validation by PCR and amplicon sequencing. We validated a previously reported ~525 KB deletion segregating in Nordic Red Cattle (Sahana et al., 2016) using PCR and amplicon sequencing. Genomic DNA was extracted as described previously by Miller et al. (Miller et al., 1988) from semen sample of two bulls carrying the deletion and two non-carriers. The PCR reaction was done with the *DyNAzyme II DNA Polymerase* (Thermo Fisher, MA, US) in a 30µl volume of 1x PCR buffer, 0.2mM dNTPs, 10pmol primer mix (forward primer: 5′- AAGCCACCACAATGAGAAGC -3′ and reverse primer: 5′- TTTGGGGTAGGAGAAGTAGGG -3′) and 50 ng of genomic DNA. The cycling conditions were the

following: 1) an initial denaturation at 95°C for 3 min, 2) 35 cycles of 30 sec denaturation (94°C), 30 sec hybridization (65.2°C), 30 sec elongation (72°C) and a final 3 min elongation (72°C). PCR products were separated on a 2% agarose gel, purified and directly sequenced using the *BigDye Terminator Cycle Sequencing Kit* (Applied Biosystems, CA, US). Electrophoresis of sequencing reactions was performed on *3500xL Genetic Analyzers* (Applied Biosystems, CA, US), and sequences were visualized with *Sequencher 5.4.6* (Gene Codes Corporation, MI, US). A 977 bp control amplification, with a primer pair within the deletion (forward primer: 5′- CCCAATGCAAAATCACAAAA -3′ and reverse primer: 5′- CCAGAAAAGCTACACTTGAACTGA -3′), was performed using the same reaction conditions as above except hybridization was performed at 59.8°C.

Analysis of population genetic properties

The population genetic properties of deletions, among the three breeds, were studied in terms of population diversity, population structure, and population differentiation. Population diversity was calculated using *"VariantsPerSampleAnnotator"* from *Genome STRiP* software, which provides distribution of variants across samples and populations. We performed principal component analysis (PCA) using *PLINK* (v1.90p) software (Purcell et al., 2007) to distinguish three cattle breeds (details in Supplementary data). We calculated V_{ST} (Redon et al., 2006) – a population stratification measure of structural variants (highly correlated with Wright's fixation index, F_{ST} (Wright, 1931)), for each deletion locus using variant allele frequency (VAF) and genotypes from pairwise comparison of one breed with the rest, such as Holstein vs Jersey+Nordic Red Cattle, and vice versa.

Functional annotation and enrichment analysis

Functional annotation of deletions were performed using *Variant Effect Predictor (VEP-87)* software (McLaren et al., 2016), and enrichment of protein domains (*InterPro* (Finn et al., 2017) and *Pfam* (Finn et al., 2016)) and pathways (*KEGG* (Kanehisa et al., 2017)) were analyzed using *STRING-v10* database (Szklarczyk et al., 2015).

Selective constraints on genes were measured from the ratio of nonsynonymous (*dN*) to synonymous (*dS*) substitution rate, i.e. *dN/dS* ratio, between cow-mouse 1-to-1 orthologs downloaded from *Ensembl* database (Yates et al., 2016) (release 87, last accessed on February 21, 2017) using *BioMart* (Kinsella et al., 2011). Here we analyzed whether *dN/dS* of genes overlapping deletions are higher (i.e. less constrained) than that of mouse lethal genes (from Dickinson et al. (Dickinson et al., 2016)) using Wilcoxson test. Reported causal genes for cattle were also retrieved from OMIA database (http://omia.angis.org.au/, last accessed on May 10, 2016) for *dN/dS* comparison.

We retrieved cattle QTL from *QTLdb* database (Hu et al., 2016) (release 31; accessed on January 6, 2017); autosomal QTL from Holstein, Jersey, Nordic Red Cattle and Ayrshire, associated to any of the six trait classes, e.g. "Reproduction", "Milk", "Production", "Exterior", "Meat and Carcass", and "Health", were

considered for QTL enrichment analysis. We calculated fold enrichment for a trait, such as for Health related QTL: (*No. of Health QTL on Deletions / Total QTL on deletions*) / (*Total Health QTL / Total QTL in the dataset*), and statistical significance using *Fisher's exact test* (two sided).

Data manipulation, visualization and statistical analysis

All statistical analyses and plots were generated in *RStudio* software (RStudio Team, 2016) running *R* software version 3.3.2 (R Core Team, 2016), unless mentioned otherwise. *BEDTools* (v2.26.0) software (Quinlan and Hall, 2010) is used for identifying the overlap between deletion calls and other genomic features, such as, *UMD3.1* assembly gaps (from *UCSC Table Browser*), CNVs from *dbVar* database (Lappalainen et al., 2013), three known deletions from (Charlier et al., 2012, Kadri et al., 2014, Sahana et al., 2016), QTL from *QTLdb*. *VCFtools* (v0.1.15) software and *PLINK* (v1.90p) software were used for analyzing the VCF file.

3.4 Results and discussion

Discovery and genotyping of deletions

Deletion discovery and genotyping were carried out using *Genome STRiP*. After filtering, we report 8,480 large deletions with genotypes in 67 Holsten, 27 Jersey, and 81 Nordic Red Cattle. The deletion size ranged from 199 bp to 773 KB with a mean of 4.5 KB (median = 1 KB), which is approximately 10 times smaller compared to 184 deletion-CNVs (mean = 44.5 KB, median = 7.7 KB) reported in a recent 777K BovineHD BeadChip (BovineHD chip) based study (Xu et al., 2016), reflecting the resolution of our sequence-based calls. Only 18% of the deletion calls have overlap (≥1 base pair) with previously reported bovine deletions (or CNV-loss) in the *dbVar* database (accessed on January 27, 2017), while remaining 82% are novel. However, ~72% of our deletion regions remained unique when compared to all CNVs (gain or loss) and copy number variable regions (CNVRs) in the database. Interestingly, majority (~80%) of these overlapping regions are from an earlier WGS-based study, where genome sequences of 27 Holstein, 17 Montbéliarde and 18 Normande bulls were analyzed (Boussaha et al., 2015). Nonetheless, we were able to broaden the accessible deletion size-range, more importantly towards smaller one unascertainable by usual SNP-array based approaches. We also report high quality genotypes for all the 8,480 deletions. Apparently, there are more low frequency variants than that of high frequency one, and the frequency distribution is very similar to humans (Mills et al., 2011) (Figure 3.1).

Previous NGS-based studies on cattle were mostly limited to SV discovery, while copy-number states (genotypes) were inferred using BovineHD chip (Chen et al., 2017), or custom SNPs array (Boussaha et al., 2015). However, in this study we estimated the copy-number at each deletion locus (per sample) from read-depth within the region using a constrained Gaussian mixture model with three classes, e.g. copy-number zero (i.e. homozygous deletion), one (i.e. heterozygous deletion) and two (i.e. homozygous reference). It is known from human studies that majority of the (bi-allelic) common SVs segregate on specific SNP

haplotypes (McCarroll et al., 2008, Conrad et al., 2010), which could be imputed with high accuracy (Handsaker et al., 2011, Handsaker et al., 2015). Thus, this approach has the potential, albeit with large reference, for accurate haplotype phasing and imputation of SVs to large cohorts of low-density chip-typed animals with no additional cost.

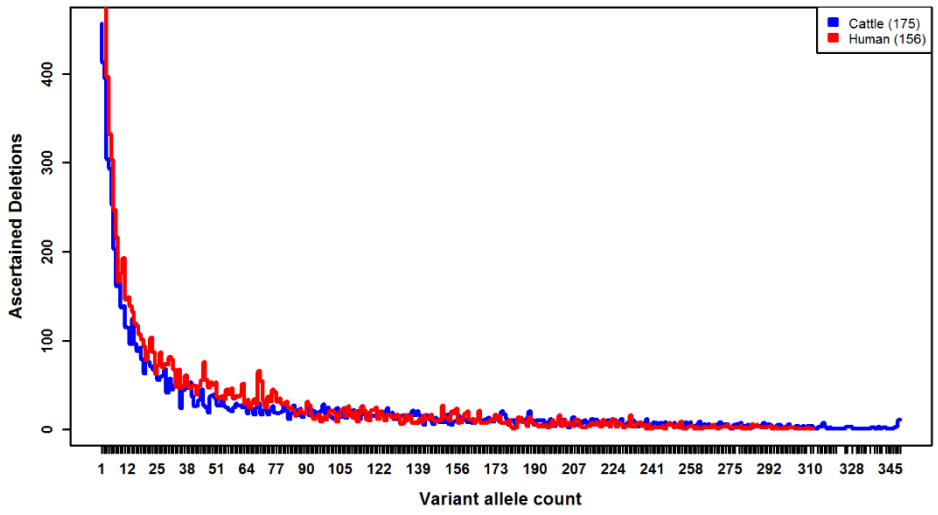


Figure 3.1. Number of ascertained deletions relative to variant allele count. *Here, variant allele frequency (VAF) is expressed in terms of variant allele count. Deletions down to an allele count of 1 (VAF=0.0026 and 0.0032, in cattle and humans, respectively) are also represented here. Human deletion calls by Mills et al.* ***(Mills et al., 2011)*** *were downloaded from 1K Genomes Project FTP server (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/paper_data_sets/companion_papers/mapping_structural_variation).*

Validation of deletions

We validated the results using three approaches: (i) using BovineHD chip intensity data, (ii) breakpoint assembly and alignment, and (iii) PCR + sequencing of amplicons.

BovineHD chip intensity. We used 777K BovineHD chip intensity data of 26 Holstein animals, both chip-typed and sequenced, to validate the deletion calls using *Genome STRiP' IntensityRankSum* test. We had partial power to investigate all deletions due to the sparsity of the array-probes (one probe per ~3.5 KB), and were underpowered to accurately verify small deletions (e.g. overlapping one-probe). Furthermore, we could only test a deletion for which at least one of the 26 samples had non-reference allele. Therefore, an estimate of false discovery rate (FDR) for the deletion call-set is provided here from the overall *p-value* distribution. In this approach, we were able to interrogate ~8.3% of the total call, majority of which contain a single array-probe within the region (Table 3.1 and Table S1). We found that deletions overlapping only one array-probe had higher FDR (11.3%) compared to two or more probes. And finally, we showed that our deletion call-set had an overall FDR of 8.8%, which is within our chosen threshold of FDR ≤ 10%.

Table 3.1. False discovery rate (FDR) estimates of Genome STRiP' deletion calls using 777K BovineHD BeadChip intensity data

Array-probe Overlap	A[¥]	B[§]	FDR*
One-probe	497	28	11.3%
>1 array-probe	206	3	2.9%
>2 array-probe	113	1	1.8%
Overall	**703**	**31**	**8.8%**

[¥]**A**=No. of sites with p-value; [§]**B**=No. of sites with p-value ≥ 0.5; *FDR estimates were based on *Wilcoxson rank sum test* using BovineHD chip intensity data of 26 Holstein animals. FDR was calculated as (**B/A ×2 ×100**).

Targeted breakpoint assembly. We next validated three randomly chosen set of ten-deletions each by assembling breakpoint sequences using TIGRA (Chen et al., 2014): 10 deletions ≤ 500 bp, 10 deletions > 500 bp but ≤1KB, and 10 deletions with VAF ≤ 0.10. Out of the thirty, we successfully resolved breakpoints of 26 deletions (~87% success rate) using a combination of TIGRA and BLAT search (Kent, 2002) (Table S2). Additionally, we assembled breakpoints of three previously reported deletions that were also segregating in our study populations, such as a ~662 KB deletion on chromosome 12 encompassing *RNASEH2B*, *GUCY1B2*, and *FAM124A*, a ~3 KB deletion on chromosome 21 encompassing *FANCI*, and a ~525 KB deletion on chromosome 23 encompassing *BTBD9*, *GLO1*, and *DNAH8*. Breakpoints of the former two deletions were previously reported (Charlier et al., 2012, Kadri et al., 2014), which exactly matched with our predicted breakpoint sequences (Figures S1abc and S2abc). While for the later, we resolved breakpoint sequences in this study (Figure 3.2ab). Overall, the success rate of our deletion-breakpoint assembly was better than the reported success rate of TIGRA on similar sized read-length (Chen et al., 2014). And *Genome STRiP*'s breakpoint predictions were on average within 20 bp of the validated breakpoint, which is within the tool's reported estimate of 1-20 bp (Handsaker et al., 2011).

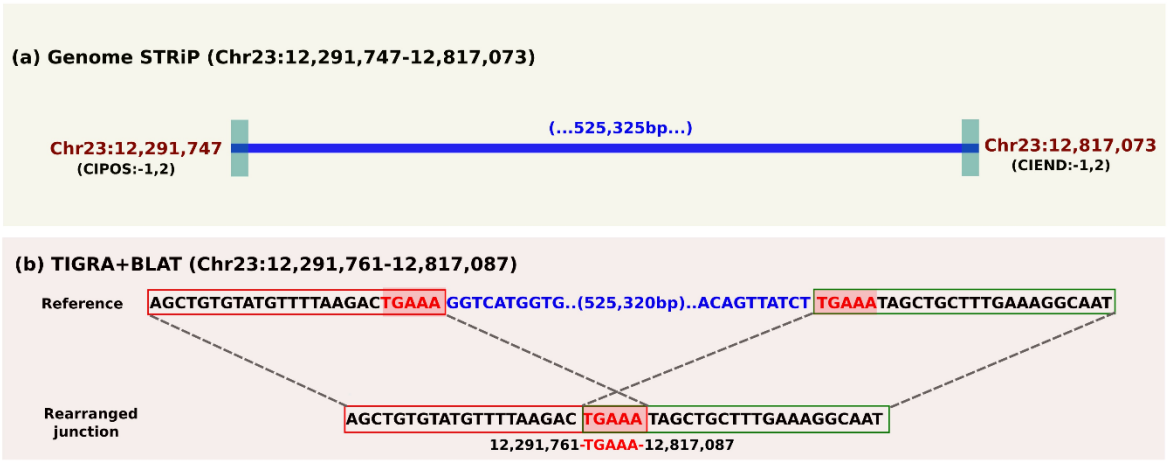


Figure 3.2. A ~525-KB deletion on chromosome 23 discovered using Genome STRiP (a), and resolved breakpoint sequences from TIGRA and BLAT search (b). *Shaded bases are a 5-bp microhomology at breakpoint junction. (This figure was drawn and modified using Inkscape version 0.91.)*

PCR and amplicon sequencing. We then experimentally validated breakpoints for "Chr23:12,291,761-12,817,087" deletion, previously reported to be associated with stillbirth in Nordic Red Cattle (Sahana et al.,

2016). Four animals were used for PCR validation: 2 carriers and 2 non-carriers. The breakpoint spanning PCR products of 359 bp were only observed in the carrier animals, while no amplicon was seen for non-carriers (Figure 3.3a). The 359 bp amplicon was then sequenced, and exact breakpoint sequences were observed (Figure 3.3b), thus confirming the breakpoint for this deletion.

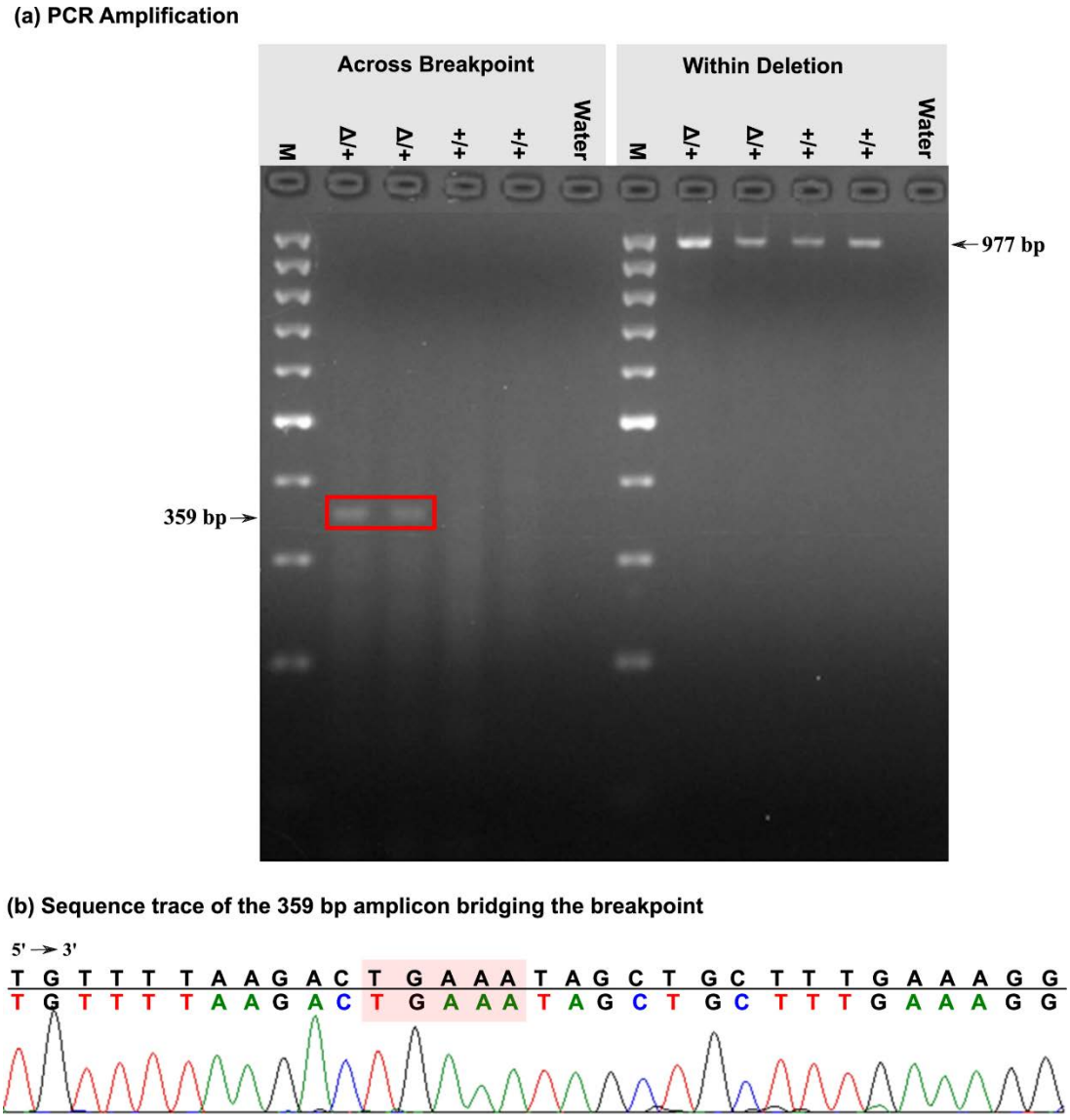


Figure 3.3. Experimental validation of the ~525 KB deletion on chromosome 23. *(a) PCR amplification across (left) and within the deletion (right) for two carrier (Δ/+) and two homozygous wild-type (+/+) animals. Water: negative control, M: molecular weight marker (GeneRuler™ 100bp DNA ladder, Fermentas). (b) Sequence trace of the 359 bp amplicon bridging the breakpoint. Shaded bases are a 5-bp microhomology at breakpoint junction. (This figure was drawn and modified using Inkscape version 0.91.)*

Population genetic properties of deletions

Population diversity. We explored population diversity among the three dairy cattle breeds from per-individual deletion-heterozygosity and homozygosity. We found that individuals from Nordic Red Cattle exhibits 3.5% and 6.4% higher deletion-heterozygosity than in Holstein and Jersey, respectively. Median numbers of heterozygote-deletion were 1272, 1229 and 1196, in Nordic Red Cattle, Holstein, and Jersey,

respectively (Figure 3.4a). In contrast to heterozygosity, Jersey animals showed highest levels of deletion-homozygosity, followed by Holstein (Figure 3.4b). Similar estimates of genetic diversity were also reported for these breeds in a SNP heterozygosity and runs-of-homozygosity analysis – where Jersey exhibited lowest (genome-wide) average nucleotide diversity (and higher number/size of runs-of-homozygosity) followed by Holstein and Nordic Red Cattle (Zhang et al., 2015). These differences could be understood from the current effective population size (N_e) of these breeds, e.g. N_e of Jersey, Holstein, and Nordic Red Cattle are 73, 99 and 106, respectively (Bovine HapMap et al., 2009); this entailed higher diversity in Nordic Red Cattle, and Holstein, than in Jersey. From singletons estimate it is also evident that Nordic Red Cattle has more rare deletions compared to Holstein and Jersey (Figure 3.4c); partly could be due to incorrect ascertainment of rare ones.

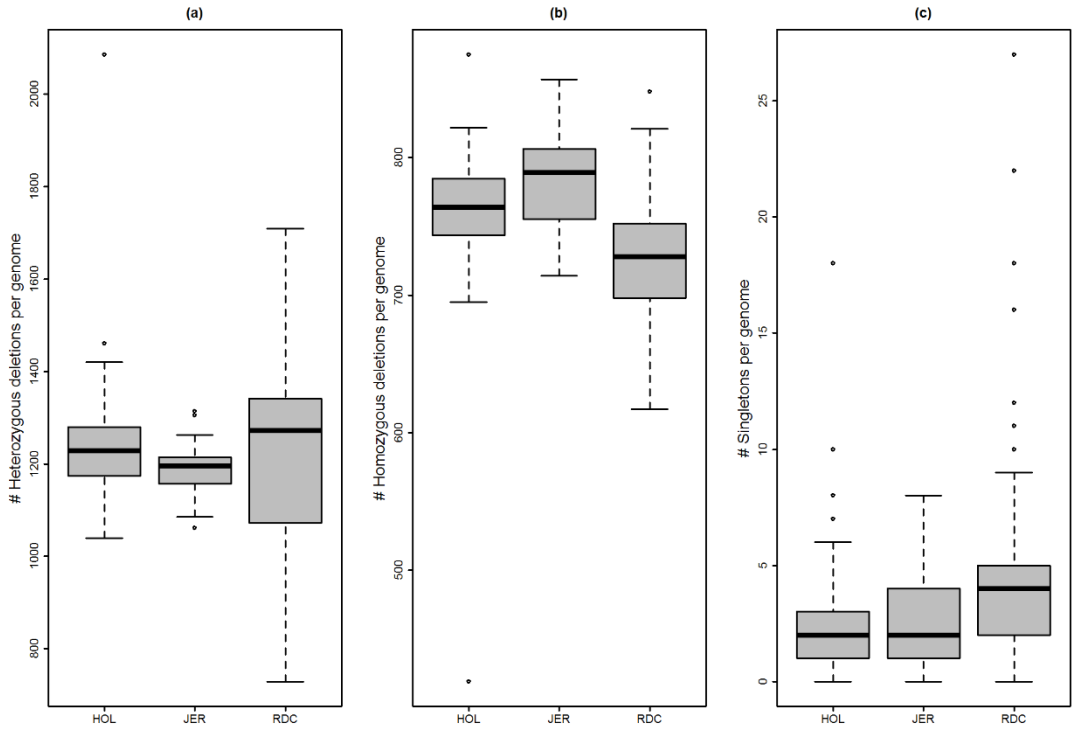


Figure 3.4. Population diversity. *(a) Heterozygous deletions per genome. (b) Homozygous deletions per genome. (c) Singletons per genome. Only high-confidence genotype calls are included. The y-axis in (a), (b), and (c) represents the number of heterozygous, homozygous, and singleton deletions per genome, respectively. HOL: Holstein; JER: Jersey; RDC: Nordic Red Cattle.*

Principal Component Analysis (PCA). We performed Principal Component Analysis (PCA) using the deletion genotypes of the samples. Around 6K deletions with VAF between 0.02 and 0.90 were used in the analysis. For comparison, we also performed PCA on ~168K bi-allelic SNPs randomly selected from 29 autosomes of the same individuals. The first two principal components (PCs) from both deletion and SNP-based PCA clearly distinguished the three breeds, and jointly explained 20% and 16.2% of the variance, respectively (Figure 3.5ab). In addition, PC3 and PC4 recapitulated substructures within Nordic Red cattle (Figure S3), and first five PCs cumulatively explained 33.6% (with the deletions) and 28.4% (with the SNPs) of the variance (Figure S4). Our deletion results agree with the known population structure of the three

breeds. Similar population structure (and substructure within Nordic Red Cattle) has been reported using genome-wide SNPs (Mao et al., 2016). Nordic Red Cattle from Denmark showed closer relationship with the Holstein in our WGS samples (Figure S5). This is consistent with the known history of Holstein interrogation in Danish Red cattle, as previously reported based on imputed WGS SNP analysis on a larger sample (Mao et al., 2016). It is also known from admixture analysis that genomes of Nordic Red Cattle are a mosaic of multiple ancestral populations, i.e. more ancestral components in Nordic Red Cattle than in Holstein and Jersey (Bovine HapMap et al., 2009, Brondum et al., 2011); our deletion-based PCA largely corroborate that.

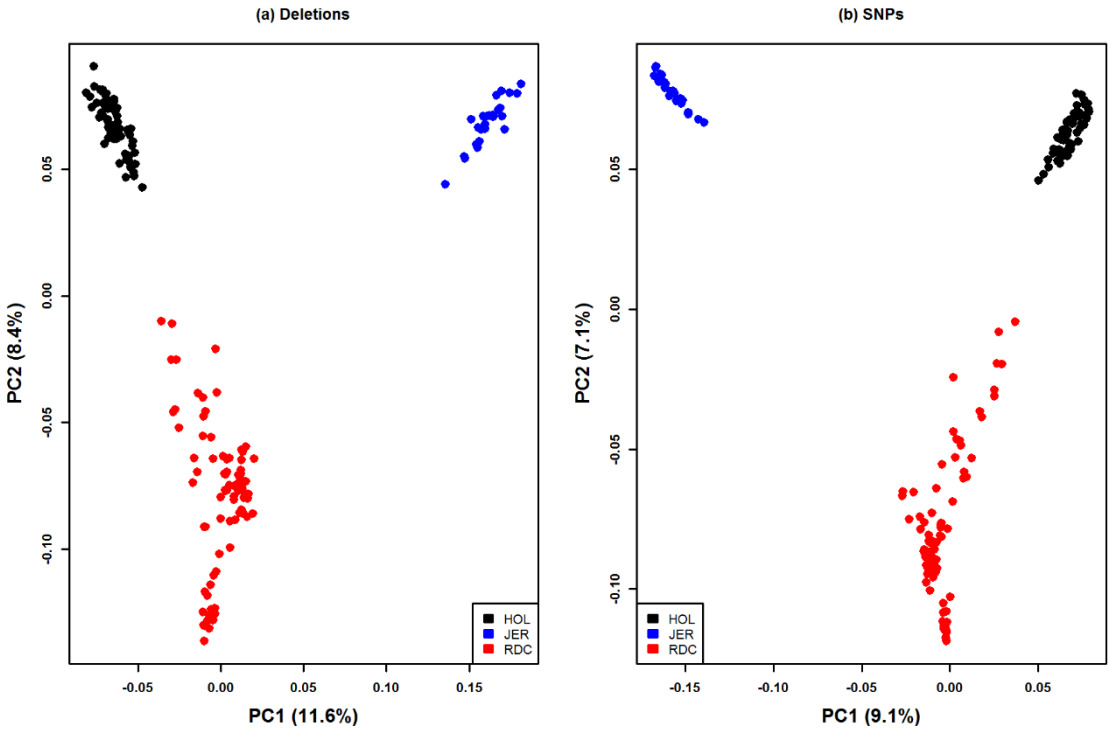


Figure 3.5. Principal Component Analysis depicting three dairy cattle breeds. *The analysis is based on (a) ~6K deletions (0.02<VAF<0.9), and (b) ~168K bi-allelic SNPs randomly selected from 29 bovine autosomes. First two principal components (PCs) from deletions and SNPs are plotted here; jointly explained 20% and 16.2% of the variance, respectively. HOL: Holstein; JER: Jersey; RDC: Nordic Red Cattle.*

***V_{ST}* analysis.** We analyzed population stratification in terms of V_{ST} (Redon et al., 2006), a measure highly correlated with Wright's fixation index (F_{ST}) (Wright, 1931), to identify population differentiation. We calculated V_{ST} for each deletion pairwise amongst the breeds (e.g. Holstein vs Jersey+Nordic Red Cattle) from VAFs. We identified 158 highly stratified deletions (pairwise $V_{ST} \geq$ mean + 4 standard deviations) among the breeds (Figure 3.6). Around 27% of these deletions overlap genic elements, i.e. exons, introns, or (upstream/downstream) untranslated regions (UTRs), and remaining 73% are intergenic variants (Tables S3-5). There were eleven sites shared between Holstein and Nordic Red Cattle, two sites between Holstein and Jersey, and one site between Nordic Red Cattle and Jersey.

Among these sites were gene variants, such as *ABCA12* (Chr2:103,682,772-103,684,297 in Holstein & Jersey) associated with growth and development (Cole et al., 2014, Xu et al., 2015), *TTC1* (Chr7:73,725,513-

73,725,918 in Holstein & Nordic Red Cattle) with cold tolerance (Howard et al., 2014), *VWA3B* (Chr11:3,521,329-3,522,551 in Holstein & Nordic Red Cattle) with milk glycosylated kappa-casein percentage (Buitenhuis et al., 2016), and were intergenic variants, such as Chr15:41,393,393-41,393,780 (in Jersey & Nordic Red Cattle) and Chr20:26,812,159-26,812,834 (in Holstein & Jersey) overlap QTL associated with calving traits (McClure et al., 2010, Sahana et al., 2011), Chr20:45,816,245-45,820,519 (in Holstein & Nordic Red Cattle) with meat and carcass trait (McClure et al., 2012), and Chr23:49,778,653-49,782,567 (in Holstein & Nordic Red Cattle) with body weight (Snelling et al., 2010). We also identified a highly differentiated fertility associated gene *DST/BPAG1* (Lobago et al., 2006, Cole et al., 2011) (Chr23:3,486,232-3,486,603) in Nordic Red Cattle. One differentially selected deletion (V_{ST}=0.28) of chromosome 3 (Chr3:12,141,822-12,170,916) overlapping *ENSBTAG00000047776* and *ENSBTAG00000024960* genes (human ortholog *CD1D*), drawn our attention (though it marginally failed our selection threshold); Holsteins exhibited VAF of 24.63% (have both homozygous and heterozygous deletion), while it is mostly homozygous for the reference allele in Nordic Red cattle (VAF=0.62%) and Jersey (VAF=0.0%). The *CD1D* gene has known function in host immune response and parasite resistance (Araujo et al., 2009, Sandri et al., 2015), and also reported differentially expressed post intra-mammary infection (Fang et al., 2017). Majority of these stratified deletion regions are novel compared to previous CNV studies in cattle, and therefore are interesting targets to investigate large deletions undergoing genetic drift or artificial selection.

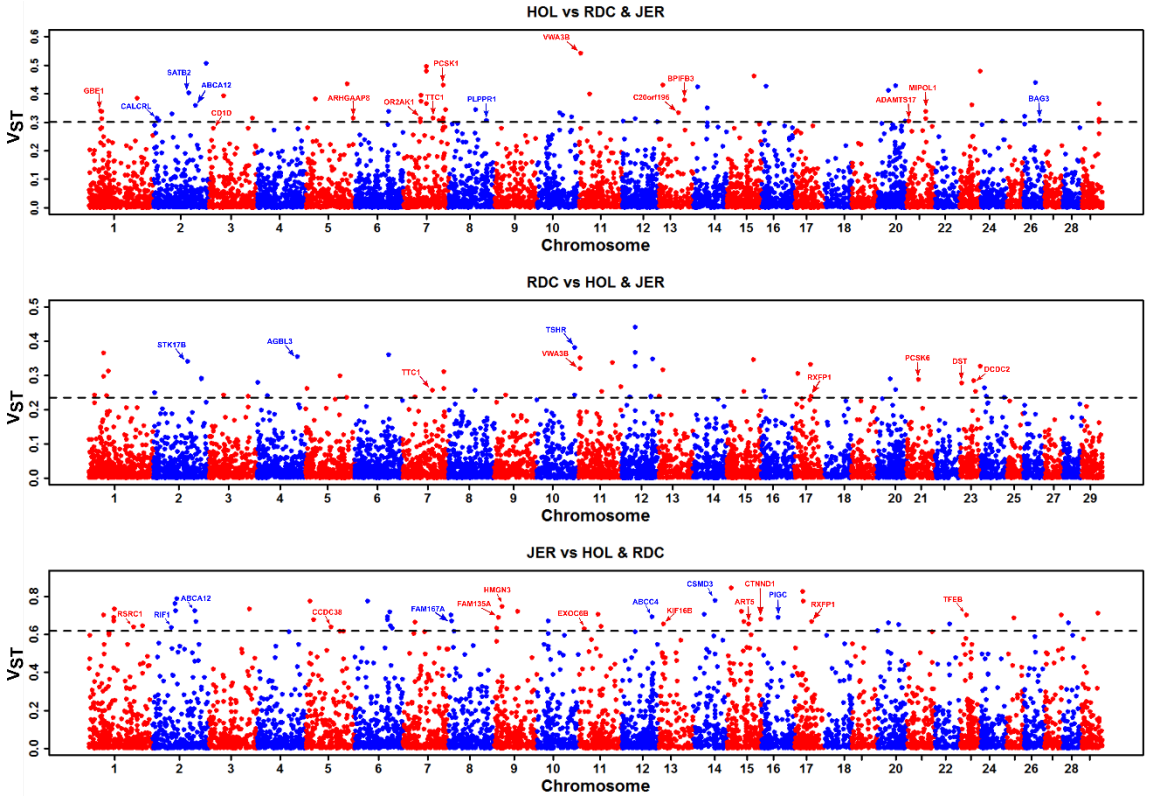


Figure 3.6. Population stratification based on V_{ST} (a measure of differentiation for structural variant, highly correlated with Wright's fixation index, F_{ST}). *Horizontal dash line indicates highly stratified deletion regions ($V_{ST} \geq$ Mean+4 Standard deviations). Highly stratified genic-deletions, e.g. overlapping exons, introns, or untranslated regions (UTRs), are highlighted with HGNC gene symbol.*

Functional impact of deletions

We annotated all the deletions using *Variant Effect Predictor (Ensembl 87)*. Around 71% (6,019 SVs) variants were intergenic and remaining 29% (2,461 SVs) overlapped genic elements, such as exons, introns, and untranslated regions (UTR). On average, high frequency gene disrupting deletions were somewhat depleted compared to intergenic variants ($VAF_{intergenic} > VAF_{genic}$, *p-value*=0.04; one-sided Wilcoxon test). Furthermore, we observed many common genic deletions. These genes are relatively less conserved, and majority has multiple paralogs (discussed later). However, deletions on known essential genes were only observed as heterozygote with relatively low VAF (<3%), and generally were private to a specific breed. For example, *FANCI* deletions (cause brachyspina (Charlier et al., 2012)) were only observed in Holstein, and *RNASEH2B* deletions (cause embryonic lethality (Kadri et al., 2014)) in Nordic Red Cattle.

Selective constraints on genes overlapping deletions. The relative abundance of high frequency genic and intergenic variants indicate that majority of these intersected genes are non-essential, and thus did not affect the viability or fecundity of the carriers. To test this hypothesis, we analyzed the selective constraints between deleted genes (overlap of any genic element) and known mouse lethal genes (from Dickinson et al. (Dickinson et al., 2016)) in terms of *dN/dS* ratio of cow-mouse 1-to-1 orthologs (Figure 3.7). Here, high *dN/dS* values indicate low selective constraints on genes, and low value indicates high constraints. We found that genes in deletions have significantly higher *dN/dS* ratios than lethal genes (*p-value* 2.3×10^{-6}; one-sided Wilcoxon test), and thus are evolutionarily less conserved. This is consistent with the rate of evolution seen in essential and non-essential genes – where mutations in essential genes were under strong purifying selection and thus evolved slowly (low *dN/dS* ratio), while non-essential genes were under relaxed selection, and hence, evolved faster (high *dN/dS* ratio) (Hurst and Smith, 1999). Nonetheless, robustness of these processes is also evident in the evolution of human essential genes. Interestingly, ~77% human essential genes could even be traced back to pre-metazoans (Blomen et al., 2015).

Nonessential genes in cattle. In total, we found 5,000 deletions for which at least one individual was homozygous. In the set, we analyzed homozygous deletions in genes to find natural gene knockouts. We found 167 deleted genes (transcript-ablation or complete deletion) corresponding to 115 independent deletions that are apparently nonessential based on the occurrences of live homozygote individuals. This is ~45% more than the previous report (Boussaha et al., 2015). Nonetheless, we found ~44% fewer genes compared to in humans (240 nonessential genes) (Sudmant et al., 2015), which could be due to the differences in sample size (175 vs 2,504 individuals) and study populations (3 vs 26 populations in human). Among these genes, ~83% (139 genes) are protein-coding, 12% pseudogenes, and the rest are different types of small RNAs (Table S6). Most of these genes belong to multigenic families and are not highly conserved (median cow-mouse *dN/dS* of 0.17 vs OMIA genes *dN/dS* of 0.11; Figure S6), as expected for homozygous deletion (Sudmant et al., 2015). Moreover, this set of genes are functionally enriched in immunoglobulin domains, olfactory receptors, and MHC classes (FDR = 2.06×10^{-22}, 2.06×10^{-22}, 7.01×10^{-6}, respectively), along with other related domains (Tables S7-9). Similar functional enrichment of nonessential genes was also seen in humans (Sudmant et al., 2015). Olfactory receptor related genes are well known for extensive

gains and losses in mammalian evolution (Niimura and Nei, 2007). And population specific copy-number variations of olfactory receptor genes were also reported in human (deletions) (Van Ziffle et al., 2011) and cattle (gains) (Lee et al., 2013). Nevertheless, this is the first report, to our knowledge, of homozygous deletion of olfactory receptor genes in cattle.

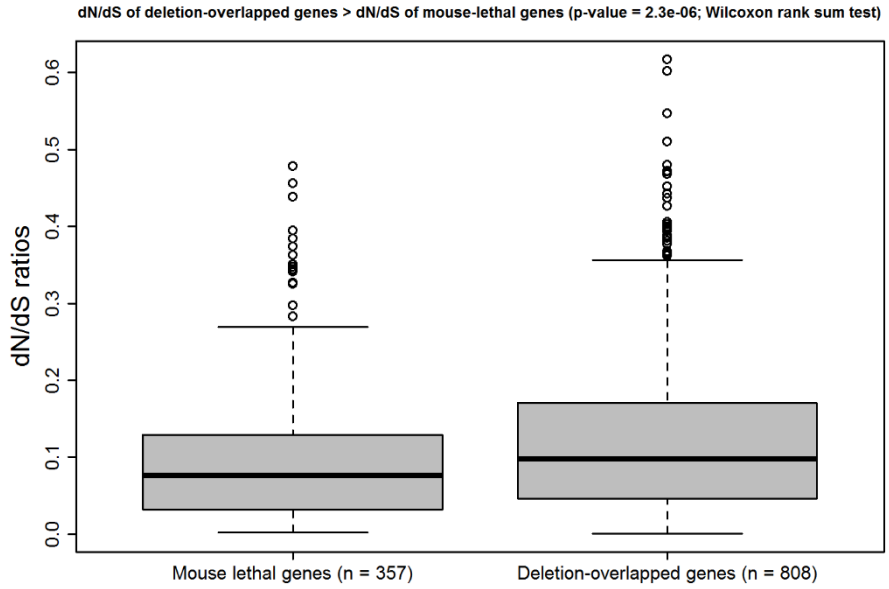


Figure 3.7. Difference between dN/dS ratios of mouse-lethal and deletion-overlapped genes in cattle. *Cow genes for which one-to-one mouse orthologs available were considered for a one-sided Wilcoxon rank-sum test. Mouse lethal genes are from Dickinson et al. (2016).*

QTL Enrichment. We next explored the enrichment (or depletion) of quantitative trait loci (QTL) on deleted regions (at least 1 bp overlap with deletion). We retrieved ~24K autosomal QTL from *QTLdb* reported to be associated with any of the six trait classes, e.g. "Health", "Reproduction", "Milk", "Exterior", "Production" and "Meat and Carcass". The association of deletions with diseases, fitness or fertility related traits is well evident (Weischenfeldt et al., 2013). Hence, we suspected enrichment of fitness and fertility related traits for our deletions. As expected, health (2 fold) and reproduction (1.5 fold) related QTL were significantly enriched, while other trait classes were highly depleted (Table 3.2). Higher enrichment of health related QTL could be driven by immune-system genes, which were also highly enriched in our dataset (discussed earlier).

Table 3.2. Enrichment of QTL on deletions

Trait Classes*	**Fold Enrichment**	***P value* (Fisher's test)**
Health	**2**	8.91×10^{-10}
Reproduction	**1.5**	7.4×10^{-11}
Milk	0.8	2.45×10^{-7}
Exterior	0.5	1.85×10^{-4}
Production	0.5	0.002
Meat and Carcass	0.5	0.058

*Trait classes are from cattleQTLdb (Hu et al., 2016). QTL from autosomes of Holsteins, Jersey, Nordic Red Cattle, and Ayrshires were considered for Fisher's exact test (two-sided).

Deletion formation mechanisms

Finally we explored the probable mechanisms of deletion formation. There are two key mechanisms of structural variants formation (for detail see review (Hastings et al., 2009, Carvalho and Lupski, 2016)); for example, recurrent SVs often result from *non-allelic homologous recombination* (NAHR) between large low-copy repeats (LCRs), and thus, contain extensive sequence homology provided by LCRs, such as segmental duplicates, at the flanking regions (Carvalho and Lupski, 2016). In contrast, non-recurrent SVs often form either by *microhomology-mediated end joining* (MMEJ) or *non-homologus end joining* (NHEJ), which requires limited to no sequence homology, and thus could be characterized by microhomologies or simple blunt ends at the breakpoint junction (Hastings et al., 2009).

Breakpoint information is crucial for understanding the mechanism, and therefore, we analyzed 29 breakpoint resolved deletions from our validation set. We found that 24 of 29 deletions contain microhomology ranging from 2-31 bp at the breakpoint, and two of which also contain insertions (S2 Table). In addition, 4 deletions exhibited non-reference insertion at breakpoint junctions, and one deletion with no apparent homology. However, the number of breakpoint sequences analyzed here were not a robust representation of our deletion call-set (less than 0.5% deletions), though selected randomly (for validation), we were able to demonstrate that majority of deletions contain microhomoloy at breakpoint, followed by few insertions, and rarely with no homology. Our results largely agree with the trend reported for large deletions in humans, e.g. 70.8% deletions exhibited microhomology/homology and 16.1% insertions at the breakpoint (Mills et al., 2011).

Limitations

This study only focused on identifying deletions in cattle because of their potential relevance to loss-of-function and embryonic lethality. However, we had limited success to identify small deletions, such as <200 bp due to reduced sensitivity of the SV caller. It is also not a comprehensive list of deletions for these samples, since we could have missed many true deletions due to sensitivity, coverage, or stringent filtering (among other reasons). Furthermore, the short read length (~100 bp) in our WGS dataset also made it difficult to resolve breakpoints from regions of long repeats.

3.5 Conclusions

Loss-of-function variants are responsible for a substantial yearly-economic loss in dairy industry, where a limited number of elite sires are in extensive use for rapid genetic gains. Mapping of such variants is essential for effective breeding planning and genomic selection. Here we showed an NGS-based analytical framework suitable for population-scale mapping of large deletions in cattle, leveraging the available WGSs. Here we described population-genetic, functional, and evolutionary properties of discovered deletions. We identified and confirmed a ~525 KB deletion on chromosome 23, causing stillbirth in Nordic Red Cattle. We demonstrated that Nordic Red Cattle had higher population diversity than Holstein and Jersey, and deletion-genotype could recapitulate genetic structure of these breeds. Natural gene knockouts are enriched for immune-related and olfactory receptor genes. We also showed that deletions are significantly enriched for

health and fertility related QTL, while depleted for production related QTL. Our population genetic and functional analysis showed promise for inclusion of SVs in genomic studies in dairy cattle. This deletion catalog will facilitate discovery, genotyping, and imputation of deletions in large cohorts of animals, and subsequent studies for gene mapping and genomic prediction of breeding values.

Authors' contributions

M.M-U. and G.S. conceived and designed the study. M.M-U. and B.G. performed in silico prediction and computational analyses. T.I-T performed experimental validation by PCR and amplicon sequencing, and wrote this method. J.V., D-J.D.K. and M.S.L. collected samples and generated sequence data. B.G., D.B., M.S.L. and G.S. supervised the study. M.M-U. wrote first draft of the manuscript. B.G., T.I-T., J.V., D-J.D.K., D.B., M.S.L. and G.S critically revised the manuscript for important intellectual content. All authors read and approved final version of the manuscript.

Acknowledgements

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

Data availability

All relevant results are within the paper and its Supplementary data files. VCF file with deletion calls could be found at https://github.com/MMesbahU/Deletions_in_cattle. Whole genome sequences of 44 samples (out of 175) are available from the NCBI Sequence Read Archive (project accession numbers SRP039339 and SRP065105). Among the 175 samples, 144 are from Run 6 of 1K bull genomes project. Rest of data are available only upon agreement with the commercial breeding organization and should be requested directly from the senior author (GS: goutam.sahana@mbg.au.dk) or the Center Director (MSL: mogens.lund@mbg.au.dk).

Supplementary data are available at DNARES online

3.6 References

Alkan, C., B. P. Coe, and E. E. Eichler. 2011. Genome structural variation discovery and genotyping. Nat Rev Genet 12:363-376. https://doi.org/10.1038/nrg2958.

Araujo, R. N., T. Padilha, D. Zarlenga, T. Sonstegard, E. E. Connor, C. Van Tassel, W. S. Lima, E. Nascimento, and L. C. Gasbarre. 2009. Use of a candidate gene array to delineate gene expression patterns in cattle selected for resistance or susceptibility to intestinal nematodes. Vet Parasitol 162:106-115. https://doi.org/10.1016/j.vetpar.2008.12.017.

Bickhart, D. M. and G. E. Liu. 2014. The challenges and importance of structural variation detection in livestock. Front Genet 5:37. https://doi.org/10.3389/fgene.2014.00037.

Blomen, V. A., P. Majek, L. T. Jae, J. W. Bigenzahn, J. Nieuwenhuis, J. Staring, R. Sacco, F. R. van Diemen, N. Olk, A. Stukalov, C. Marceau, H. Janssen, J. E. Carette, K. L. Bennett, J. Colinge, G. Superti-Furga, and T. R. Brummelkamp. 2015. Gene essentiality and synthetic lethality in haploid human cells. Science 350:1092-1096. https://doi.org/10.1126/science.aac7557.

Boussaha, M., D. Esquerre, J. Barbieri, A. Djari, A. Pinton, R. Letaief, G. Salin, F. Escudie, A. Roulet, S. Fritz, F. Samson, C. Grohs, M. Bernard, C. Klopp, D. Boichard, and D. Rocha. 2015. Genome-Wide Study of Structural Variants in Bovine Holstein, Montbeliarde and Normande Dairy Breeds. PLoS One 10:e0135931. https://doi.org/10.1371/journal.pone.0135931.

Bovine HapMap, C., R. A. Gibbs, J. F. Taylor, C. P. Van Tassell, W. Barendse, K. A. Eversole, C. A. Gill, R. D. Green, D. L. Hamernik, S. M. Kappes, S. Lien, L. K. Matukumalli, J. C. McEwan, L. V. Nazareth, R. D. Schnabel, G. M. Weinstock, D. A. Wheeler, P. Ajmone-Marsan, P. J. Boettcher, A. R. Caetano, J. F. Garcia, O. Hanotte, P. Mariani, L. C. Skow, T. S. Sonstegard, J. L. Williams, B. Diallo, L. Hailemariam, M. L. Martinez, C. A. Morris, L. O. Silva, R. J. Spelman, W. Mulatu, K. Zhao, C. A. Abbey, M. Agaba, F. R. Araujo, R. J. Bunch, J. Burton, C. Gorni, H. Olivier, B. E. Harrison, B. Luff, M. A. Machado, J. Mwakaya, G. Plastow, W. Sim, T. Smith, M. B. Thomas, A. Valentini, P. Williams, J. Womack, J. A. Woolliams, Y. Liu, X. Qin, K. C. Worley, C. Gao, H. Jiang, S. S. Moore, Y. Ren, X. Z. Song, C. D. Bustamante, R. D. Hernandez, D. M. Muzny, S. Patil, A. San Lucas, Q. Fu, M. P. Kent, R. Vega, A. Matukumalli, S. McWilliam, G. Sclep, K. Bryc, J. Choi, H. Gao, J. J. Grefenstette, B. Murdoch, A. Stella, R. Villa-Angulo, M. Wright, J. Aerts, O. Jann, R. Negrini, M. E. Goddard, B. J. Hayes, D. G. Bradley, M. Barbosa da Silva, L. P. Lau, G. E. Liu, D. J. Lynn, F. Panzitta, and K. G. Dodds. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324:528-532. https://doi.org/10.1126/science.1167936.

Brondum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15:728. https://doi.org/10.1186/1471-2164-15-728.

Brondum, R. F., E. Rius-Vilarrasa, I. Stranden, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. J Dairy Sci 94:4700-4707. https://doi.org/10.3168/jds.2010-3765.

Buitenhuis, B., N. A. Poulsen, G. Gebreyesus, and L. B. Larsen. 2016. Estimation of genetic parameters and detection of chromosomal regions affecting the major milk proteins and their post translational modifications in Danish Holstein and Danish Jersey cattle. BMC Genet 17:114. https://doi.org/10.1186/s12863-016-0421-2.

Carvalho, C. M. and J. R. Lupski. 2016. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet 17:224-238. https://doi.org/10.1038/nrg.2015.25.

Charlier, C., J. S. Agerholm, W. Coppieters, P. Karlskov-Mortensen, W. Li, G. de Jong, C. Fasquelle, L. Karim, S. Cirera, N. Cambisano, N. Ahariz, E. Mullaart, M. Georges, and M. Fredholm. 2012. A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina. PLoS One 7:e43085. https://doi.org/10.1371/journal.pone.0043085.

Charlier, C., W. Li, C. Harland, M. Littlejohn, W. Coppieters, F. Creagh, S. Davis, T. Druet, P. Faux, F. Guillaume, L. Karim, M. Keehan, N. K. Kadri, N. Tamma, R. Spelman, and M. Georges. 2016. NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. Genome Res 26:1333-1341. https://doi.org/10.1101/gr.207076.116.

Chen, K., L. Chen, X. Fan, J. Wallis, L. Ding, and G. Weinstock. 2014. TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. Genome Res 24:310-317. https://doi.org/10.1101/gr.162883.113.

Chen, L., A. J. Chamberlain, C. M. Reich, H. D. Daetwyler, and B. J. Hayes. 2017. Detection and validation of structural variations in bovine whole-genome sequence data. Genetics Selection Evolution 49:13. https://doi.org/10.1186/s12711-017-0286-5.

Cole, J. B., D. J. Null, and P. M. VanRaden. 2016. Phenotypic and genetic effects of recessive haplotypes on yield, longevity, and fertility. J Dairy Sci 99:7274-7288. https://doi.org/10.3168/jds.2015-10777.

Cole, J. B., B. Waurich, M. Wensch-Dorendorf, D. M. Bickhart, and H. H. Swalve. 2014. A genome-wide association study of calf birth weight in Holstein cattle using single nucleotide polymorphisms and phenotypes predicted from auxiliary traits. J Dairy Sci 97:3156-3172. https://doi.org/10.3168/jds.2013-7409.

Cole, J. B., G. R. Wiggans, L. Ma, T. S. Sonstegard, T. J. Lawlor, Jr., B. A. Crooker, C. P. Van Tassell, J. Yang, S. Wang, L. K. Matukumalli, and Y. Da. 2011. Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. BMC Genomics 12:408. https://doi.org/10.1186/1471-2164-12-408.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Wellcome Trust Case Control, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704-712. https://doi.org/10.1038/nature08516.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brondum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46:858-865. https://doi.org/10.1038/ng.3034.

Dickinson, M. E., A. M. Flenniken, X. Ji, L. Teboul, M. D. Wong, J. K. White, T. F. Meehan, W. J. Weninger, H. Westerberg, H. Adissu, C. N. Baker, L. Bower, J. M. Brown, L. B. Caddle, F. Chiani, D. Clary, J. Cleak, M. J. Daly, J. M. Denegre, B. Doe, M. E. Dolan, S. M. Edie, H. Fuchs, V. Gailus-Durner, A. Galli, A. Gambadoro, J. Gallegos, S. Guo, N. R. Horner, C. W.

Hsu, S. J. Johnson, S. Kalaga, L. C. Keith, L. Lanoue, T. N. Lawson, M. Lek, M. Mark, S. Marschall, J. Mason, M. L. McElwee, S. Newbigging, L. M. Nutter, K. A. Peterson, R. Ramirez-Solis, D. J. Rowland, E. Ryder, K. E. Samocha, J. R. Seavitt, M. Selloum, Z. Szoke-Kovacs, M. Tamura, A. G. Trainor, I. Tudose, S. Wakana, J. Warren, O. Wendling, D. B. West, L. Wong, A. Yoshiki, C. International Mouse Phenotyping, L. Jackson, I. C. d. l. S. Infrastructure Nationale Phenomin, L. Charles River, M. R. C. Harwell, P. Toronto Centre for, I. Wellcome Trust Sanger, R. B. Center, D. G. MacArthur, G. P. Tocchini-Valentini, X. Gao, P. Flicek, A. Bradley, W. C. Skarnes, M. J. Justice, H. E. Parkinson, M. Moore, S. Wells, R. E. Braun, K. L. Svenson, M. H. de Angelis, Y. Herault, T. Mohun, A. M. Mallon, R. M. Henkelman, S. D. Brown, D. J. Adams, K. C. Lloyd, C. McKerlie, A. L. Beaudet, M. Bucan, and S. A. Murray. 2016. High-throughput discovery of novel developmental phenotypes. Nature 537:508-514. https://doi.org/10.1038/nature19356.

Fang, L., G. Sahana, G. Su, Y. Yu, S. Zhang, M. S. Lund, and P. Sorensen. 2017. Integrating Sequence-based GWAS and RNA-Seq Provides Novel Insights into the Genetic Basis of Mastitis and Milk Production in Dairy Cattle. Sci Rep 7:45560. https://doi.org/10.1038/srep45560.

Finn, R. D., T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H. Y. Chang, Z. Dosztanyi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. Tosatto, C. H. Wu, I. Xenarios, L. S. Yeh, S. Y. Young, and A. L. Mitchell. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 45:D190-D199. https://doi.org/10.1093/nar/gkw1107.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44:D279-285. https://doi.org/10.1093/nar/gkv1344.

Handsaker, R. E., J. M. Korn, J. Nemesh, and S. A. McCarroll. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43:269-276. https://doi.org/10.1038/ng.768.

Handsaker, R. E., V. Van Doren, J. R. Berman, G. Genovese, S. Kashin, L. M. Boettger, and S. A. McCarroll. 2015. Large multiallelic copy number variations in humans. Nat Genet 47:296-303. https://doi.org/10.1038/ng.3200.

Hastings, P. J., J. R. Lupski, S. M. Rosenberg, and G. Ira. 2009. Mechanisms of change in gene copy number. Nat Rev Genet 10:551-564. https://doi.org/10.1038/nrg2593.

Hou, Y., G. E. Liu, D. M. Bickhart, M. F. Cardone, K. Wang, E. S. Kim, L. K. Matukumalli, M. Ventura, J. Song, P. M. VanRaden, T. S. Sonstegard, and C. P. Van Tassell. 2011. Genomic characteristics of cattle copy number variations. BMC Genomics 12:127. https://doi.org/10.1186/1471-2164-12-127.

Howard, J. T., S. D. Kachman, W. M. Snelling, E. J. Pollak, D. C. Ciobanu, L. A. Kuehn, and M. L. Spangler. 2014. Beef cattle body temperature during climatic stress: a genome-wide association study. Int J Biometeorol 58:1665-1672. https://doi.org/10.1007/s00484-013-0773-5.

Hu, Z. L., C. A. Park, and J. M. Reecy. 2016. Developmental progress and current status of the Animal QTLdb. Nucleic Acids Res 44:D827-833. https://doi.org/10.1093/nar/gkv1233.

Hurst, L. D. and N. G. Smith. 1999. Do essential genes evolve slowly? Curr Biol 9:747-750.

Jansen, S., B. Aigner, H. Pausch, M. Wysocki, S. Eck, A. Benet-Pages, E. Graf, T. Wieland, T. M. Strom, T. Meitinger, and R. Fries. 2013. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics 14:446. https://doi.org/10.1186/1471-2164-14-446.

Kadri, N. K., G. Sahana, C. Charlier, T. Iso-Touru, B. Guldbrandtsen, L. Karim, U. S. Nielsen, F. Panitz, G. P. Aamand, N. Schulman, M. Georges, J. Vilkki, M. S. Lund, and T. Druet. 2014. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. PLoS Genet 10:e1004049. https://doi.org/10.1371/journal.pgen.1004049.

Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353-D361. https://doi.org/10.1093/nar/gkw1092.

Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent. 2004. The UCSC Table Browser data retrieval tool. Nucleic Acids Res 32:D493-496. https://doi.org/10.1093/nar/gkh103.

Kent, W. J. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12:656-664. https://doi.org/10.1101/gr.229202. Article published online before March 2002.

Kinsella, R. J., A. Kahari, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, P. Kersey, and P. Flicek. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030. https://doi.org/10.1093/database/bar030.

Lappalainen, I., J. Lopez, L. Skipper, T. Hefferon, J. D. Spalding, J. Garner, C. Chen, M. Maguire, M. Corbett, G. Zhou, J. Paschall, V. Ananiev, P. Flicek, and D. M. Church. 2013. DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res 41:D936-941. https://doi.org/10.1093/nar/gks1213.

Lee, K., D. T. Nguyen, M. Choi, S. Y. Cha, J. H. Kim, H. Dadi, H. G. Seo, K. Seo, T. Chun, and C. Park. 2013. Analysis of cattle olfactory subgenome: the first detail study on the characteristics of the complete olfactory receptor repertoire of a ruminant. BMC Genomics 14:596. https://doi.org/10.1186/1471-2164-14-596.

Li, H. and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079. https://doi.org/10.1093/bioinformatics/btp352.

Liu, G. E., Y. Hou, B. Zhu, M. F. Cardone, L. Jiang, A. Cellamare, A. Mitra, L. J. Alexander, L. L. Coutinho, M. E. Dell'Aquila, L. C. Gasbarre, G. Lacalandra, R. W. Li, L. K. Matukumalli, D. Nonneman, L. C. Regitano, T. P. Smith, J. Song, T. S. Sonstegard, C. P. Van Tassell, M. Ventura, E. E. Eichler, T. G. McDaneld, and J. W. Keele. 2010. Analysis of copy number variations among diverse cattle breeds. Genome Res 20:693-703. https://doi.org/10.1101/gr.105403.110.

Lobago, F., H. Gustafsson, M. Bekana, J. F. Beckers, and H. Kindahl. 2006. Clinical features and hormonal profiles of cloprostenol-induced early abortions in heifers monitored by ultrasonography. Acta Vet Scand 48:23. https://doi.org/10.1186/1751-0147-48-23.

Mao, X., G. Sahana, D. J. De Koning, and B. Guldbrandtsen. 2016. Genome-wide association studies of growth traits in three dairy cattle breeds using whole-genome sequence data. J Anim Sci 94:1426-1437. https://doi.org/10.2527/jas.2015-9838.

McCarroll, S. A., F. G. Kuruvilla, J. M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M. H. Shapero, P. I. de Bakker, J. B. Maller, A. Kirby, A. L. Elliott, M. Parkin, E. Hubbell, T. Webster, R. Mei, J. Veitch, P. J. Collins, R. Handsaker, S. Lincoln, M. Nizzari, J. Blume, K. W. Jones, R. Rava, M. J. Daly, S. B. Gabriel, and D. Altshuler. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40:1166-1174. https://doi.org/10.1038/ng.238.

McClure, M. C., N. S. Morsci, R. D. Schnabel, J. W. Kim, P. Yao, M. M. Rolf, S. D. McKay, S. J. Gregg, R. H. Chapple, S. L. Northcutt, and J. F. Taylor. 2010. A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. Anim Genet 41:597-607. https://doi.org/10.1111/j.1365-2052.2010.02063.x.

McClure, M. C., H. R. Ramey, M. M. Rolf, S. D. McKay, J. E. Decker, R. H. Chapple, J. W. Kim, T. M. Taxis, R. L. Weaber, R. D. Schnabel, and J. F. Taylor. 2012. Genome-wide association analysis for quantitative trait loci influencing Warner-Bratzler shear force in five taurine cattle breeds. Anim Genet 43:662-673. https://doi.org/10.1111/j.1365-2052.2012.02323.x.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303. https://doi.org/10.1101/gr.107524.110.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. Ritchie, A. Thormann, P. Flicek, and F. Cunningham. 2016. The Ensembl Variant Effect Predictor. Genome Biol 17:122. https://doi.org/10.1186/s13059-016-0974-4.

Miller, S. A., D. D. Dykes, and H. F. Polesky. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16:1215.

Mills, R. E., K. Walter, C. Stewart, R. E. Handsaker, K. Chen, C. Alkan, A. Abyzov, S. C. Yoon, K. Ye, R. K. Cheetham, A. Chinwalla, D. F. Conrad, Y. Fu, F. Grubert, I. Hajirasouliha, F. Hormozdiari, L. M. Iakoucheva, Z. Iqbal, S. Kang, J. M. Kidd, M. K. Konkel, J. Korn, E. Khurana, D. Kural, H. Y. Lam, J. Leng, R. Li, Y. Li, C. Y. Lin, R. Luo, X. J. Mu, J. Nemesh, H. E. Peckham, T. Rausch, A. Scally, X. Shi, M. P. Stromberg, A. M. Stutz, A. E. Urban, J. A. Walker, J. Wu, Y. Zhang, Z. D. Zhang, M. A. Batzer, L. Ding, G. T. Marth, G. McVean, J. Sebat, M. Snyder, J. Wang, K. Ye, E. E. Eichler, M. B. Gerstein, M. E. Hurles, C. Lee, S. A. McCarroll, J. O. Korbel, and P. Genomes. 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470:59-65. https://doi.org/10.1038/nature09708.

Niimura, Y. and M. Nei. 2007. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS One 2:e708. https://doi.org/10.1371/journal.pone.0000708.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-575. https://doi.org/10.1086/519795.

Quinlan, A. R. and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842. https://doi.org/10.1093/bioinformatics/btq033.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. 2006. Global variation in copy number in the human genome. Nature 444:444-454. https://doi.org/10.1038/nature05329.

RStudio Team. 2016. RStudio: Integrated Development Environment for R. RStudio, Inc., Boston, MA.

Sahana, G., B. Guldbrandtsen, and M. S. Lund. 2011. Genome-wide association study for calving traits in Danish and Swedish Holstein cattle. J Dairy Sci 94:479-486. https://doi.org/10.3168/jds.2010-3381.

Sahana, G., T. Iso-Touru, X. Wu, U. S. Nielsen, D. J. de Koning, M. S. Lund, J. Vilkki, and B. Guldbrandtsen. 2016. A 0.5-Mbp deletion on bovine chromosome 23 is a strong candidate for stillbirth in Nordic Red cattle. Genet Sel Evol 48:35. https://doi.org/10.1186/s12711-016-0215-z.

Sandri, M., B. Stefanon, and J. J. Loor. 2015. Transcriptome profiles of whole blood in Italian Holstein and Italian Simmental lactating cows diverging for genetic merit for milk protein. J Dairy Sci 98:6119-6127. https://doi.org/10.3168/jds.2014-9049.

Schutz, E., C. Wehrhahn, M. Wanjek, R. Bortfeld, W. E. Wemheuer, J. Beck, and B. Brenig. 2016. The Holstein Friesian Lethal Haplotype 5 (HH5) Results from a Complete Deletion of TBF1M and Cholesterol Deficiency (CDH) from an ERV-(LTR) Insertion into the Coding Region of APOB. PLoS One 11:e0154602. https://doi.org/10.1371/journal.pone.0154602.

Snelling, W. M., M. F. Allan, J. W. Keele, L. A. Kuehn, T. McDaneld, T. P. Smith, T. S. Sonstegard, R. M. Thallman, and G. L. Bennett. 2010. Genome-wide association study of growth in crossbred beef cattle. J Anim Sci 88:837-848. https://doi.org/10.2527/jas.2009-2257.

Sudmant, P. H., T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Hsi-Yang Fritz, M. K. Konkel, A. Malhotra, A. M. Stutz, X. Shi, F. Paolo Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. J. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Y. Lam, X. Jasmine Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. M. Kidd, Y. Kong, E. W. Lameijer, S. McCarthy, P. Flicek, R. A. Gibbs, G. Marth, C. E. Mason, A. Menelaou, D. M. Muzny, B. J. Nelson, A. Noor, N. F. Parrish, M. Pendleton, A. Quitadamo, B. Raeder, E. E. Schadt, M. Romanovitch, A. Schlattl, R. Sebra, A. A. Shabalin, A. Untergasser, J. A. Walker, M. Wang, F. Yu, C. Zhang, J. Zhang, X. Zheng-Bradley, W. Zhou, T. Zichner, J. Sebat, M. A. Batzer, S. A. McCarroll, C. Genomes Project, R. E. Mills, M. B. Gerstein, A. Bashir, O. Stegle, S. E. Devine, C. Lee, E. E. Eichler, and J. O. Korbel. 2015. An integrated map of structural variation in 2,504 human genomes. Nature 526:75-81. https://doi.org/10.1038/nature15394.

Szklarczyk, D., A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43:D447-452. https://doi.org/10.1093/nar/gku1003.

Van Ziffle, J., W. Yang, and F. F. Chehab. 2011. Homozygous deletion of six olfactory receptor genes in a subset of individuals with Beta-thalassemia. PLoS One 6:e17327. https://doi.org/10.1371/journal.pone.0017327.

Weischenfeldt, J., O. Symmons, F. Spitz, and J. O. Korbel. 2013. Phenotypic impact of genomic structural variation: insights from and for human disease. Nat Rev Genet 14:125-138. https://doi.org/10.1038/nrg3373.

Wright, S. 1931. Evolution in Mendelian Populations. Genetics 16:97-159.

Xu, L., D. M. Bickhart, J. B. Cole, S. G. Schroeder, J. Song, C. P. Tassell, T. S. Sonstegard, and G. E. Liu. 2015. Genomic signatures reveal new evidences for selection of important traits in domestic cattle. Mol Biol Evol 32:711-725. https://doi.org/10.1093/molbev/msu333.

Xu, L., J. B. Cole, D. M. Bickhart, Y. Hou, J. Song, P. M. VanRaden, T. S. Sonstegard, C. P. Van Tassell, and G. E. Liu. 2014. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. BMC Genomics 15:683. https://doi.org/10.1186/1471-2164-15-683.

Xu, L., Y. Hou, D. M. Bickhart, Y. Zhou, H. A. Hay el, J. Song, T. S. Sonstegard, C. P. Van Tassell, and G. E. Liu. 2016. Population-genetic properties of differentiated copy number variations in cattle. Sci Rep 6:23161. https://doi.org/10.1038/srep23161.

Yalcin, B., K. Wong, A. Agam, M. Goodson, T. M. Keane, X. Gan, C. Nellaker, L. Goodstadt, J. Nicod, A. Bhomra, P. Hernandez-Pliego, H. Whitley, J. Cleak, R. Dutton, D. Janowitz, R. Mott, D. J. Adams, and J. Flint. 2011. Sequence-based characterization of structural variation in the mouse genome. Nature 477:326-329. https://doi.org/10.1038/nature10432.

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

G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino, and P. Flicek. 2016. Ensembl 2016. Nucleic Acids Res 44:D710-716. https://doi.org/10.1093/nar/gkv1157.

Zarrei, M., J. R. MacDonald, D. Merico, and S. W. Scherer. 2015. A copy number variation map of the human genome. Nat Rev Genet 16:172-183. https://doi.org/10.1038/nrg3871.

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

Chapter 4.

Joint imputation of whole-genome sequence variants and large chromosomal deletions in cattle

Md Mesbah-Uddin,[1,2*] Bernt Guldbrandtsen,[1] Mogens Sandø Lund,[1] Didier Boichard,[2] and Goutam Sahana,[1*]

[1]Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark

[2]GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[*]Corresponding authors: mdmesbah@gmail.com and goutam.sahana@mbg.au.dk

Journal of Dairy Science, 2019. 102(12): p. 11193-11206.

Supplementary materials are available at https://doi.org/10.3168/jds.2019-16946.

4.1 Abstract

Genotype imputation, often focused on single-nucleotide polymorphisms (SNPs) and small insertions and deletions (indels; size <100 bp), is a crucial step for association mapping and estimation of genomic breeding values. Here, we present strategies to impute genotypes for large chromosomal deletions (size >100 bp), along with SNPs and indels in cattle. The pipelines include a strategy for extending the whole-genome sequence reference panel for large deletions, a two-step genotype refinement approach using Beagle4 and SHAPEIT2 software, and finally, joint imputation of SNPs, indels and large deletions to the existing SNP array-typed population using Minimac3 software. Using these pipelines we achieved an imputation accuracy of r^2 >0.6 at minor allele frequencies as low as 0.7% for SNPs and indels, and 0.2% for large deletions. This highlights the potential of our approach to build haplotype reference panel and impute different classes of sequence variants across a wide allele frequency-spectrum with high accuracy.

Key words: Gaussian mixture model, imputation, structural variant, copy-number variation, chromosomal deletion

4.2 Introduction

The population-scale whole-genome resequencing in different cattle breeds has generated a comprehensive catalog of DNA polymorphisms, including single-nucleotide polymorphisms (SNPs), small insertions and deletions (indels) (Daetwyler et al., 2014), copy-number variations (CNVs) (Letaief et al., 2017), and structural variants (SVs) (Boussaha et al., 2015, Chen et al., 2017). In livestock, there is a growing interest in understanding the phenotypic impacts of SVs, e.g., large deletions and duplications, because a substantial proportion of SVs overlap genes and may have a biological effect (Bickhart and Liu, 2014). Such genomic studies require SV genotypes in large cohorts of animals, which are rarely available. However, one way to generate such SV datasets is to design a customized SNP array using the precise SV breakpoints and genotype the population directly. The custom part of the EuroG10K BeadChip (Boichard et al., 2018) is an example of this. Alternatively, one can perform imputation for common bi-allelic SVs in an approach similar to that of bi-allelic SNPs and indels—since common SVs are often tagged well by the surrounding SNP haplotypes (as shown by studies in humans, e.g., Conrad et al. (2010) and Handsaker et al. (2015)). The later approach will allow us to impute SVs into the large cohorts of animals genotyped using the existing SNP arrays.

However, the majority of whole-genome sequence (WGS) based studies in cattle were either constrained by (small) sample size (Shin et al., 2014, Gao et al., 2017, Mielczarek et al., 2017) or (unavailability of) SV genotypes (Chen et al., 2017). Therefore, building a suitable reference panel for imputing SVs into large cohorts of SNP array-typed cattle population remains a challenge. Earlier, we reported ~8,500 large deletions (size: 199 bp-773 Kb) with genotype likelihoods in 175 WGS animals from three Nordic dairy cattle breeds, namely, Nordic Holstein (HOL), Danish Jersey (JER), and Nordic Red Dairy Cattle (RDC) (Mesbah-Uddin et al., 2018a). We were interested to extend our large deletion (deletion, hereafter) reference panel of 175 animals by including additional sequenced animals from corresponding breeds from shared Variant Call

Format (VCF) files of the 1,000 Bull Genomes Project (1KBGP). In this study, we present a strategy for computing deletion genotype likelihoods using the SNP read depth information from a VCF file when access to BAM files are not available, building a haplotype reference panel and jointly imputing WGS SNPs, indels, and deletions.

4.3 Methods

SNP Genotype Dataset

A total of 13,307 JER, HOL and RDC were genotyped using either Illumina BovineSNP50 (50k, hereafter) or BovineHD BeadChip (777k, hereafter). The 50k and the 777k datasets comprised 9,932 (838 JER, 5,371 HOL, and 3,723 RDC) and 3,375 (835 JER, 1,215 HOL, and 1,325 RDC) animals, respectively.

From both datasets, we excluded SNPs that were monomorphic, had a GenCall <0.60, or deviated from Hardy-Weinberg proportions (*P value* $\leq 1 \times 10^{-5}$) in any of the three breeds. We also removed SNPs, which overlapped a deletion locus from our deletion dataset (see *"WGS Dataset"* subsection "Deletions"). After filtering, the 50k and 777k datasets had 47,237 and 697,296 SNPs, respectively.

WGS Dataset

In total, 772 animals were available in our WGS dataset (92 JER, 132 RDC and 548 HOL). Among these animals, 175 and 597 were from the Nordic sequence project (Brøndum et al., 2014) and the *Run-6* of the 1KBGP (Daetwyler et al., 2014), respectively. GATK v1.6 (McKenna et al., 2010) and SAMtools 0.1.18 mpileup (Li et al., 2009) software were used to call SNPs and indels from the Nordic and the 1KBGP WGS dataset, respectively (for detail, see Brøndum et al. (2014) and Daetwyler et al. (2014)).

SNPs and indels. In this study, we only considered Phred-scaled genotype likelihoods (PL) for bi-allelic SNPs and indels, which occurred in both the Nordic and the 1KBGP dataset. We removed a marker from the dataset when any of the following conditions was met, namely: (i) minor allele count (MAC) less than 5 (vcftools' --mac 5), (ii) deviation from Hardy-Weinberg proportion in any one breed (*P-value* $\leq 1 \times 10^{-5}$), (iii) missing genotypes in more than 10% of the animals (--max-missing 0.1), (iv) marker within 5 bp from one another (--thin 5), (v) marker quality less than 50 (--minQ 50), (vi) markers within 100 bp of gaps in the UMD3.1 bovine genome assembly (Zimin *et al.* 2009), or (vii) markers within 1 Kb of a deletion locus from our deletion list (discussed later). After filtering, 14,800,299 bi-allelic markers remained in our WGS dataset: 14,678,220 SNPs and 122,079 indels.

Deletions. In addition, Mesbah-Uddin et al. (2018a) estimated PL for ~8,500 deletions on Nordic samples from the sequences aligned to UMD3.1 (BAM files) using GenomeSTRiP-2 software (Handsaker et al., 2011). Among them, we selected 5,798 deletions after a thinning of 100 Kb using VCFtools' "--thin 100000" option, i.e., excluding deletions within 100 Kb distance from one other. From this dataset, we extracted deletion PL for the Nordic animals, while for the remaining 597 animals, we estimated PL using a modified version of the read depth (RD) genotyping method used by Mesbah-Uddin et al. (2018b), as explained below.

Estimating Deletion Genotype Likelihoods from VCF file

We estimated genotype likelihoods for the selected deletions on 597 additional animals from the 1KBGP (*Run-6*) using RD data from VCF file.

Calculation of expected RD. Studies showed that GC bias, i.e., RD variation due to differences in the percentage of guanine (G) and cytosine (C) bases in the region, can confound the inference of copy-number from RD (Abyzov et al., 2011, Benjamini and Speed, 2012). For the genome assembly UMD3.1, we first calculated the GC% in bins of 100 bp after excluding potential copy-number variable regions, such as sex chromosomes, mitochondrial sequences, unplaced contigs, assembly gaps, repeat sequences, and CNVs and SVs from DGVa database (last accessed on January 22, 2019, from: https://www.ebi.ac.uk/dgva). Next, for each animal, we extracted RD data from the corresponding genomic intervals from VCF files, and calculated the expected RD (μ_{GC}) and variance (var_{GC}) for different GC% (for details, see **Supplementary Methods**). There was substantial variations in RD against different GC% between samples (Supplementary Figure S1). To account for this heterogeneous GC% bias for a given deletion locus, the expected RD was assumed to be the average RDs from its GC% bin, instead of the overall genomic coverage. For example, when interrogating a 10 Kb deletion locus with 30% GC content, we used the genomic average over all 30% GC bins as the expected RD for each corresponding animal.

Gaussian mixture model. To compute genotype likelihoods for a deletion, we first extracted the SNPs (say n SNPs in total) within the deleted segment and retrieved read depth data from the allelic-depth (AD) tag of the VCF file. Here, we only considered bi-allelic SNPs with QUAL ≥30 and a thinning of 10 bp. Next, we fitted a Gaussian mixture model (GMM) to the data, assuming a linear relationship between the observed RD within the locus and (unobserved) copy-number (CN) of the locus, such as,

$$p(RD_i) = \sum_{k=0}^{2} w_k \, N(RD_i \mid mu_k, var_k)$$

Here, *RD* is a vector of n data points corresponding to the n SNPs; *k* (0 to 2) is an indicator variable for the Gaussian component; w_k is the relative weight for CN classes (e.g., w_0, w_1, w_2 for CN_0, CN_1 and CN_2, respectively); mu_k and var_k are the expectation and variance of the corresponding Gaussian distribution. The GMM parameters were estimated using an expectation-maximization (EM) algorithm as follows:

Initialization of the EM. Assuming a deletion locus, we constrained GMM to fit exactly three components with a fixed mu_k for each *k* such that $mu_0 = 0$ for CN_0, $mu_1 = \frac{\mu_{GC}}{2}$ for CN_1, and $mu_2 = \mu_{GC}$ for CN_2. We initialized w_k with three equal weights of 1/3, and var_k with the value of 0.2, $\frac{var_{GC}}{2}$, and var_{GC}, for CN_0, CN_1 and CN_2, respectively.

E-step. Given the parameters, the expected values for the latent variable Z_{ik} were estimated as:

$$Z_{ik} = \frac{w_k \, N(RD_i \mid mu_k, \, var_k)}{\sum_{k=0}^{2} w_k \, N(RD_i \mid mu_k, \, var_k)}$$

M-step. The membership weights and variances were updated using the Z_{ik} values from the E-step (while keeping mu_k fixed):

$$w_k = \frac{1}{n}\sum_{i=1}^{n} Z_{ik}$$

$$var_k = \frac{\sum_{i=1}^{n} Z_{ik}(RD_i - mu_k)^2}{\sum_{i=1}^{n} Z_{ik}}$$

For a given deletion, parameters w_k and var_k were estimated iteratively from the observed *RD* data of each animal using the EM algorithm until convergence. The scripts used to prepare the RD data from the VCF file and to estimate deletion PL could be accessed from: "https://github.com/MMesbahU/ImputeDelPipeline/tree/master/read_Depth_genotyping".

Refining the genotypes. Next, to refine the deletion genotypes along with the WGS SNPs and indels, we combined PL estimates for all these variants and performed chromosome-wise phasing using Beagle v4.r1274 software (Browning and Browning, 2016) (hereafter referred to as Beagle; for details see "*Phasing*").

Evaluation of the deletion genotyping pipeline. We evaluated the pipeline for Brachyspina-associated deletion located on chromosome 21 between the positions 21,184,869 and 21,188,202 bp (Charlier et al., 2012). In this dataset, we had 113 Holstein with known Brachyspina carrier-status. There were 13 carriers and 100 non-carriers of the deletion. Using the GMM approach described before, we first estimated deletion PL using the SNP RD from the VCF file. After excluding variants within the deletion, we extracted PL for WGS SNPs and indels within ±1 Mb of the deletion. Next, to refine the deletion genotypes, we performed phasing using the surrounding haplotypes using Beagle (with same parameters as discussed in the "*Phasing*" step). Finally, the genotype concordance (true vs. Beagle's best guess genotypes) was manually inspected.

Phasing

Following the method used by Delaneau et al. (2014), we phased our genotype dataset using a combination of software Beagle and SHAPEIT v2.r837 (Delaneau et al., 2013a) (hereafter referred to as SHAPEIT). First, providing the WGS PLs as input, we ran Beagle for 10 burn-in and 15 sampling iterations with a window size of 12,000 markers and an overlap of 2,000 markers between consecutive windows. From this step, we obtained posterior genotype probabilities (GP) for each bi-allelic variant for the three possible genotypes, e.g., homozygous reference, heterozygous (or hemizygous for deletion), and homozygous for alternate allele (or homozygous deletion). We fixed the genotype for a variant as known when Beagle GP ≥0.99, and for the remaining variants, we re-called the genotype using SHAPEIT. To initialize phasing and calling in SHAPEIT, we used the haplotypes generated in the Beagle step. After this initialization, SHAPEIT was run for 12 pruning stages of four iterations each, followed by 20 main MCMC sampling iterations. We used a window size of 0.1 Mb for the WGS variants, and 2 Mb for the 50k and 777k variants, following the SHAPEIT guidelines (last accessed on May 8, 2019, from:

https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#gcall). In each window, 600 conditioning haplotypes were generated, among which, 400 were Hamming distance conditioning (--states 400) and the remaining 200 were chosen randomly (--states-random 200) to improve the mixing of the MCMC. However, both the 50k and 777k dataset were already phased using Beagle. We took the Beagle haplotypes and GPs, and performed phasing (and calling) using SHAPEIT, as described above.

Imputation

The imputation was performed in two steps using Minimac3 v.2.0.1 software (Das et al., 2016) (hereafter referred to as Minimac) with default parameters. In the first step, we imputed 9,932 animals from 50k to 777k using the 777k haplotype reference panel consisted of 3,375 animals. There were few animals common in both 50 k and 777 k dataset or 777 k and WGS dataset. Here, we kept the sample with highest genotype density (i.e. WGS >777k >50k) in the reference panel and remove the sample from the corresponding imputation target dataset. We then combined both imputed and genotyped 777k dataset to prepare a non-redundant target of 12,966 animals (1,640 JER, 6,371 HOL and 4,955 RDC), after excluding the overlap between the 777k and WGS dataset. Next, using the haplotype reference panel of 772 WGS animals we imputed our target population from 777k to WGS SNPs, indels and deletions.

Evaluation of the Reference Panel

First, we prepared a phased reference panel using Beagle with the following inputs: (Phred-scaled) genotype likelihoods (PLs) for WGS SNPs and indels in 772 animals, deletion PLs in 175 animals from the study by Mesbah-Uddin et al. (2018a), deletion PLs estimated for 597 animals using auxiliary read depth data from VCF files with missing deletion PLs coded as "." for EM non-convergence. Second, to get the final reference panel, we re-phased and refined genotypes at loci with Beagle GP less than 0.99 using SHAPEIT. We refer to the first and second reference panels as Ref_{Beagle} and $Ref_{Beagle_SHAPEIT}$, respectively. We assessed the performance of this two-step approach on chromosome 29. The performance differences between the two reference panels, $Ref_{Beagle_SHAPEIT}$ vs. Ref_{Beagle}, were tested using two-tailed paired t-test.

Cross validation. We prepared two reference panels of chromosome 29: Ref_{Beagle} and $Ref_{Beagle_SHAPEIT}$. The leave-out-trials were performed on 112 deletions with MAC ≥5 common in both dataset. For each deletion, we extracted WGS markers within 1 Mb on either side of the deletion, after excluding markers within and ±500 bp surrounding regions. Next, we removed the phasing for all the markers, and performed 77 leave-out-trials by randomly masking deletion genotypes for a set of non-overlapping (i.e. sampled without replacement) 10 animals in each trial (12 animals in the last trial). We used the same random seed to mask the genotypes of samples in both dataset. These analyses were performed using Beagle with the same parameters as in the "*Phasing*" step.

Here, we considered the deletion genotypes in the corresponding reference panel as the true genotype for that scenario. To evaluate the imputation accuracy at each deletion locus, we extracted the imputed and true genotypes for the corresponding animals from the 77 trials, and performed Pearson's correlation (***r***) of allele

dosages (coded as 0, 1, or 2 for homozygous reference, hemizygous or homozygous deletion, respectively) between the reference and the imputed dataset. Finally, we reported the squared version of the correlation coefficient (r^2) in the Results section.

Downstream imputation performance. For chromosome 29, we also assessed downstream performances of the two reference panels (Ref_{Beagle} and $Ref_{Beagle_SHAPEIT}$) to impute jointly the target 777k dataset to WGS SNPs, indels, and deletions using Minimac. Here, we evaluated the performance based on the imputation accuracy estimates ($\hat{r}^2$) provided in the Minimac "info" output file. Under the assumption that genotype counts for a poorly imputed marker will shrink towards its expectation, Minimac calculates imputation accuracy as follows (last accessed on May 8, 2019, from: https://genome.sph.umich.edu/wiki/Minimac3_Info_File#Rsq):

$$\hat{r}^2 = \frac{\frac{1}{2n} \times \sum_{i=1}^{2n} (D_i - \hat{p})^2}{\hat{p}(1 - \hat{p})}$$

Here, $\hat{p}$ = the estimated alternate allele frequency, D_i = the imputed alternate allele probability at the i^{th} haplotype, and n = the number of samples in the imputation target panel.

Tools used for data manipulation and analyses

We used following tools for data preparation and analyses: R software version 3.1.2 (R Core Team, 2014), Python software version 3.7 (https://www.python.org/), VCFtools v0.1.12a (Danecek et al., 2011), BEDTools v2.26.0 (Quinlan and Hall, 2010) and BCFtools v1.7 (https://github.com/samtools/BCFtools).

Principal Component Analysis

To distinguish the three breeds present in our dataset, we performed a principal component (PC) analysis using the imputed WGS variants with MAF >1%. Within this MAF threshold, we randomly selected 1% SNPs, 10% indels, and 100% deletions from each of the 29 autosomes. The final dataset contained 137,064 SNPs, 11,710 indels, and 4,670 deletions. The analysis was performed using PLINK (v2.00a2 AVX2) software (Chang et al., 2015). For each of the three variant classes, we presented the first two PCs and along with the proportion of variance explained by the corresponding PC.

4.4 Results and discussion

Deletion Genotyping using RD data from the 1KBGP VCF file

The WGS reference panel was extended from 175 animals to 772 animals by data from the 1KBGP. In the extended reference panel, genotype likelihoods (PL) for 14,800,299 bi-allelic markers (SNPs + indels) on all the animals were available. However, the deletion PL for 5,798 selected markers were only available on

175 animals. For the remaining 597 animals, we estimated PL by fitting the GMM with three components to the RD data extracted from the VCF file. We then used surrounding haplotypes to refine deletion genotypes in animals with ambiguous (or missing) PL estimates.

We evaluated the deletion genotyping pipeline for classifying animals for their carrier status for the Brachyspina-associated deletion (Charlier et al., 2012) in a panel of 113 animals with known carrier status including 13 hemizygotes and 100 homozygotes for the reference allele. We found that our two-step approach—estimating PL using WGS SNP RD data from the VCF file followed by a genotype refinement step using Beagle—accurately classified deletion genotypes in all the animals (13 hemizygotes and 100 homozygotes for the reference allele).

The WGS Reference Panel of 772 Animals

Following previous reports (Brøndum et al., 2014, Pausch et al., 2017), where the authors showed that a multi-breed reference panel provides better downstream imputation performance in cattle compared with a single-breed reference panel, we built a multi-breed WGS reference panel of 772 animals that included 92 JER, 132 RDC and 548 HOL. We applied the GMM pipeline to estimate the genotype likelihoods for all the deletions in the additional animals included from the 1KBGP. Next, to refine the genotypes and get posterior genotype probability (GP), we performed phasing by chromosome using Beagle with the PL estimates for all SNPs, indels and deletions as inputs.

However, studies in humans have shown that an initial phasing of WGS genotypes using Beagle followed by fixing the genotype with high posterior probability (e.g., Beagle GP >0.995) and re-calling the remaining markers genotypes using SHAPEIT "-call" provides further improvement in the reference panel (Delaneau et al., 2014, Genomes Project et al., 2015). Besides, SHAPEIT can also improve haplotype phasing by leveraging on the haplotype sharing within and between populations with heterogeneous ancestry (Delaneau et al., 2013b). Here, both our WGS reference and SNP array-typed target populations consist of animals with diverse ancestry (see Fig. 3 from Bouwman et al. (2018) for examples of haplotype diversity in JER, HOL, and RDC; see Zhang et al. (2018) for a detailed account in RDC). In addition, on average ~2.7% of the total WGS genotypes in our dataset had Beagle GP <0.99 (Table S1). Furthermore, there were 7.3% (1.1 of 14.7 million) SNPs, 18.3% (22.4 of 122.1 k) indels and 66.6% (3.9 of 5.8 k) deletions where more than 10% genotypes were below this threshold (Table S1). Therefore, we expected that by using SHAPEIT we could further improve the reference panel by improving (low-quality) genotype calls and haplotype phasing. Based on this assumption, we next refined the reference panel using SHAPEIT, where all genotypes were re-phased, and re-called when Beagle GP was less than 0.99.

To evaluate the effect of this two-step refinement approach on imputation accuracy we prepared $\text{Ref}_{\text{Beagle}}$ and $\text{Ref}_{\text{Beagle_SHAPEIT}}$ of the chromosome 29. First, we performed leave-out-trials on 112 deletions with MAC ≥5 from this chromosome and evaluated the imputation accuracy using (squared) Pearson's correlation (r^2) between true vs. imputed dosages of 0, 1, or 2 for homozygous reference, hemizygous or homozygous deletion, respectively (see **Methods** "Cross validation"). After the trials, five deletions became homozygous

for the reference allele. These were excluded from subsequent analysis. Overall, the average gain in accuracy was r^2 = 0.081 (95% CI: 0.053, 0.109) with $Ref_{Beagle_SHAPEIT}$ vs. Ref_{Beagle} ($P = 7.9 \times 10^{-8}$; two-tailed paired t-test). The performance gain was even higher (r^2 = 0.105 [CI: 0.069, 0.141], $P = 1.7 \times 10^{-7}$) for variants where more than 10% of the genotypes were under the GP threshold (Table 4.1). Furthermore, 76 vs. 90% of the deletions achieved r^2 >0.80 using the reference Ref_{Beagle} vs. $Ref_{Beagle_SHAPEIT}$, respectively (Figure 4.1a).

Next, we evaluated the downstream performances of the two reference panels to impute jointly a target population of ~13k animals with 777k genotypes to WGS SNPs, indels and deletions of chromosome 29 using Minimac. As expected, the reference panel $Ref_{Beagle_SHAPEIT}$ showed better downstream imputation performance for all variant types compared to that of Ref_{Beagle} (Figure 4.1b, 4.2ab, and Table 4.1). The average gains in accuracy in terms of Minimac r^2 were 0.032 (CI: 0.012, 0.053; P = 0.002) for deletions, 0.020 (CI: 0.019, 0.020; $P < 1.0 \times 10^{-16}$) for SNPs and 0.034 (CI: 0.031, 0.037; $P < 1.0 \times 10^{-16}$) for indels (Table 4.1). However, after Beagle step, ~8% of the SNPs, 20% of the indels and 70% of the deletions had more than 10% of genotypes below the GP threshold (Table S1). Indeed, for these markers we observed a clear difference between before vs. after SHAPEIT step. For example, the average gain in accuracy after the SHAPEIT step was 0.041 (CI: 0.011, 0.070; P = 0.007) for deletions, 0.058 (CI: 0.056, 0.059; $P < 1.0 \times 10^{-16}$) for SNPs and 0.076 (CI: 0.065, 0.087; $P < 1.0 \times 10^{-16}$) for indels (Table 4.1). Besides, 87 vs. 93% of the deletions (Figure 4.1b), 87 vs. 89% of the SNPs (Figure 4.2a), and 78 vs. 83% of the indels (Figure 4.2b) achieved r^2 >0.80 using the reference panel Ref_{Beagle} vs. $Ref_{Beagle_SHAPEIT}$, respectively. It is worth noting that for the majority of the imputed variants, allele frequencies differed between the two reference panels. Nevertheless, the average r^2 in different 5% MAF bins were also higher with $Ref_{Beagle_SHAPEIT}$ vs. Ref_{Beagle} (Figure S2 and Table S2). Thus, it is likely that SHAPEIT "-call" step improved genotype calls, providing an improved WGS reference panel for the chromosome. These results are in agreement with the reported performance gains in humans (Delaneau et al., 2014).

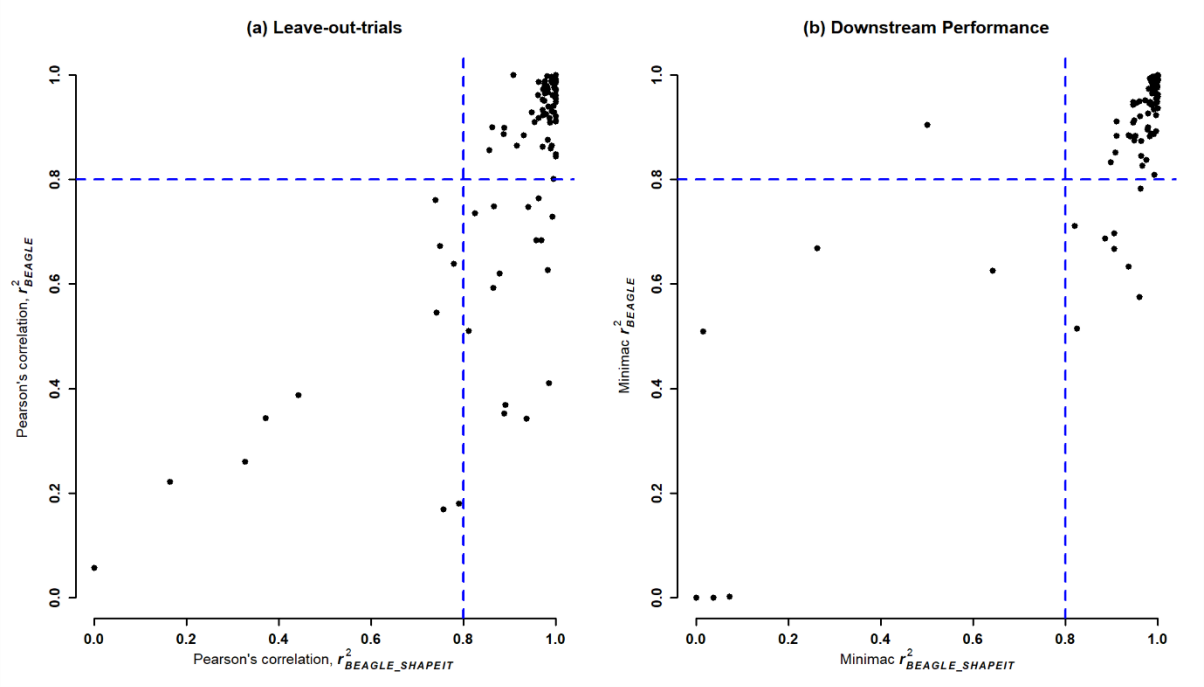


Figure 4.1. Imputation accuracy for 107 deletions on chromosome 29 using the reference panels Ref_{Beagle} vs. $Ref_{Beagle_SHAPEIT}$. *(a) Leave-out-trials. For each deletion, 77 trials were performed: 76 trials of 10 animals each and 1 trial of 12 animals. Imputation accuracy was calculated using the Pearson's correlation, r^2, between the true and imputed alternative allele dosages. (b) Downstream performance. Using the two reference panels, a target population of ~13k animals with the 777k genotypes was imputed for the deletions (along with other WGS markers) of chromosome 29 using Minimac. Each dot represent one deletion; dashed lines indicate r^2 value of 0.80. Ref_{Beagle}: reference panel phased using Beagle; $Ref_{BeagleSHAPEIT}$: reference panel phased using Beagle followed by re-phasing using SHAPEIT.*

Imputation of WGS SNPs, Indels and Deletions to the Chip-Typed Cattle Population

Our imputation target was a population of ~13k animals (1,640 JER, 6,371 HOL and 4,955 RDC) that were genotyped using either the 50k or the 777k BeadChip. We prepared this dataset using the same pipeline: two-step phasing followed by imputation from the 50k to the 777k genotypes using Minimac (see **Methods** "*Phasing*" and "*Imputation*"). Next, for the 29 bovine autosomes, we jointly imputed this target 777k dataset to WGS SNPs, indels and deletions using the reference panel $Ref_{Beagle_SHAPEIT}$ comprised of 772 animals. In Figure 4.4a-c, we present the MAF vs. Minimac r^2 estimates for the imputed WGS markers. The average imputation accuracy for SNPs, indels and deletions were 0.928 (SD 0.207), 0.900 (0.217) and 0.904 (0.220), respectively. Here, irrespective of the marker type, the common variants (MAF >5%) were well imputed with an average r^2 >0.90 (Table 4.2). Furthermore, low frequency variants (1%≤ MAF ≤5%) were also well imputed with an average r^2 value of 0.889 (0.250) for SNPs, 0.874 (0.245) for indels and 0.927 (0.154) for deletions. As for the rare variants (MAF <1%), we observed a sharp reduction in imputation accuracy compared to that of low frequency variants: differences of 0.361, 0.325 and 0.214 were seen for SNPs, indels and deletions, respectively (Table 4.2). Similar low imputation accuracies for rare variants were also reported previously (Brøndum et al., 2014, Pausch et al., 2017). Rare variants are often novel in the population, have low linkage-disequilibrium with neighboring markers and poorly represented in the reference panel, if at all.

For these variants, we further stratified the r^2 into MAF bins of 0.1% (Table 4.2). We found that both the SNPs and indels with MAF >0.5% had an imputation accuracy r^2 >0.50. The reduction was mainly due to variants in the bottom half (in terms of MAF) of this category. In contrast to SNPs and indels, rare deletions with MAF = 0.2% were also imputed with high accuracy (r^2 = 0.633), and for the lowest 0.1% MAF bin (i.e. 0< MAF ≤0.1%), r^2 was 0.313 (Table 4.2). Nevertheless, compared to the previous report on this population (see Figure S2 from Wu et al. (2016)), we observed substantial gains in accuracy in a wide range of MAF bins (Figure 4.3, Table 4.2 and S3). These results also showed that, given a suitable reference panel, Minimac could jointly impute various classes of bi-allelic sequence variants—in a frequency spectrum range from rare to common—to the existing chip-typed population with high accuracy.

Principal Component Analysis

Next, to assess the usability of this dataset in population-genetic studies, we performed principal component analysis using a random sample of the imputed variants with MAF >1%. Here we verified that the three classes of imputed markers were able to distinguish the three population present in our dataset in a similar way. Indeed, with the top two principal components from SNPs, indels and deletions, we were able to illustrate the population structure present in our dataset (Figure 4.4).

4.5 Conclusion

In this study, we presented strategies to incorporate large chromosomal deletions in population genetic studies, along with SNPs and indels. First, we showed an approach to estimate deletion genotype likelihood using the read depth data from the VCF file. Second, using a two-step genotype refinement, we built WGS reference panels that included SNPs, indels, and deletions. Finally, we imputed our 777k population to full sequence. We found that taking genotype uncertainty in the reference panel into account leads to substantial gains in accuracy. These results highlighted the feasibility of building a WGS reference panel comprising different classes of sequence variants and jointly imputing them to the existing SNP chip-typed large cohorts of animals. In addition, we showed that, as WGS resources, such as the 1KBGP, both include animals with shallow and deep sequencing depth, refining the low-confidence genotypes yielded a substantial gain in imputation accuracy. Furthermore, such an approach also provided accurate genotypes for rare variants—which often include putative causal variants. This, in turn, is expected to improve downstream association signals.

This study was focused only on the imputations of deletions along with SNPs and indels. With such an accurate imputation procedure even for rare and low frequency variants, it becomes possible to explore their potential biological effects by genome-wide association study and to study their predictive ability in genomic evaluation. This requires further investigation and will be presented in another study. It is likely that the strategies and pipelines demonstrated here will facilitate joint imputation of CNVs, SVs, SNPs and indels, and subsequent population genetic studies in cattle. Nevertheless, this imputation resource will be used in association mapping and estimation of (genomic) breeding values in Nordic dairy cattle breeds.

Table 4.1. Comparison of imputation performances of two reference panels, Ref_{Beagle} vs. $Ref_{Beagle_SHAPEIT}$ [1], on test chromosome 29

Evaluation	WGS Markers [3]	Overall (average) performance r^2 (SD)		Average r^2 difference between the reference panels $Ref_{Beagle_SHAPEIT}$ vs. Ref_{Beagle}: [2]					
		Ref_{Beagle}	$Ref_{Beagle_SHAPEIT}$	All markers		Markers with less than 10% genotypes under the threshold [4]		Markers with more than 10% genotypes under the threshold	
				Mean r^2 (CI)	*P-value* [5]	Mean r^2 (CI)	*P-value*	Mean r^2 (CI)	*P-value*
LoT [6]	**Deletions**	0.844 (0.223)	0.925 (0.165)	0.081 (0.053, 0.109)	7.9×10^{-8}	0.028 (-0.008, 0.064	0.123	0.105 (0.069, 0.141)	1.7×10^{-7}
Downstream imputation [7]	**Deletions**	0.894 (0.187)	0.926 (0.201)	0.032 (0.012, 0.053)	0.002	0.014 (0.005, 0.022)	0.003	0.041 (0.011, 0.070)	0.007
	SNPs	0.894 (0.246)	0.914 (0.226)	0.020 (0.019, 0.020)	[8]	0.016 (0.016, 0.017)	[8]	0.058 (0.056, 0.059)	[8]
	Indels	0.855 (0.254)	0.888 (0.229)	0.034 (0.031, 0.037)	[8]	0.024 (0.021, 0.026)	[8]	0.076 (0.065, 0.087)	[8]

[1] Ref_{Beagle}: reference panel phased using Beagle; $Ref_{BeagleSHAPEIT}$: reference panel phased using Beagle followed by re-phasing using SHAPEIT

[2] Performance gains between the two reference panels were calculated using R software function *t.test(*$Ref_{Beagle_SHAPEIT}$*,* Ref_{Beagle}*, alternative = "two.sided", paired = TRUE, conf.int = TRUE)*

[3] WGS markers of chromosome 29: 107 deletions, 2,723 indels and 326,838 SNPs

[4] The chosen Beagle GP threshold was 0.99

[5] On each dataset, three different tests were performed; therefore, a *P-value* threshold of 0.017 (0.05/3) was selected at 5% level of significance, after multiple-testing correction

[6] LoT: Leave-out-trials

[7] Downstram imputation performance of the two reference panels to impute from the 777k to the WGS SNPs, indels and deletions of chromosome 29 using Minimac

[8] The *P-value* $<1.0 \times 10^{-16}$

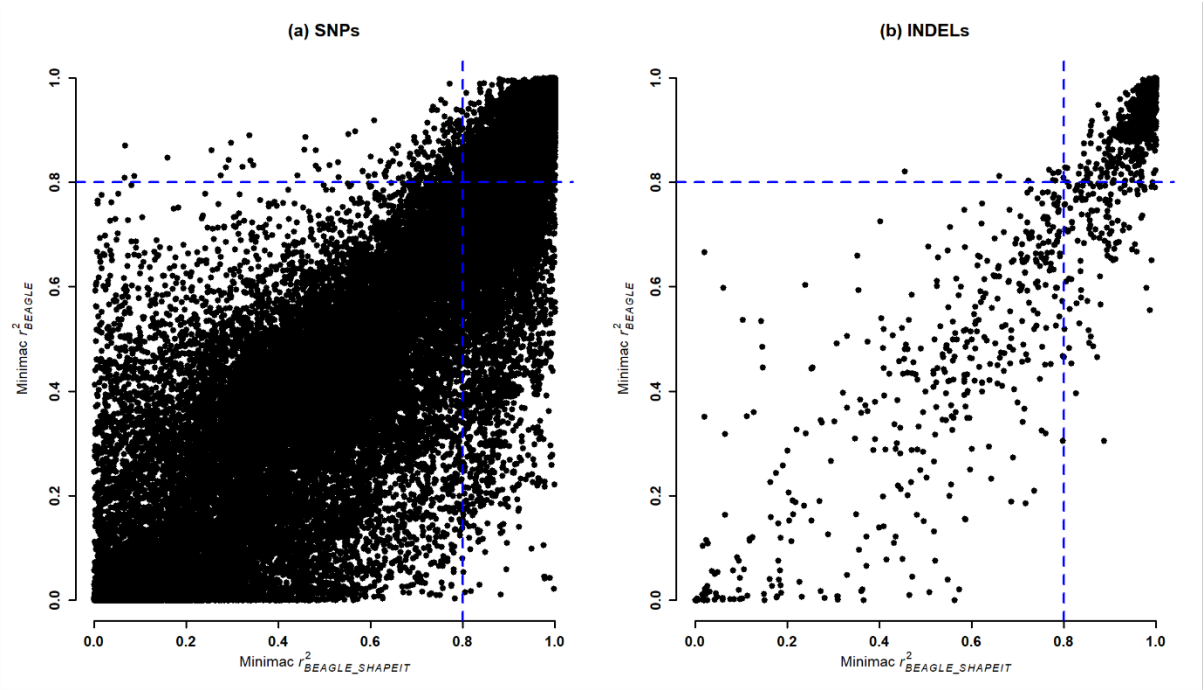


Figure 4.2. Downstream imputation performance between Ref$_{Beagle}$ vs. Ref$_{Beagle_SHAPEIT}$ on test chromosome 29. *(a) SNPs. In total, 326,838 SNPs were imputed from this chromosome; each dot represents a SNP. (b) INDELs. There were 2,723 imputed indels in this chromosome; each dot represents an indel. The x- and y-axis represents corresponding Minimac r^2 value; dashed lines indicate Minimac r^2 value of 0.80. Ref$_{Beagle}$: reference panel phased using Beagle; Ref$_{BeagleSHAPEIT}$: reference panel phased using Beagle followed by re-phasing using SHAPEIT.*

Table 4.2. Accuracy of imputation from the 777k genotypes to the WGS SNPs, indels and deletions of the 29 bovine autosomes

Category	Minimac r^2 (SD)		
	SNPs	**INDELs**	**Deletions**
Overall	0.928 (0.207)	0.900 (0.217)	0.904 (0.220)
Common variants (MAF >5%)	0.954 (0.163)	0.927 (0.176)	0.957 (0.148)
Low frequency variants (1≤ MAF ≤5%)	0.889 (0.250)	0.874 (0.245)	0.927 (0.154)
Rare variants (MAF <1%)	0.527 (0.392)	0.549 (0.370)	0.713 (0.343)
0.1%	0.119 (0.183)	0.173 (0.222)	0.313 (0.365)
0.2%	0.201 (0.231)	0.262 (0.265)	0.633 (0.364)
0.3%	0.275 (0.269)	0.349 (0.271)	0.785 (0.280)
0.4%	0.330 (0.306)	0.388 (0.294)	0.764 (0.294)
0.5%	0.443 (0.348)	0.498 (0.328)	0.822 (0.243)
0.6%	0.503 (0.371)	0.552 (0.351)	0.819 (0.217)
0.7%	0.601 (0.370)	0.617 (0.353)	0.846 (0.232)
0.8%	0.653 (0.377)	0.671 (0.356)	0.803 (0.275)
0.9%	0.718 (0.355)	0.691 (0.350)	0.880 (0.180)
1.0%	0.731 (0.357)	0.722 (0.341)	0.871 (0.207)

Total number of imputed WGS SNPs = 14,070,960, Indels = 122,054, and deletions = 5,730

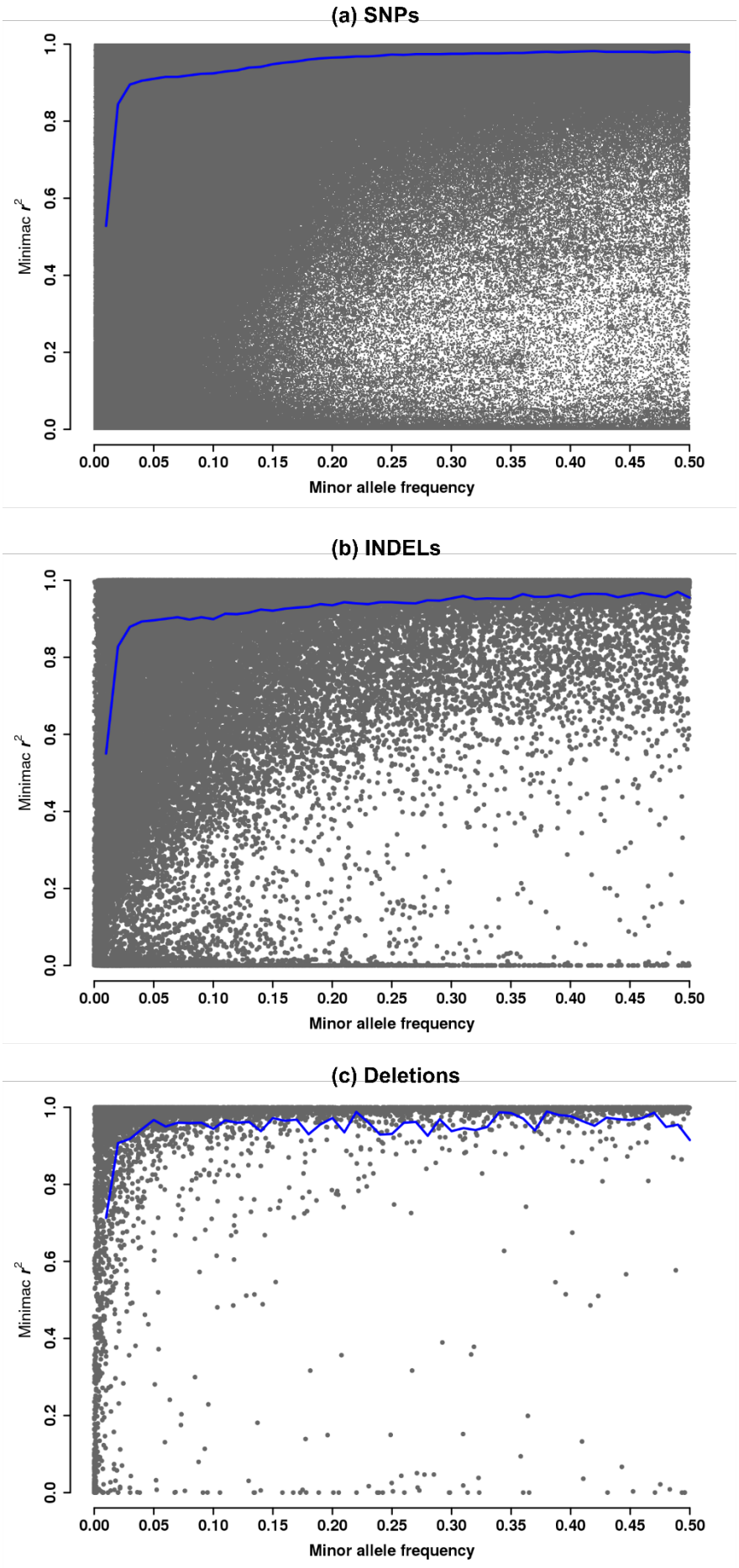

(a) SNPs
Minimac r^2
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
Minor allele frequency
(b) INDELs
Minimac r^2
Minor allele frequency
(c) Deletions
Minimac r^2
Minor allele frequency

Figure 4.3. Imputation accuracy of WGS SNPs, indels and deletions on the 29 bovine autosomes. *Each dot represents an imputed marker with MAF >0. There are (a) 14,070,960 SNPs, (b) 122,054 indels and (c) 5,730 deletions. Here, the imputation accuracy were greater than 0.80 for 91% SNPs, 85% indels and 87% deletions. The solid line in each scatterplot represents the average Minimac r^2 in MAF bins of 1% (see Table S3 for corresponding r^2 values with standard deviations). For the imputed variants, the Pearson's correlation coefficients, r, between the MAF in the reference panel vs. the target population were 0.98 for SNPs, 0.97 for indels and 0.98 for deletions. The observed average differences in MAF, between target vs. reference, were 0.0023 (two-tailed t-test P <1.0×10^{-16}) for SNPs, 0.0036 (P = 6.7×10^{-12}) for indels and 0.0009 (P = 0.73) for deletions.*

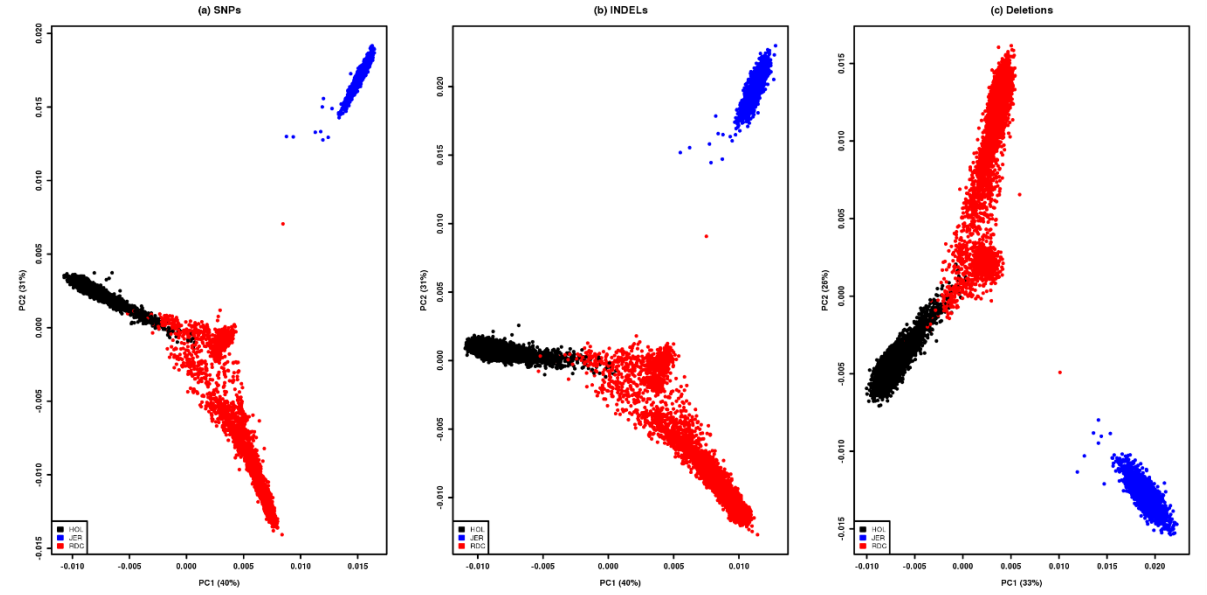


Figure 4.4. Principal component analysis using imputed WGS SNPs, indels and deletion. *Here, the analyses were performed on (a) ~137k SNPs, (b) ~12k indels and ~4.5k deletions using PLINK v2. These variants were randomly selected from the imputed WGS variants with MAF >1%. PC: principal component; HOL: Holstein; JER: Jersey; RDC: Nordic Red Dairy cattle.*

Acknowledgement

This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B). Md Mesbah-Uddin benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate 'EGS-ABG'. Authors also acknowledge the 1,000 Bull Genomes Project for sharing the VCF files. We are also grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources.

Availability of data and materials

All relevant data are included in the manuscript and its additional files. WGS SNPs and INDEL are available at http://www.1000bullgenomes.com/doco/1000bulls_v6_annotated_snps.tab.gz, http://www.1000bullgenomes.com/doco/1000bulls_v6_annotated_indels.tab.gz, and large deletions are at ftp://ftp.ebi.ac.uk/pub/databases/dgva/estd234_Mesbah-Uddin_et_al_2017/. The imputation pipelines are available at https://github.com/MMesbahU/ImputeDelPipeline.git.

Authors' contributions

MMU, BG, MSL and GS conceived and designed the study. MMU performed computational analyses. MMU drafted the manuscript. BG, MSL, DB and GS jointly supervised the study. BG, DB and GS critically revised the manuscript. All authors read and approved the final manuscript.

4.6 References

Abyzov, A., A. E. Urban, M. Snyder, and M. Gerstein. 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21:974-984. https://doi.org/10.1101/gr.114876.110.

Benjamini, Y. and T. P. Speed. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40:e72. https://doi.org/10.1093/nar/gks001.

Bickhart, D. M. and G. E. Liu. 2014. The challenges and importance of structural variation detection in livestock. Front Genet 5:37. https://doi.org/10.3389/fgene.2014.00037.

Boichard, D., M. Boussaha, A. Capitan, D. Rocha, C. Hozé, M. P. Sanchez, T. Tribout, R. Letaief, P. Croiseau, C. Grohs, W. Li, C. Harland, C. Charlier, M. S. Lund, G. Sahana, M. Georges, S. Barbier, W. Coppieters, S. Fritz, and B. Guldbrandtsen. 2018. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. Page 675 in Proc. Proc. World Congr. Genet. Appl. Livest. Prod., Auckland, New Zealand. AL Rae Centre for Genetics and Breeding,

Massey University, Palmerston North, New Zealand. http://www.wcgalp.org/system/files/proceedings/2018/experience-large-scale-use-eurogenomics-custom-snp-chip-cattle.pdf.

Boussaha, M., D. Esquerre, J. Barbieri, A. Djari, A. Pinton, R. Letaief, G. Salin, F. Escudie, A. Roulet, S. Fritz, F. Samson, C. Grohs, M. Bernard, C. Klopp, D. Boichard, and D. Rocha. 2015. Genome-Wide Study of Structural Variants in Bovine Holstein, Montbeliarde and Normande Dairy Breeds. PLoS One 10:e0135931. https://doi.org/10.1371/journal.pone.0135931.

Bouwman, A. C., H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, H. Pausch, R. F. Brøndum, P. J. Bowman, B. Thomsen, B. Guldbrandtsen, M. S. Lund, B. Servin, D. J. Garrick, J. Reecy, J. Vilkki, A. Bagnato, M. Wang, J. L. Hoff, R. D. Schnabel, J. F. Taylor, A. A. E. Vinkhuyzen, F. Panitz, C. Bendixen, L. E. Holm, B. Gredler, C. Hozé, M. Boussaha, M. P. Sanchez, D. Rocha, A. Capitan, T. Tribout, A. Barbat, P. Croiseau, C. Drögemüller, V. Jagannathan, C. Vander Jagt, J. J. Crowley, A. Bieber, D. C. Purfield, D. P. Berry, R. Emmerling, K. U. Götz, M. Frischknecht, I. Russ, J. Sölkner, C. P. Van Tassell, R. Fries, P. Stothard, R. F. Veerkamp, D. Boichard, M. E. Goddard, and B. J. Hayes. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet 50:362-367. https://doi.org/10.1038/s41588-018-0056-5.

Brøndum, R. F., B. Guldbrandtsen, G. Sahana, M. S. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics 15:728. https://doi.org/10.1186/1471-2164-15-728.

Browning, B. L. and S. R. Browning. 2016. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 98:116-126. https://doi.org/10.1016/j.ajhg.2015.11.020.

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

Charlier, C., J. S. Agerholm, W. Coppieters, P. Karlskov-Mortensen, W. Li, G. de Jong, C. Fasquelle, L. Karim, S. Cirera, N. Cambisano, N. Ahariz, E. Mullaart, M. Georges, and M. Fredholm. 2012. A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina. PLoS One 7:e43085. https://doi.org/10.1371/journal.pone.0043085.

Chen, L., A. J. Chamberlain, C. M. Reich, H. D. Daetwyler, and B. J. Hayes. 2017. Detection and validation of structural variations in bovine whole-genome sequence data. Genet Sel Evol 49:13. https://doi.org/10.1186/s12711-017-0286-5.

Conrad, D. F., D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. Macarthur, J. R. Macdonald, I. Onyiah, A. W. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Wellcome Trust Case Control, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles. 2010. Origins and functional impact of copy number variation in the human genome. Nature 464:704-712. https://doi.org/10.1038/nature08516.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerre, O. Bouchez, M. N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46:858-865. https://doi.org/10.1038/ng.3034.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and G. Genomes Project Analysis. 2011. The variant call format and VCFtools. Bioinformatics 27:2156-2158. https://doi.org/10.1093/bioinformatics/btr330.

Das, S., L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P. R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, and C. Fuchsberger. 2016. Next-generation genotype imputation service and methods. Nat Genet 48:1284-1287. https://doi.org/10.1038/ng.3656.

Delaneau, O., B. Howie, A. J. Cox, J. F. Zagury, and J. Marchini. 2013a. Haplotype estimation using sequencing reads. Am J Hum Genet 93:687-696. https://doi.org/10.1016/j.ajhg.2013.09.002.

Delaneau, O., J. Marchini, C. Genomes Project, and C. Genomes Project. 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun 5:3934. https://doi.org/10.1038/ncomms4934.

Delaneau, O., J. F. Zagury, and J. Marchini. 2013b. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 10:5-6. https://doi.org/10.1038/nmeth.2307.

Gao, Y., J. Jiang, S. Yang, Y. Hou, G. E. Liu, S. Zhang, Q. Zhang, and D. Sun. 2017. CNV discovery for milk composition traits in dairy cattle using whole genome resequencing. BMC Genomics 18:265. https://doi.org/10.1186/s12864-017-3636-3.

Genomes Project, C., A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. 2015. A global reference for human genetic variation. Nature 526:68-74. https://doi.org/10.1038/nature15393.

Handsaker, R. E., J. M. Korn, J. Nemesh, and S. A. McCarroll. 2011. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43:269-276. https://doi.org/10.1038/ng.768.

Handsaker, R. E., V. Van Doren, J. R. Berman, G. Genovese, S. Kashin, L. M. Boettger, and S. A. McCarroll. 2015. Large multiallelic copy number variations in humans. Nat Genet 47:296-303. https://doi.org/10.1038/ng.3200.

Letaief, R., E. Rebours, C. Grohs, C. Meersseman, S. Fritz, L. Trouilh, D. Esquerre, J. Barbieri, C. Klopp, R. Philippe, V. Blanquet, D. Boichard, D. Rocha, and M. Boussaha. 2017. Identification of copy number variation in French dairy and beef breeds using next-generation sequencing. Genet Sel Evol 49:77. https://doi.org/10.1186/s12711-017-0352-z.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079. https://doi.org/10.1093/bioinformatics/btp352.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303. https://doi.org/10.1101/gr.107524.110.

Mesbah-Uddin, M., B. Guldbrandtsen, T. Iso-Touru, J. Vilkki, D. J. De Koning, D. Boichard, M. S. Lund, and G. Sahana. 2018a. Genome-wide mapping of large deletions and their population-genetic properties in dairy cattle. DNA Res 25:49-59. https://doi.org/10.1093/dnares/dsx037.

Mesbah-Uddin, M., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2018b. Genotype call for chromosomal deletions using read-depth from whole genome sequence variants in cattle. Page 662 in Proc. World Congr. Genet. Appl. Livest. Prod., Auckland, New Zealand. AL Rae Centre for Genetics and Breeding, Massey University, Palmerston North, New Zealand. http://www.wcgalp.org/system/files/proceedings/2018/genotype-call-chromosomal-deletions-using-read-depth-whole-genome-sequence-variants-cattle.pdf.

Mielczarek, M., M. Fraszczak, R. Giannico, G. Minozzi, J. L. Williams, K. Wojdak-Maksymiec, and J. Szyda. 2017. Analysis of copy number variations in Holstein-Friesian cow genomes based on whole-genome sequence data. J Dairy Sci 100:5515-5525. https://doi.org/10.3168/jds.2016-11987.

Pausch, H., I. M. MacLeod, R. Fries, R. Emmerling, P. J. Bowman, H. D. Daetwyler, and M. E. Goddard. 2017. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet Sel Evol 49:24. https://doi.org/10.1186/s12711-017-0301-x.

Quinlan, A. R. and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842. https://doi.org/10.1093/bioinformatics/btq033.

R Core Team. 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Shin, D. H., H. J. Lee, S. Cho, H. J. Kim, J. Y. Hwang, C. K. Lee, J. Jeong, D. Yoon, and H. Kim. 2014. Deleted copy number variation of Hanwoo and Holstein using next generation sequencing at the population level. BMC Genomics 15:240. https://doi.org/10.1186/1471-2164-15-240.

Wu, X., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2016. Association analysis for feet and legs disorders with whole-genome sequence variants in 3 dairy cattle breeds. J Dairy Sci 99:7221-7231. https://doi.org/10.3168/jds.2015-10705.

Zhang, Q., M. P. L. Calus, M. Bosse, G. Sahana, M. S. Lund, and B. Guldbrandtsen. 2018. Human-Mediated Introgression of Haplotypes in a Modern Dairy Cattle Breed. Genetics 209:1305-1317. https://doi.org/10.1534/genetics.118.301143.

Chapter 5.

Genome-wide association study with imputed whole-genome sequence variants including large deletions for female fertility in three Nordic dairy breeds

Md Mesbah-Uddin,[1,2*] Bernt Guldbrandtsen,[1] Aurélien Capitan[2,3], Mogens Sandø Lund,[1] Didier Boichard,[2] and Goutam Sahana,[1]

[1]Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark

[2]GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[3]Allice, 75595 Paris, France

[*]Corresponding author: mdmesbah@gmail.com

Work in Progress and Manuscript in preparation

5.1 Abstract

Female fertility is an important economic trait in dairy cattle production. In this study, we performed genome-wide association studies for eight fertility traits in Nordic Holstein (HOL), Jersey (JER) and Nordic Red Dairy cattle (RDC). The traits included were fertility index (FI) and seven component traits of FI, namely, number of inseminations per conception in heifers (AISh) or in cows (AISc), interval from calving to first insemination (ICF), interval from first to last insemination in heifers (IFLh) or in cows (IFLc), non-return rate in heifers (NRRh) or in cows (NRRc). To identify fertility associated quantitative trait loci (QTL) we performed single-marker analyses followed by stepwise selection of independent markers using conditional and joint association (COJO) analyses.

From single-marker analyses for fertility traits, we reported genome-wide significant association (*P-value* $<5\times10^{-8}$) of 30,384 SNPs, 178 indels and 3 deletions in HOL, 17 SNPs in JER, and 23,481 SNPs, 189 indels and 13 deletions in RDC. After COJO analyses of FI associated markers, we identified 37 and 23 independent associations in HOL and RDC, respectively. For these QTL region (top marker ±500 kb), we proposed several candidate genes that had functional annotations such as embryonic lethality, male and female infertility, oocyte degeneration, abnormal estrous cycle, decreased ovulation rate, and other fertility related phenotypes in mouse, cattle, and zebrafish.

Inclusion of these QTL information in genomic prediction is expected to provide gains in accuracy. However, to maximize this gain, direct genotyping of these top markers could be considered by including them in the SNP array used for routine genotyping. Finally, validation of these QTL in an independent population should also be considered to ascertain the true nature of these associations.

Key words: GWAS, fertility, deletion, Holstein, Jersey, Nordic Red, dairy cattle

5.2 Introduction

Over the years, genome-wide association studies (GWASs) have provided a better understanding of underlying genetic architectures for many economical traits in dairy cattle. These include milk yield (Iso-Touru et al., 2016), milk compositions (Gebreyesus et al., 2019, Sanchez et al., 2019), milking speed (Jardim et al., 2018, Marete et al., 2018), clinical mastitis (Kadri et al., 2015, Cai et al., 2018), body conformation (Abo-Ismail et al., 2017), stature (Bouwman et al., 2018), and so on. Fertility is also an important economic trait that was difficult to select before genomic selection due to its very low heritability. GWAS for female fertility provided an opportunity to detect quantitative trait loci (QTL) (Hoglund et al., 2014, Hoglund et al., 2015a, Hoglund et al., 2015b) and to include these QTL in genomic prediction (Brøndum et al., 2015). However, the accuracy of prediction for female fertility is lower compared with production traits. There is a potential to identify QTL or markers in strong linkage disequilibrium (LD) with QTL using GWAS for female fertility to utilize the QTL information in breeding value prediction. Two factors could be considered

to improve GWAS signals even with a fixed size of the mapping population. The first one is to improve accuracy of imputation, and the second is to include other classes of genetic markers, besides SNPs and indels, such as chromosomal deletions, duplications, insertions, to name a few. In earlier studies, we attempted to address both factors and reported substantial gains in accuracy of imputation for WGS SNPs, indels and large chromosomal deletions (thereafter named deletions) in three Nordic dairy cattle breeds, namely, HOL, JER, and RDC (presented in **Chapter 3** and **4** of this thesis). In this study, we performed GWAS for eight fertility traits in these three breeds using imputed WGS dataset comprise of SNP, indel and deletion.

5.3 Methods

Genotypes and phenotypes

We used the imputed WGS dataset comprised of SNPs, indels and deletions from **Chapter 4**, where we performed a two-step imputation from the Illumina BovineSNP50 to BovineHD BeadChip (777k, hereafter) genotypes, followed by joint imputation from the 777k genotypes to WGS SNPs, indels and deletions. The phasing was performed using a combination of Beagle v4.r1274 (Browning and Browning, 2016) and SHAPEIT v2.r837 (Delaneau et al., 2013) software following the method reported in (Delaneau et al., 2014), and joint imputation was performed using Minimac3 v.2.0.1 software (Das et al., 2016). Overall, average imputation accuracies (Minimac r^2) were 0.93 for SNPs, 0.90 for indels and 0.90 for deletions. From this imputation dataset (which had several filters for low quality markers including deviation from Hardy-Weinberg proportion, *P-value* $\leq 1 \times 10^{-5}$), we considered markers with minor allele frequency (MAF) >1% and Minimac r^2 value ≥ 0.1 for GWAS. In the final dataset, we had 5,596 HOL, 1,215 JER and 4,507 RDC bulls with both genotypes and phenotypes (de-regressed estimated breeding values). At this MAF and Minimac r^2, the genotype dataset comprised 12,174,287 (12,076,004 SNPs; 94,581 indels; 3,702 deletions) markers in HOL, 10,135,228 (10,057,293 SNPs; 74,829 indels; 3,106 deletions) markers in JER and 12,855,372 (12,748,285 SNPs, 102,843 indels, 4,244 deletions) markers in RDC. For each breed, the phenotype dataset comprised of de-regressed breeding values for eight fertility traits: namely, number of inseminations per conception in heifers (AISh) or in cows (AISc), interval (number of days) from calving to first insemination in cows (ICF), interval (number of days) from first to last insemination in heifers (IFLh) or in cows (IFLc), fertility index (FI), and non-return rate in heifers (NRRh) or in cows (NRRc) (for details see (NAV, 2013)). Summary of these phenotypes was presented in Table 5.1.

Genome-wide association study

We performed single-marker association analysis using following mixed linear model:

$$\boldsymbol{y} = \mathbf{1}_n\mu + b\boldsymbol{x} + \sum_{i=1}^{10} c_i \boldsymbol{PC_i} + \boldsymbol{g} + \boldsymbol{\varepsilon}$$

where **y** is a vector of phenotypic records for fertility traits, $\mathbf{1}_n$ is a vector of ones, μ is the mean of the given trait, b is the additive allele substitution effect of the candidate marker to be tested, $\boldsymbol{x}$ is a vector of allele dosages (coded as 0, 1 or 2), $\boldsymbol{PC_i}$ is a vector of the principal component i (first 10 PCs were considered) and c_i is the corresponding coefficients, $\boldsymbol{g}$ is a vector of polygenic effect with $\boldsymbol{g} \sim N\left(0, \boldsymbol{G}\sigma_g^2\right)$, where $\boldsymbol{G}$ is the genomic relationship matrix calculated using the 777 k markers without the test chromosome (leave-one-chromosome-out (LOCO) approach), and $\boldsymbol{\varepsilon}$ is the vector of residual effect with $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{I}\sigma_\varepsilon^2)$, where $\boldsymbol{I}$ is an identity matrix. Differences in reliabilities of de-regressed proofs were neglected because of the large number of daughters. The analysis was performed using GCTA software (version 1.92.1 beta6) (Yang et al., 2011a). To declare an association with the trait, we used a genome-wide significant threshold of *P-value* $<5 \times 10^{-8}$, considering 1 million independent test at 5% level of significance.

We estimated the genomic inflation, *lambda*, for the single-marker GWAS *P-values*, using GenABEL software "*estlamda*" function (Aulchenko et al., 2007).

Conditional and joint association analysis

Next, to identify (additional) independent signal(s) from each significant GWAS locus, we performed stepwise model selection using GCTA-COJO analysis (Yang et al., 2012). Here, we used a LD threshold of 0.80 to identify independent marker(s), and assumed complete independence for markers ±10 Mb away from a conditional marker. After each iteration the marker with lowest *P-value* was selected when it passed our GWAS threshold of $<5 \times 10^{-8}$ and was included to conditional marker list to perform next iteration. When there were multiple markers with same (lowest) *P-value*, the first marker based on chromosome position was arbitrarily chosen. Finally, joint association analysis was perform on markers selected from the same chromosome.

Identification of candidate genes

We used bovine reference genome assembly UMD3.1 (Zimin et al., 2009, Elsik et al., 2016) for all the genomic coordinates presented in this study. To identify candidate genes within a QTL, we used *Ensembl* (release 94) webserver (Zerbino et al., 2018) and manually inspected 1 Mb genomic region centering at the top marker (i.e. top marker ±500 kb surrounding region). From *Ensembl*, we retrieved annotations for genes and known phenotypes associated with those genes. However, when multiple genes were found to be located within the interval, for reporting, we preferred genes with fertility related phenotypes in other species (if available).

5.4 Results and Discussion

We performed single-marker association analysis for eight female fertility traits (Table 5.1) in Holsten, Jersey and Nordic Red Dairy cattle using imputed WGS SNPs, indels and deletions. Association test statistics were inflated (Table 5.2) when we only accounted for relatedness using genomic relationship matrix (GRM) derived from the 777 k markers in a LOCO approach. Therefore, following Jiang et al. (2019), we adjusted

for both population stratification, using the first ten principal components (PCs) (calculated from a random set of 10% markers selected from the 29 autosomes) as fixed effects, and relationships among animals using the LOCO approach. Genomic inflation rates, *lambda*, before and after fitting the PCs were presented in Table 5.2. We observed a decrease in *lambda* values after the inclusion of PCs as cofactor, although it was still very far from the standard value of "1" under the assumption of the distribution of test statistics under the null-hypothesis. However, when a trait is under polygenic inheritance, i.e. when multiple (small effect) causal variants underlie the trait variation (also known as infinitesimal model), a substantial inflation of GWAS test statistics is expected even in the absence of population structure (Yang et al., 2011b). The extent of this inflation is governed by several factors, namely, heritability of the trait, study sample size, pattern of linkage-disequilibrium, and number of causal variants (for detail see Yang et al. (2011b)). The genomic inflation seen in this study despite the correction for population stratification and relatedness could be explained by the nature of polygenic inheritance of these fertility traits we analyzed, e.g. several non-zero effect variants in Bayesian mixture models as reported by Brøndum et al. (2015). The phenotypes used here are de-regressed breeding values for bulls with high reliability and are equivalent to a high heritability trait. The long-range LD pattern seen in the study population (de Roos et al., 2008, van den Berg et al., 2016) could also explain inflated *lambda* values observed in this study. Among the three breeds, HOL had the highest sample size and therefore, highest power to identify QTL, was also the breed where we observed highest *lambda* values. However, the adjustment for both population stratification and relatedness could be very conservative; nonetheless, we expect that the test statistics will be robust against spurious associations and false discoveries.

The summary of single-marker associations, for eight female fertility traits in HOL, JER and RDC, is presented in Table 5.3 and corresponding Manhattan plots are presented in Figure 5.1, 5.2 and 5.3, respectively.

Deletions associated with female fertility traits

In HOL, we identified three deletions that had significant genome-wide association (*P-value* $<5\times10^{-8}$) with ICF, FI and NRRc, respectively (Table 5.4). We identified three candidate genes, ***ACSS1***, ***CFAP61*** and ***HELLS***, within these deletion loci (±500 kb), that are known to be associated with embryonic lethality, premature death, and male infertility in mouse (Table 5.5).

In RDC, 13 deletions reached genome-wide significance; associated phenotypes include seven fertility traits (Table 5.4). We did not find significant association between deletions and ICF. Among these 13 deletions, strongest association (*p-value* = 3.5×10^{-42}) was seen for a previously known deletion on chromosome 12 located between position 20,100,648 and 20,763,119 bp (esv4015629) that cause embryonic lethality in cattle (Kadri et al., 2014). The esv4015629 had significant association with all fertility traits analyzed, except ICF as expected. Kadri et al. (2014) also did not find association between esv4015629 and ICF. Several deletion loci showed association with multiple fertility traits. For example, FI, IFLc and AISc are the three highly correlated traits (correlation of 0.97 and 0. 91 between FI vs. IFLc, and FI vs. AISc, respectively (NAV,

2013)) that had seven associated deletion loci in common (Table 5.4). Several genes within these loci have known fertility related phenotypes, including embryonic lethality, in mouse (Table 5.5).

In JER, no deletion locus reached genome-wide significance. The top five associations included esv4014584 (Chr13:50733796-50735009; associated with ICF, *P-value* = 3.2×10^{-6}), esv4018559 (Chr6:33504998-33506227; associated with AISc, *P-value* = 3.4×10^{-6}), esv4019122 (Chr6:33738269-33738743; associated with AISc, *P-value* = 3.4×10^{-6}), esv4017388 (Chr1:145570019-145570346; associated with NRRc, *P-value* = 8.5×10^{-6}), and esv4017421 (Chr3:7140763-7141120; associated with NRRh, *P-value* = 1.8×10^{-5}). Among these five deletions, only esv4017388 located within a QTL associated with NRRc (Figure 5.1h). Candidate genes for this locus include ***ADARB1*** (ENSBTAG00000017486; 52 kb upstream of esv4017388) and ***POFUT2*** **(**ENSBTAG00000007818; 467 kb downstream of esv4017388). In mouse, ***ADARB1*** causes postnatal lethality (MGI: 891999), and ***POFUT2*** causes embryonic lethality during organogenesis as well as lethality between implantation and somite formation (MGI: 1916863).

Fertility index (FI) associated SNPs and indels

For SNPs and indels, we only presented FI associated independent GWAS loci. The seven FI component traits need further analyses to identify independent as well as pleotropic loci. COJO and multi-trait Meta analyses for these traits are in progress.

FI associated loci in HOL. In HOL, 30,384 SNPs and 178 indels in total had reached GWAS significant *P-value* threshold for the eight fertility traits (Table 5.3 and Figure 5.1). After COJO analysis, we identified 37 FI associated independent GWAS loci out of these significant markers (Table 5.6). In Table 5.7, we presented the candidate genes for these loci. Top three FI associated loci are rs434006863 in chromosome 13 (*P-value* = 1.5×10^{-22}), rs380439408 in chromosome 24 (*P-value* = 5.4×10^{-14}) and rs380495923 in chromosome 10 (*P-value* = 2.0×10^{-12}). The candidate gene for rs434006863, rs380439408 and rs380495923 are ***MPP7***, ***CDH2*** and ***TBPL2***, respectively. ***MPP7*** has known role in bone mass density in zebrafish [ZFIN: ZDB-GENE-991209-8]; ***CDH2*** causes embryonic lethality in mouse [MGI: 88355]; and ***TBPL2*** gene decreases rate of embryo development in zebrafish [ZFIN: ZDB-GENE-040520-3], and causes female infertility and impairs folliculogenesis in mouse [MGI: 2684058] (for detail, see Table 5.7).

FI associated loci in JER. For JER, 17 SNPs were significantly associated with either ICF, IFLc or NRRc (Table 5.3 and Figure 5.2). In an earlier study, Hoglund et al. (2015b) reported six FI associated QTL with *P-value* <5.6×10^{-9} for this JER population. However, in this study, we did not find any significant association with FI for JER. Two factors, namely, sample size and control for genomic inflation, could explain this situation. For JER, we had ~4 times fewer samples compared with HOL and RDC, and QTL detection power was therefore lower; however, sample size between this two studies did not changed significantly (1,225 vs. 1,211 in this study). Difference between studies can be explained by the way to account for genomic inflation, much more stringent in this study: *lambda* of 1.88 in Hoglund et al. (2015b) vs. 1.27 in this study.

FI associated loci in RDC. Finally, in RDC, we identified 23,481 SNPs and 189 indels that had genome-wide significant association with fertility traits (Table 5.3). After COJO analysis, we reported 23 FI associated independent GWAS loci, nine of which are located in chromosome 12 (Table 5.8). The LD pattern (correlation) of these nine lead markers is presented in Table 5.9. Functional annotations of the genes located within these QTL region (top markers ±500 kb) is presented in Table 5.10. Interestingly, the candidate genes within these loci also had known fertility related functions in cattle, mouse, and zebrafish (Table 5.10).

5.5 Conclusion

In this study, we identified several novel QTL for female fertility in Nordic dairy cattle. With our imputed data, we were able to identify a previously known deletion QTL in RDC; however, GWAS signals for this QTL were stronger with our imputed data than in the previous report. Other deletion loci also reached GWAS threshold but were not among the top signals of those loci. Those deletions could be in strong LD with the QTL. Interestingly, majority of the candidate genes within the reported QTL regions had established fertility related function in mouse and zebrafish. QTL identified in this study will be used in subsequent study to elucidate the effect of these markers in prediction accuracy for female fertility.

Acknowledgement

We are grateful to the Nordic Cattle Genetic Evaluation (NAV), Aarhus, Denmark for providing the phenotypic data used in this study and Viking Genetics, Randers, Denmark for proving blood/semen samples for genotyping. This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B). Md Mesbah-Uddin benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate 'EGS-ABG'. Authors also acknowledge the 1000 Bull Genomes Project for sharing the VCF files. Authors are also grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources.

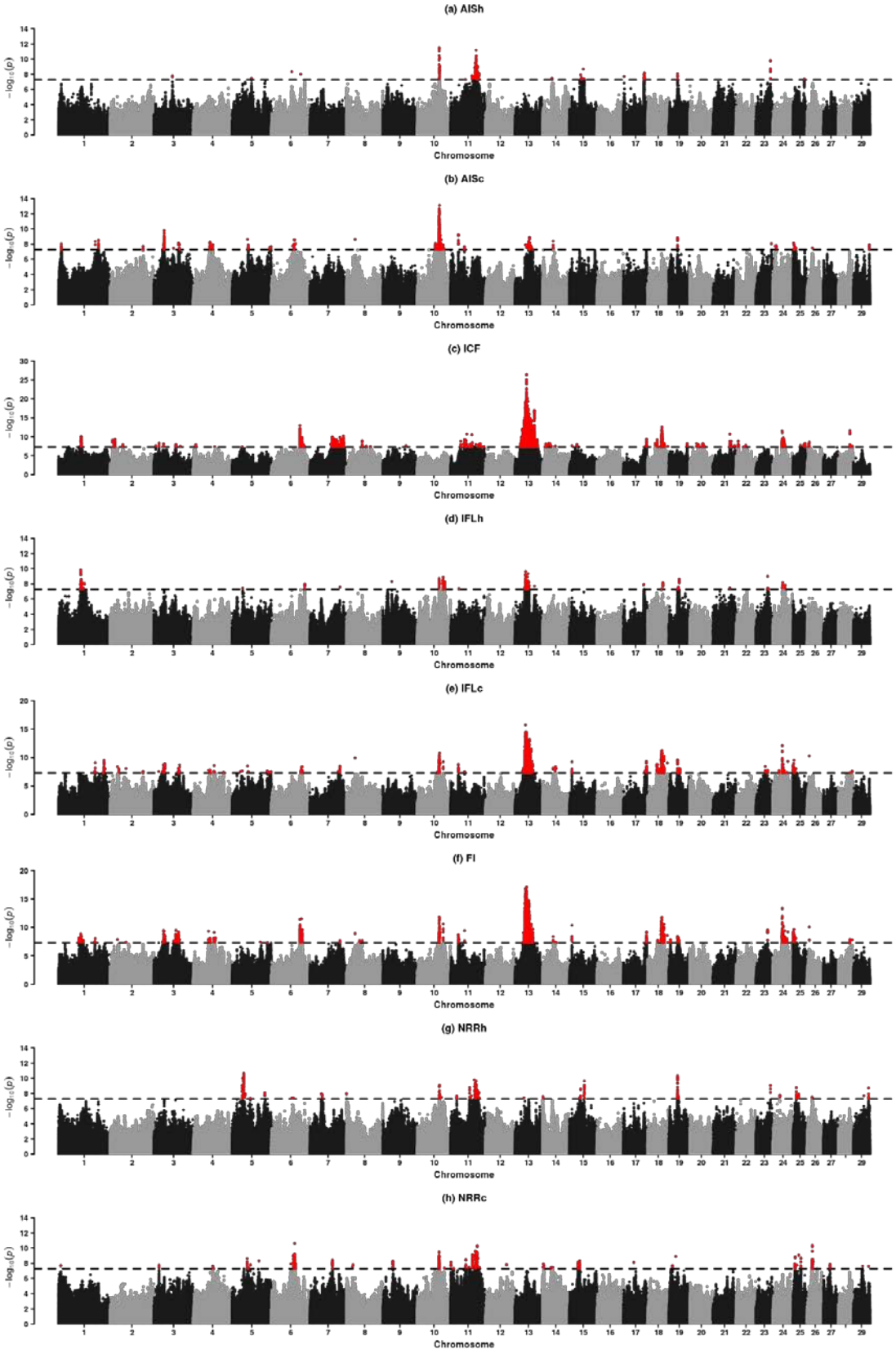


Figure 5.1. Manhattan plots of single-marker GWAS in Holstein cattle. *Each dot represent one WGS marker; dashed horizontal line is GWAS significance threshold of* 5×10^{-8}.

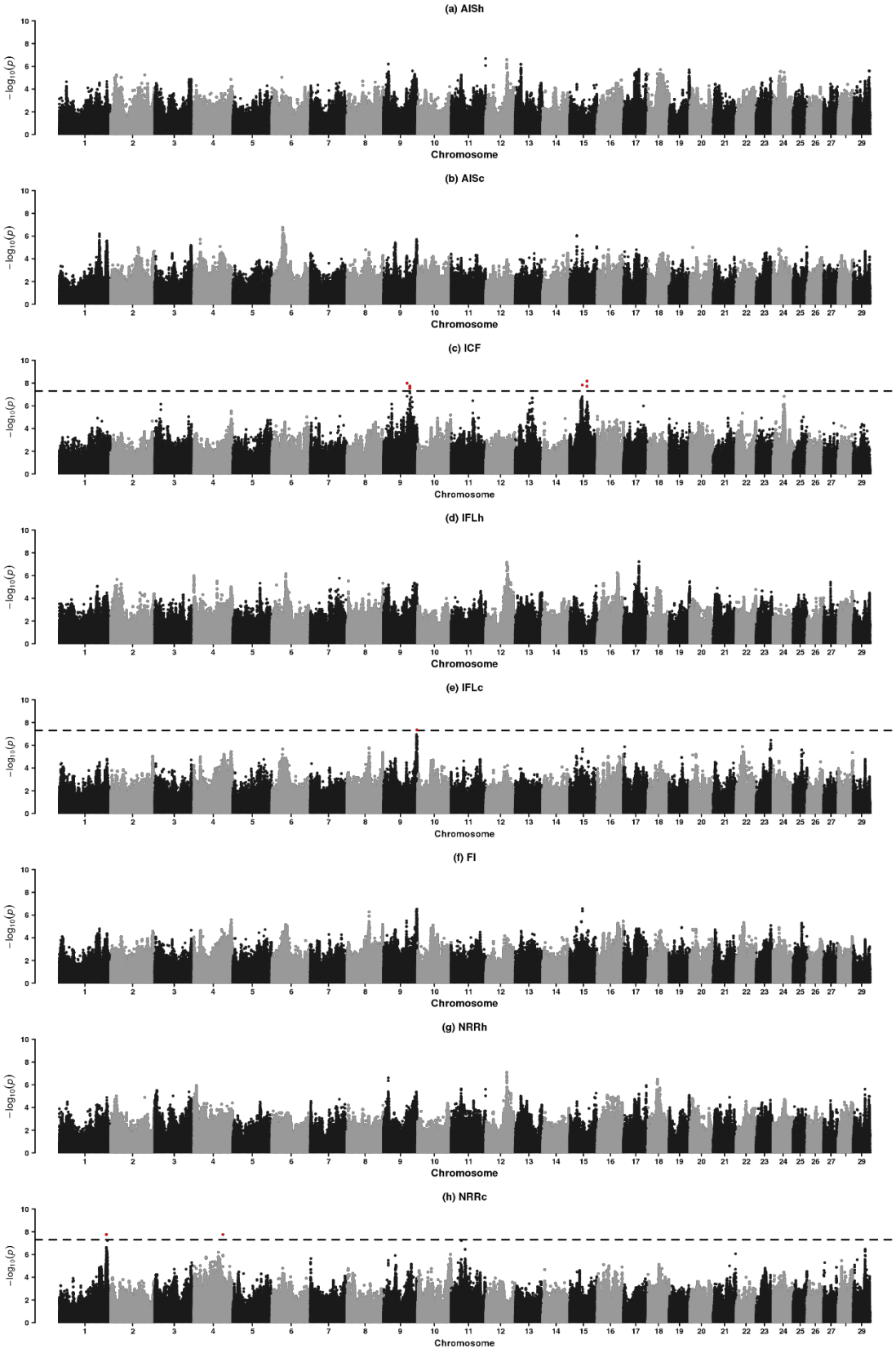


Figure 5.2. Manhattan plots of single-marker GWAS in Jersey cattle. *Each dot represent one WGS marker; dashed horizontal line is GWAS significance threshold of* 5×10^{-8}.

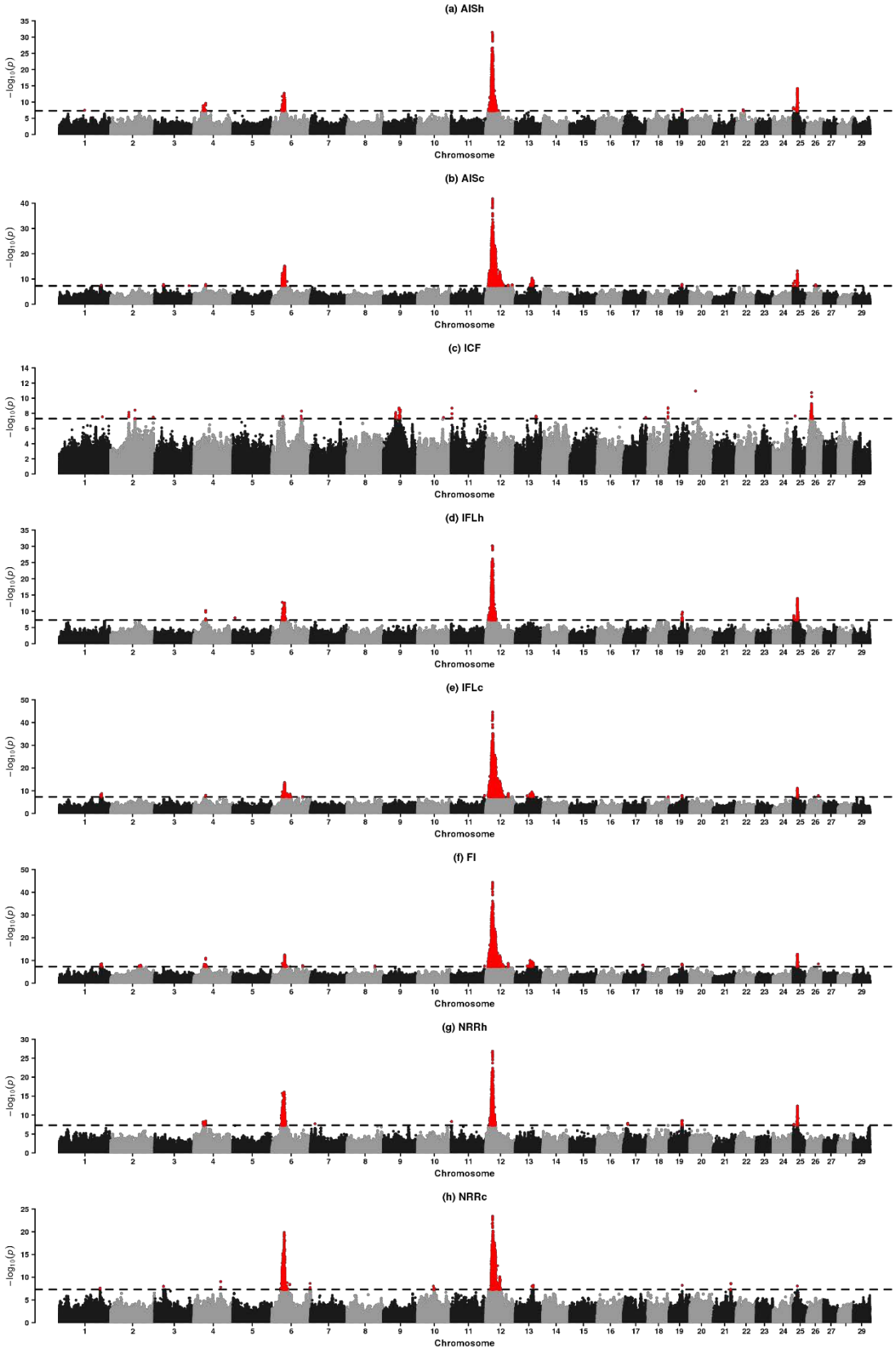


Figure 5.3. Manhattan plots of single-marker GWAS in Nordic Red Dairy cattle. *Each dot represent one WGS marker; dashed horizontal line is GWAS significance threshold of* 5×10^{-8}.

.

Table 5.1. Summary statistics of female fertility traits (de-regressed breeding values) analyzed

Trait[1]	Holstein			Jersey			Nordic Red Dairy Cattle		
	Number	Mean (SD)	Range	Number	Mean (SD)	Range	Number	Mean (SD)	Range
AISc	5,342	100.30 (11.38)	65-142	1,179	99.54 (10.11)	69-126	4,291	102.57 (10.56)	73-136
AISh	5,487	100.06 (10.00)	72-131	1,204	100.22 (12.70)	60-133	4,345	95.27 (12.62)	59-132
ICF	5,342	98.38 (11.19)	65-131	1,186	103.96 (10.9)	75-135	4,299	99.21 (10.20)	69-130
IFLc	5,361	98.1 (12.09)	60-146	1,192	99.74 (11.53)	65-136	4,309	101.63 (10.17)	73-134
IFLh	5,478	98.75 (10.97)	68-133	1,199	101.66 (11.37)	68-134	4,341	95.18 (11.64)	62-129
FI	5,577	98.86 (10.77)	68-142	1,211	101.14 (11.42)	69-131	4,471	100.44 (9.67)	74-129
NRRc	5,353	100.77 (10.21)	72-135	1,189	97.10 (9.67)	67-126	4,309	102.34 (10.56)	73-136
NRRh	5,487	101.65 (11.38)	70-137	1,207	99.95 (10.86)	65-130	4,351	96.76 (10.98)	65-129

[1]AISc: number of inseminations per conception in cows, AISh: number of inseminations per conception in heifers, ICF: Interval (number of days) from calving to first insemination (in cows), IFLc: interval (number of days) from first to last insemination in cows, IFLh: interval (number of days) from first to last insemination in heifers, FI: fertility index, NRRc: non-return rate in cows, NRRh: non-return rate in heifers.

Table 5.2. Genomic inflation rate (lambda) before and after correcting for population stratification using first ten principal components (PCs)

Trait[1]	Holstein		Jersey		Nordic Red Dairy Cattle	
	Without PCs	With PCs	Without PCs	With PCs	Without PCs	With PCs
AISh	1.76	1.65	1.39	1.29	1.41	1.37
AISc	2.04	1.83	1.31	1.24	1.56	1.50
ICF	2.38	2.19	1.36	1.29	1.63	1.57
IFLh	1.93	1.81	1.37	1.29	1.42	1.37
IFLc	2.24	2.02	1.32	1.25	1.59	1.53
FI	2.26	2.05	1.33	1.27	1.59	1.52
NRRh	1.99	1.84	1.37	1.29	1.41	1.38
NRRc	2.20	1.98	1.39	1.28	1.53	1.48

[1]AISc: number of inseminations per conception in cows, AISh: number of inseminations per conception in heifers, ICF: Interval (number of days) from calving to first insemination (in cows), IFLc: interval (number of days) from first to last insemination in cows, IFLh: interval (number of days) from first to last insemination in heifers, FI: fertility index, NRRc: non-return rate in cows, NRRh: non-return rate in heifers.

Table 5.3. Summary of single-marker association results (markers with P-value <5 × 10^{-8})

Trait	Holstein	Jersey	Nordic Red Dairy Cattle
AISh	583	0	10,004
AISc	2,012	0	16,089
ICF	21,539	8	280
IFLh	932	0	7,302
IFLc	7,645	1	16,289
FI	11,091	0	14,916
NRRh	1,057	0	8,036
NRRc	2,008	8	9,182

[1]AISc: number of inseminations per conception in cows, AISh: number of inseminations per conception in heifers, ICF: Interval (number of days) from calving to first insemination (in cows), IFLc: interval (number of days) from first to last insemination in cows, IFLh: interval (number of days) from first to last insemination in heifers, FI: fertility index, NRRc: non-return rate in cows, NRRh: non-return rate in heifers.

Table 5.4. Fertility associated large deletions (P-value <5×10^{-8})

Chr	Marker ID	Position	Size of Deletion	Trait	A1/A2	Freq	b	se	p
Holstein cattle									
13	esv4017813	13:42,945,993-42,946,953	961 bp	ICF	DEL/G	0.271	-1.050	0.186	1.7×10^{-8}
13	esv4018227	13:40,037,936-40,039,382	2.45 kb	FI	DEL/C	0.482	-0.894	0.155	8.1×10^{-9}
26	esv4015626	26:16,220,854-16,223,815	2.96 kb	NRRc	DEL/T	0.097	1.418	0.248	1.0×10^{-8}
Nordic Red Dairy cattle									
6	esv4013561	6:31,705,882-31,706,925	1.04 kb	NRRh	DEL/C	0.287	-1.341	0.242	3.0×10^{-8}
6	esv4018710	6:37,999,532-37,999,979	448 bp	**NRRh**, NRRc	DEL/C	0.352	-1.525	0.236	9.4×10^{-11}
6	esv4019275	6:39,473,114-39,473,611	498 bp	NRRc	DEL/A	0.389	1.260	0.215	4.9×10^{-9}
6	esv4011295	6:39,711,454-39,712,517	1.06 kb	AISc	DEL/A	0.423	-1.208	0.215	1.9×10^{-8}
6	esv4014663	6:39,856,457-39,856,658	202 bp	**NRRc**, AISc	DEL/A	0.466	1.502	0.211	1.2×10^{-12}
12	esv4019299	12:17163364-17163728	365 bp	**FI**, AISh, AISc, IFLh, IFLc, NRRh	DEL/G	0.250	-1.655	0.226	2.2×10^{-13}
12	esv4015629	12:20,100,648-20,763,119	662 kb	**FI**, AISh, AISc, IFLh, IFLc, NRRh, NRRc	DEL/G	0.137	-3.985	0.293	3.5×10^{-42}
12	esv4012561	12:20,570,478-20,571,118	641 bp	**FI**, IFLc, IFLh, AISc	DEL/A	0.092	-2.010	0.323	5.0×10^{-10}
12	esv4017726	12:22,961,548-22,961,822	275 bp	**IFLc**, AISc, FI	DEL/A	0.435	1.394	0.207	1.8×10^{-11}
12	esv4013857	12:35,299,066-35,299,339	274 bp	IFLc	DEL/T	0.127	-1.789	0.315	1.4×10^{-8}
12	esv4015687	12:43,999,233-44,009,032	9.8 kb	**IFLc**, AISc, NRRc, FI	DEL/T	0.102	-2.255	0.331	9.8×10^{-12}
25	esv4014328	25:12,757,355-12,758,442	1.09 kb	**FI**, AISh, IFLh, AISc, IFLc	DEL/G	0.059	-2.432	0.401	1.3×10^{-9}
25	esv4014654	25:13,705,829-13,706,486	658 bp	**AISh**, AISc, IFLc, IFLh, FI, NRRh	DEL/A	0.050	-3.833	0.522	2.1×10^{-13}

Chr: Chromosome; Marker ID: deletion IDs are from Database of Genomic Variants archive (DGVa), were extracted from Ensembl release 94; Position base pair positions in UMD3.1; A1/A2: Effect allele/other allele; Freq: frequency of A1 allele; *b*: allele substitution effect, se: standard error of b estimates; P: p-value from single marker GWAS. Traits: number of inseminations per conception in cows (AISc) and in heifers (AISh); interval (number of days) from calving to first insemination (ICF); interval from first to last insemination in cows (IFLc) or in heifers (IFLh); fertility index (FI); non-return rate in cows (NRRc) or in heifers (NRRh). When a deletion was associated with multiple trait, then GWAS summary for trait with strongest signal (bold face) was presented in the table. No significant deletion was found in JER breed.

Table 5.5. Annotation of fertility associated deletion loci

Marker ID	Gene	Annotation
Holstein		
esv4017813	(+127 kb) ACSS1 ENSBTAG00000004281	**ACSS1**: postnatal growth retardation, premature death, in mouse [MGI:1915988]
esv4018227	CFAP61 ENSBTAG00000009291	**CFAP61**: male infertility in mouse [MP:0001925; MGI:1926024]
esv4015626	(-363 kb) HELLS ENSBTAG00000005979	**HELLS**: postnatal lethality, preweaning lethality, neonatal lethality, perinatal lethality, in mouse [MGI:106209]
Nordic Red Dairy cattle		
esv4013561	(+115 kb) SMARCAD1 ENSBTAG00000017061	**SMARCAD1**: abnormal embryo size [MP:0001697], decreased oocyte number, decreased fetal size, reduced female fertility, reduced male fertility, premature death, abnormal fertility/fecundity, postnatal growth retardation, neonatal lethality, postnatal lethality, perinatal lethality, preweaning lethality [MGI:95453], in mouse
esv4018710	ABCG2 ENSBTAG00000017704 ; (+100 kb) PKD2 ENSBTAG00000020031	**ABCG2**: serum uric acid concentration [OMIM 138900]; **PKD2**: embryonic lethality during organogenesis, lethality throughout fetal growth and development, abnormal direction of embryo turning, in mouse [MGI:1099818]
esv4019275	(-481 kb) LCORL ENSBTAG00000046561	**LCORL**: birth weight in rat
esv4011295	gene-desert	–
esv4014663	gene-desert	–
esv4019299	(-274 kb) HTR2A ENSBTAG00000013498	**HTR2A**: prolactin secretion in turkey [GO:0070459]; decreased Paneth cell number, and abnormal intestinal smooth muscle morphology in mouse [MGI:109521]; disease of metabolism, hyperglycemia [RGD: 1624376], pancreatitis [RGD: 1624393], congestive heart failure [RGD: 1600503], and hypoxia [RGD: 1624392], in rats
esv4015629	RNASEH2B ENSBTAG00000020149 (100% deleted) ; GUCY1B2 ENSBTAG00000007457 (100% deleted); FAM124A ENSBTAG00000046408 (~79% deleted)	**RNASEH2B**: embryonic growth retardation, preweaning lethality, perinatal lethality, decreased embryo size, in mouse [MGI:1914403]
esv4012561	upstream of GUCY1B2 ENSBTAG00000007457; (-153 kb) RNASEH2B ENSBTAG00000020149	**RNASEH2B**: embryonic growth retardation, preweaning lethality, perinatal lethality, decreased embryo size, in mouse [MGI:1914403]
esv4017726	LHFPL6 ENSBTAG00000034033; (+360) FREM2 ENSBTAG00000017032	**LHFPL6**: prostate adenocarcinoma, Hepatocellular Carcinoma, Melanoma, bladder carcinoma, in humans [Cancer Gene Census https://cancer.sanger.ac.uk/census]; **FREM2**: Fraser syndrome in humans [OMIM:219000; 617666]; malformed fin in zebrafish [ZFIN: ZDB-GENE-081119-3]; absent kidney, abnormal embryo development, prenatal lethality, perinatal lethality, neonatal lethality, in mouse [MGI:2444465]
esv4013857	(-423 kb) SACS ENSBTAG00000001867; (+349 kb) FGF9 ENSBTAG00000048237	**SACS**: embryonic lethality prior to organogenesis [MP:0013292], prenatal lethality prior to heart atrial septation [MP:0013294], prenatal lethality [MP:0002080], in mouse;

		FGF9: embryonic lethality prior to tooth bud stage, abnormal embryo size, decreased fetal weight, abnormal fetal cardiomyocyte proliferation, abnormal mesenchymal cell proliferation, respiratory failure, neonatal lethality, preweaning lethality, postnatal lethality, in mouse [MGI:104723]
esv4015687	(+287 kb) KLHL1 ENSBTAG00000008647	thin cerebellar molecular layer, abnormal gait, impaired coordination, in mouse [MGI:2136335]
esv4014328	(+165 kb) ERCC4 ENSBTAG00000021773; (+314 kb) MRTFB ENSBTAG00000008728	**ERCC4** : abnormal cell morphology, abnormal liver morphology, decreased body weight, postnatal growth retardation, lethality at weaning, in mouse [MGI:1354163]; **MRTFB**: lethality throughout fetal growth and development, perinatal lethality, in mouse [MGI:3050795]
esv4014654	(+72 kb) RRN3 ENSBTAG00000004804	**RRN3**: embryonic lethality during organogenesis, decreased embryo size, embryonic growth retardation, in mouse [MGI:1925255]

Annotations are based on UMD3.1 and Ensembl release 94. Information in the parenthesis represents the distance between the top marker and the start position of the corresponding gene; MGI: Mouse Genome Informatics (http://www.informatics.jax.org/); RGD: Rat Genome Database (https://rgd.mcw.edu/wg/); MP: Mouse Phenotype (https://www.mousephenotype.org/); OMIM: Online Mendelian Inheritance in Man (https://www.omim.org/)

Table 5.6. Fertility index associated loci in Holstein cattle identified from GWAS with imputed sequence variants

Chr	Top Marker	Position	A1/A2	Freq	Single-marker analysis			Conditional and joint analysis		
					b	se	p	bJ	bJ_se	pJ
1	rs137209107	68,642,776	A/G	0.143	-1.377	0.228	1.5×10^{-9}	-1.377	0.229	1.7×10^{-9}
1	rs384339926	113,461,025	G/A	0.129	1.285	0.224	1.0×10^{-8}	1.285	0.225	1.1×10^{-8}
2	rs207792133	23,885,602	C/T	0.131	-1.284	0.226	1.4×10^{-8}	-1.284	0.227	1.5×10^{-8}
2	rs43306949	50,233,158	G/A	0.159	-1.147	0.209	3.9×10^{-8}	-1.147	0.209	4.2×10^{-8}
3	rs137588365	29,789,003	A/G	0.376	-0.997	0.159	3.6×10^{-10}	-0.997	0.160	4.1×10^{-10}
3	rs42253052	68,281,044	A/G	0.313	1.071	0.170	3.3×10^{-10}	1.071	0.171	3.8×10^{-10}
3	rs109192619	79,679,521	T/C	0.479	-0.902	0.159	1.4×10^{-8}	-0.902	0.159	1.6×10^{-8}
4	rs135727662	47,510,332	A/G	0.489	0.954	0.153	4.8×10^{-10}	0.954	0.154	5.4×10^{-10}
4	rs385277008	65,156,839	G/A	0.172	1.270	0.207	7.9×10^{-10}	1.270	0.207	9.0×10^{-10}
5	rs210968504	86,998,676	G/T	0.182	-1.127	0.205	3.7×10^{-8}	-1.127	0.205	4.0×10^{-8}
6	rs110866176	93,466,329	C/G	0.449	-1.096	0.157	3.0×10^{-12}	-1.096	0.158	3.7×10^{-12}
7	rs473589185	93,099,110	T/G	0.060	-1.827	0.327	2.3×10^{-8}	-1.827	0.328	2.5×10^{-8}
8	rs382943070	27,286,359	T/C	0.112	-1.491	0.245	1.2×10^{-9}	-1.491	0.246	1.4×10^{-9}
8	rs208577191	49,271,809	G/A	0.323	0.953	0.170	2.2×10^{-8}	0.953	0.171	2.4×10^{-8}
10	rs380495923	68,425,703	T/A	0.320	-1.186	0.168	1.6×10^{-12}	-1.186	0.169	2.0×10^{-12}
10	rs134897493	81,159,267	T/C	0.354	-1.123	0.168	2.4×10^{-11}	-1.123	0.169	2.9×10^{-11}
11	rs134324803	23,631,737	G/C	0.242	-1.087	0.181	2.0×10^{-9}	-1.087	0.182	2.3×10^{-9}
11	rs209793386	42,987,167	T/G	0.168	1.307	0.209	3.9×10^{-10}	1.307	0.210	4.4×10^{-10}
13	rs384444008	25,178,930	G/T	0.214	1.040	0.190	4.6×10^{-8}	1.301	0.196	3.0×10^{-11}
13	rs385034802	29,587,693	T/C	0.445	0.345	0.159	2.9×10^{-2}	0.990	0.168	4.2×10^{-9}
13	rs434006863	36,616,682	T/C	0.148	1.849	0.215	8.8×10^{-18}	2.192	0.224	1.5×10^{-22}
13	rs136213414	51,160,223	A/G	0.157	-1.437	0.215	2.6×10^{-11}	-1.437	0.216	3.1×10^{-11}
14	rs42685926	34,562,085	C/G	0.132	1.361	0.232	4.4×10^{-9}	1.361	0.233	4.9×10^{-9}
15	rs109077219	8,194,073	T/G	0.395	1.054	0.160	4.4×10^{-11}	1.054	0.161	5.2×10^{-11}
17	rs41853464	72,010,870	A/G	0.371	-0.999	0.162	7.0×10^{-10}	-0.999	0.162	7.9×10^{-10}
18	rs382112789	31,104,391	T/G	0.175	-1.166	0.208	2.0×10^{-8}	-1.166	0.208	2.2×10^{-8}
18	rs450967176	44,576,200	T/G	0.117	-1.717	0.244	1.9×10^{-12}	-1.717	0.245	2.3×10^{-12}
18	rs382522431	61,080,660	A/T	0.135	1.380	0.232	2.8×10^{-9}	1.380	0.233	3.1×10^{-9}
19	rs383235276	4,888,887	T/C	0.253	1.053	0.186	1.4×10^{-8}	1.053	0.186	1.6×10^{-8}
19	rs41911242	26,618,825	C/T	0.316	-0.978	0.166	3.8×10^{-9}	-0.978	0.166	4.2×10^{-9}
23	rs715807267	36,483,718	A/G	0.100	1.630	0.259	2.9×10^{-10}	1.630	0.259	3.3×10^{-10}
24	rs380439408	29,556,826	A/C	0.233	-1.413	0.187	4.0×10^{-14}	-1.413	0.188	5.4×10^{-14}
24	rs134963365	46,786,442	T/G	0.485	0.988	0.160	6.2×10^{-10}	0.988	0.160	7.1×10^{-10}
24	rs379672347	62,207,595	G/T	0.248	-1.010	0.181	2.6×10^{-8}	-1.010	0.182	2.9×10^{-8}

25	rs459173018	2,987,055	T/C	0.045	2.311	0.365	2.3×10^{-10}	2.311	0.366	2.7×10^{-10}
26	rs42083046	7,416,303	C/T	0.199	-1.268	0.195	8.6×10^{-11}	-1.268	0.196	1.0×10^{-10}
28	rs137496077	36,027,888	G/A	0.488	-0.917	0.162	1.4×10^{-8}	-0.917	0.162	1.5×10^{-8}

Chr: Chromosome; Position: base pair positions in UMD3.1; A1/A2: Effect allele/other allele; Freq: frequency of A1 allele; *b*: allele substitution effect, se: standard error of b estimates; P: p-value from single marker GWAS. Subscript 'J' indicates values from conditional and joint analysis.

Table 5.7. Annotation of fertility index associated loci in Holstein cattle

Top Marker	SO	Gene	Annotation
rs137209107	1	MYLK ENSBTAG00000014567	**MYLK**: Familial thoracic aortic aneurysm in humans [OMIM:613780]
rs384339926	1	MME ENSBTAG00000002075	**MME**: abnormal circulating protein level, dermatitis, amyloidosis [MGI:97004]
rs207792133	1	RAPGEF4 ENSBTAG00000020984	**RAPGEF4**: abnormal calcium ion homeostasis, abnormal sarcoplasmic reticulum morphology, ventricular tachycardia [MGI:1917723]
rs43306949	2	(-497 kb) SNRPD1 ENSBTAG00000012376	**SNRPD1**: viability of whole organism in zebrafish [ZFIN: ZDB-GENE-020419-14]
rs137588365	1	PHTF1 ENSBTAG00000019615	**PHTF1**: abnormal heart morphology, abnormal pancreas morphology, abnormal skin morphology [MP:0000266, 0001944, 0002060]
rs42253052	2	(-208 kb) ST6GALNAC5 ENSBTAG00000007309	**ST6GALNAC5**: rheumatoid arthritis in rats
rs109192619	1	MGC137454 ENSBTAG00000030852; PDE4B ENSBTAG00000008636; (+467 kb) LEPR ENSBTAG00000005910	**PDE4B**: abnormal spinal cord morphology, abnormal CNS synaptic transmission, in mouse [MGI:99557]; **LEPR**: reduced female fertility, delayed estrous cycle, abnormal female reproductive system morphology, decreased mature ovarian follicle number, premature death, male infertility, absent estrus, abnormal estrous cycle, anovulation, in mouse [MGI:104993]
rs135727662	2	(+125 kb) NAMPT ENSBTAG00000015509; (-57 kb) SYPL1 ENSBTAG00000019794	**NAMPT**: abnormal embryo size[MP:0001697], embryonic lethality prior to organogenesis [MGI:1929865], embryonic lethality prior to tooth bud stage [MP:0013293], in mouse; **SYPL1**: male infertility in mouse [MGI:108081; MP:0001925]
rs385277008	1	PPP1R17 ENSBTAG00000001976	**PPP1R17**: familial hypercholesterolemia [OMIM:143890]
rs210968504	1	SOX5 ENSBTAG00000022360	**SOX5**: failure of palatal shelf elevation, neonatal lethality, respiratory failure, in mouse [MGI:98367]
rs110866176	2	(+22 kb) SOWAHB ENSBTAG00000038148; (-125 kb) SHROOM3 ENSBTAG00000019633	**SOWAHB**: exencephaly, wavy neural tube, abnormal nervous system morphology, in mouse [MGI:1925338]; **SHROOM3**: abnormal embryo development, absent neurocranium, wavy neural tube, perinatal lethality, exencephaly, in mouse [MGI:1351655]
rs473589185	2	(+154 kb) ARRDC3 ENSBTAG00000007116	**ARRDC3**: embryonic lethality, perinatal lethality, abnormal energy homeostasis, in mouse [MGI:2145242]
rs382943070	2	(-97 kb) CNTLN ENSBTAG00000001847	–
rs208577191	1	TMC1 ENSBTAG00000018778; (-367 kb) ZFAND5 ENSBTAG00000009417	**ZFAND5**: neonatal lethality in mouse [MGI:1278334]
rs380495923	2	(-287 kb) TBPL2 ENSBTAG00000018277	**TBPL2**: decreased rate of embryo development in zebrafish [ZFIN:ZDB-GENE-040520-3]; impaired ovarian folliculogenesis, oocyte degeneration, female infertility, absent zona pellucida, in mouse [MGI:2684058]
rs134897493	2	(-38 kb) ACTN1 ENSBTAG00000018255; (-188 kb) ZFP36L1 ENSBTAG00000025434	**ACTN1**: Bleeding disorder, platelet-type, 15 [OMIM:615193]; **ZFP36L1**: embryonic growth retardation, embryonic lethality during organogenesis,

			abnormal placenta morphology, abnormal embryonic erythropoiesis, decreased embryo size, open neural tube, in mouse [MGI:107946]
rs134324803	2	Gene desert	–
rs209793386	2	(+187 kb) BCL11A ENSBTAG00000016534	**BCL11A**: neonatal lethality in mouse [MGI:106190]
rs384444008	2	(+164 kb) KIAA1217 ENSBTAG00000018395	**KIAA1217**: abnormal caudal vertebrae morphology, abnormal intervertebral disk development, in mouse [MGI:95454]
rs385034802	2	(-10 kb) FAM107B ENSBTAG00000010023	**FAM107B**: abnormal coat/ hair morphology, decreased body weight, abnormal mouth morphology, in mouse [MGI:1913790]
rs434006863	1	MPP7 ENSBTAG00000017354	**MPP7**: decreased mass density in bone tissue [ZFIN: ZDB-GENE-991209-8]
rs136213414	1	HAO1 ENSBTAG00000019811	**HAO1**: Prostatic Neoplasms; Nephrolithiasis, Calcium Oxalate in rats
rs42685926	1	C14H8orf34 ENSBTAG00000022588	–
rs109077219	1	PGR ENSBTAG00000024648	**PGR:** Progesterone resistance in humans [OMIM:264080]; absent corpus luteum, anovulation, female infertility, in mouse [MGI:97567]
rs41853464	2	(-49 kb) MORC2 ENSBTAG00000009963; (+31 kb) SMTN ENSBTAG00000003100	**MORC2**: preweaning lethality, male infertility, female infertility, in mouse [MGI:3045293; MP:0011100]; **SMTN**: postnatal lethality, abnormal enterocyte morphology, postnatal growth retardation, in mouse [MGI:1354727]
rs382112789	2	Gene desert	–
rs450967176	2	(-171 kb) KCTD15 ENSBTAG00000017956; (+225 kb) LSM14A ENSBTAG00000000630	**KCTD15**: abnormal embryo size [MP:0001697], preweaning lethality [MGI:2385276], in mouse; **LSM14A**: preweaning lethality in mouse [MGI:1914320]
rs382522431	2	>10 genes with ±500 kb	–
rs383235276	2	–	–
rs41911242	2	(+20 kb) NLRP1 ENSBTAG00000020433	**NLRP1**: premature death in mouse [MGI:2684861]
rs715807267	2	(-190 kb) SOX4 ENSBTAG00000046556	**SOX4**: prenatal lethality in mouse [MGI:98366]
rs380439408	2	(-316 kb) CDH2 ENSBTAG00000021190	**CDH2**: abnormal embryo turning, wavy neural tube, embryonic growth retardation, embryonic lethality during organogenesis, prenatal lethality, in mouse [MGI:88355]
rs134963365	1	LOXHD1 ENSBTAG00000015210	**LOXHD1**: Deafness in humans [OMIM:613079], mouse [MGI:1914609] and rat
rs379672347	2	(-50 kb) KDSR ENSBTAG00000007723; (+41 kb) SERPINB5 ENSBTAG00000002019	**KDSR**: Spinal muscular atrophy in Bos taurus [OMIA 000939-9913]; **SERPINB5**: embryonic lethality between implantation and somite formation, embryonic epiblast cell degeneration, in mouse [MGI:109579]
rs459173018	2	(-10 kb) SLX4 ENSBTAG00000004509; (+182 kb) CREBBP ENSBTAG00000026403	**SLX4**: delayed embryonic viscerocranium morphogenesis in zebrafish [ZFIN:ZDB-GENE-050208-359]; lethality throughout fetal growth and development, abnormal DNA repair, perinatal lethality, postnatal growth retardation, preweaning lethality, reduced female fertility, in mouse [MGI:106299]; **CREBBP**: embryonic lethality during organogenesis, prenatal lethality, postnatal lethality, decreased embryo size, embryonic growth retardation, exencephaly, in mouse [MGI:1098280]
rs42083046	1	PRKG1 ENSBTAG00000018404	**PRKG1**: postnatal lethality [MGI:108174], preweaning lethality [MP:0011110], in mouse

rs137496077	[2]	(+49 kb) TSPAN14 ENSBTAG00000003907	**TSPAN14**: preweaning lethality in mouse [MP:0011100]

[1]intron variant; [2]intergenic variant. Annotations are based on UMD3.1 and Ensembl release 94. Information in the parenthesis represents the distance between the top marker and the start position of the corresponding gene. MGI: Mouse Genome Informatics (http://www.informatics.jax.org/); RGD: Rat Genome Database (https://rgd.mcw.edu/wg/); MP: Mouse Phenotype (https://www.mousephenotype.org/); OMIM: Online Mendelian Inheritance in Man (https://www.omim.org/); OMIA: Online Mendelian Inheritance in Animals; ZFIN: The Zebrafish Information Network (http://zfin.org/).

Table 5.8. Fertility index associated loci in Nordic Red Dairy cattle identified from GWAS with imputed sequence variants

Chr.	Top Marker	Position	A1/A2	Freq.	Single-marker analysis			Conditional and joint analysis		
					b	se	p	bJ	bJ_se	pJ
1	rs135868307	131,241,313	G/T	0.413	-1.198	0.202	2.9×10^{-9}	-1.198	0.203	3.3×10^{-9}
2	rs450253115	93,470,575	G/C	0.047	-2.559	0.450	1.3×10^{-8}	-2.787	0.453	7.6×10^{-10}
2	rs42490886	97,240,584	T/C	0.138	-1.498	0.279	8.1×10^{-8}	-1.646	0.281	4.7×10^{-9}
4	rs210258782	36,974,279	T/C	0.194	-1.618	0.238	1.0×10^{-11}	-1.618	0.239	1.3×10^{-11}
6	rs110759619	39,633,323	G/A	0.481	1.373	0.189	3.3×10^{-13}	1.373	0.190	4.4×10^{-13}
6	rs379594438	95,384,366	G/A	0.055	2.381	0.421	1.6×10^{-8}	2.381	0.423	1.8×10^{-8}
8	rs470078356	87,251,892	T/C	0.013	-4.800	0.861	2.4×10^{-8}	-4.800	0.864	2.7×10^{-8}
11	rs132823348	103,372,929	A/G	0.190	-1.343	0.242	2.7×10^{-8}	-1.343	0.243	3.0×10^{-8}
12	rs386105326	11,016,415	G/A	0.113	-2.336	0.308	3.2×10^{-14}	-2.336	0.310	4.6×10^{-14}
12	rs385475503 (rs797826108)	21,149,563	G/T	0.141	-4.087	0.290	4.9×10^{-45}	-11.123	0.445	8.0×10^{-138}
12	rs384604017	29,738,986	T/C	0.285	-0.482	0.221	2.9×10^{-2}	9.110	0.464	7.0×10^{-86}
12	rs133494863	29,790,646	C/T	0.274	0.734	0.224	1.0×10^{-3}	-5.561	0.359	3.6×10^{-54}
12	rs432066150	30,453,718	G/A	0.291	-0.962	0.226	2.1×10^{-5}	3.602	0.316	4.7×10^{-30}
12	Chr12:31178612	31,178,612	C/T	0.128	-2.820	0.298	3.1×10^{-21}	-14.778	0.636	2.5×10^{-119}
12	Chr12:43320610	43,320,610	A/G	0.087	-2.424	0.338	7.1×10^{-13}	-2.424	0.340	9.4×10^{-13}
12	rs377982349	55,272,444	C/T	0.086	-1.944	0.348	2.3×10^{-8}	-1.944	0.349	2.6×10^{-8}
12	rs133220791	69,485,528	T/C	0.234	-1.374	0.230	2.2×10^{-9}	-1.374	0.231	2.5×10^{-9}
13	rs207566536	46,143,358	G/A	0.136	-1.794	0.277	1.0×10^{-10}	-1.794	0.279	1.2×10^{-10}
13	rs378749244	56,756,424	T/C	0.475	1.246	0.206	1.4×10^{-9}	1.246	0.207	1.6×10^{-9}
17	rs136268337	59,753,524	T/C	0.353	1.200	0.210	1.1×10^{-8}	1.200	0.211	1.2×10^{-8}
19	rs135743550	39,835,929	A/G	0.302	-1.275	0.216	3.5×10^{-9}	-1.275	0.217	4.1×10^{-9}
25	rs382027875	13,426,672	T/C	0.055	-3.091	0.420	1.8×10^{-13}	-3.091	0.422	2.4×10^{-13}
26	rs481408336	35,317,193	A/G	0.061	2.147	0.363	3.3×10^{-9}	2.147	0.364	3.7×10^{-9}

Chr: Chromosome; Position: base pair positions in UMD3.1; A1/A2: Effect allele/other allele; Freq: frequency of A1 allele; *b*: allele substitution effect, se: standard error of b estimates; P: p-value from single marker GWAS. Subscript 'J' indicates values from conditional and joint analysis.

Table 5.9. Linkage disequilibrium (r) pattern among the top markers of chromosome 12 from Table 5.8

SNP	BP	rs386105326	rs385475503	rs384604017	rs133494863	rs432066150	Chr12:31178612	Chr12:43320610	rs377982349	rs133220791
rs386105326	11,016,415	1	0	0	0	0	0	0	0	0
rs385475503	21,149,563	0	1	0.335	-0.196	0.358	0	0	0	0
rs384604017	29,738,986	0	0.335	1	0.329	0.393	0.600	0	0	0
rs133494863	29,790,646	0	-0.196	0.329	1	-0.259	-0.236	0	0	0
rs432066150	30,453,718	0	0.358	0.393	-0.259	1	0.601	0	0	0
Chr12:31178612	31,178,612	0	0	0.600	-0.236	0.601	1	0	0	0
Chr12:43320610	43,320,610	0	0	0	0	0	0	1	0	0
rs377982349	55,272,444	0	0	0	0	0	0	0	1	0
rs133220791	69,485,528	0	0	0	0	0	0	0	0	1

Table 5.10. Annotation of fertility index associated loci in Nordic Red Dairy cattle

Top Marker	SO	Gene	Annotation
rs135868307	1	FOXL2 ENSBTAG00000031277	**FOXL2**: impaired ovarian folliculogenesis, postnatal lethality, female infertility, in mouse [MGI:1349428]
rs450253115	2	–	–
rs42490886	2	(-262 kb) PIKFYVE ENSBTAG00000002177	**PIKFYVE**: embryonic lethality before implantation, abnormal embryonic/fetal subventricular zone morphology, embryonic lethality between somite formation and embryo turning, in mouse [MGI:1335106]
rs210258782	2	(-346 kb) SEMA3A ENSBTAG00000018133; (+321 kb) SEMA3E ENSBTAG00000014920	**SEMA3A**: decreased survivor rate, lethality at weaning, premature death, prenatal lethality, abnormal posture, in mouse [MGI:107558]; **SEMA3E**: abnormal somite development, in mouse [MGI:1340034]
rs110759619	2	–	–
rs379594438	1	BMP2K ENSBTAG00000001126	**BMP2K**: abnormal lens morphology in mouse [MGI:2155456]
rs470078356	3	SPTLC1 ENSBTAG00000002220	**SPTLC1**: prenatal lethality, in mouse [MGI:1099431]
rs132823348	2	–	more than 20 genes within ±500 kb
rs386105326	2	(+156 kb) SUGT1 ENSBTAG00000002137	**SUGT1**: Abortion due to haplotype FH4 in Fleckvieh cattle [OMIA:001960-9913]; embryonic lethality prior to tooth bud stage [MP:0013293], embryonic lethality prior to organogenesis [MGI:1915205], in mouse
rs385475503	3	WDFY2 ENSBTAG00000008053; (-176 kb) INTS6 ENSBTAG00000002970; +314 kb) ATP7B ENSBTAG00000010353; (+338 kb) ALG11 ENSBTAG00000010337	**INTS6**: delayed embryo development in zebrafish [ZFIND: ZDB-GENE-070906-1], embryonic lethality between implantation and somite formation in mouse [MGI:1202397]; **ATP7B**: postnatal lethality, reduced female fertility, postnatal growth retardation, in mouse [MGI:103297]; **ALG11**: embryonic lethality prior to organogenesis [MP:0013292], preweaning lethality [MP:0011100], in mouse
rs384604017	2	(-13 kb) B3GLCT ENSBTAG00000033412; (-458 kb) RXFP2 ENSBTAG00000015132; (+421 kb) ALOX5AP ENSBTAG00000013201	**B3GLCT**: Peters plus syndrome in humans [OMIM:261540]; **RXFP2**: arrest of spermatogenesis, male infertility, decreased male germ cell number, in mouse [MGI:2153463]; **ALOX5AP**: edematous pericardium and yolk, increased occurrence of cell death, in zebrafish [ZFIN: ZDB-GENE-030131-9322]
rs133494863	2	same as rs384604017	–
rs432066150	2	(-101 kb) HMGB1 ENSBTAG00000018103; (+66 kb) KATNAL1 ENSBTAG00000009340	**HMGB1**: premature death, decreased glycogen catabolism rate, postnatal growth retardation, neonatal lethality, in mouse [MGI:96113]; **KATNAL1**: male infertility, abnormal spermatogenesis, abnormal spermatid morphology, in mouse [MGI:2387638]

Chr12:31178612	[3]	MTUS2 ENSBTAG00000001094; (-174 kb) SLC7A1 ENSBTAG00000018577	**SLC7A1**: abnormal cell physiology, anemia, decreased body size, neonatal lethality, in mouse [MGI:88117]
Chr12:43320610	[2]	–	–
rs377982349	[2]	(+412 kb) SPRY2 ENSBTAG00000001774	**SPRY2**: premature death, postnatal lethality, in mouse [MGI:1345138]
rs133220791	[2]	–	–
rs207566536	[2]	–	–
rs378749244	[2]	–	–
rs136268337	[3]	KSR2 ENSBTAG00000044119	**KSR2**: obese, increased percent body fat/body weight, in mouse [MGI:3610315]
rs135743550	[3]	SRCIN1 ENSBTAG00000013469	**SRCIN1**: impaired synaptic plasticity [MGI:1933179]
rs382027875	[2]	(-354 kb) MRTFB ENSBTAG00000008728	**MRTFB**: lethality throughout fetal growth and development, perinatal lethality, in mouse [MGI:3050795]
rs481408336	[3]	ABLIM1 ENSBTAG00000004899	**ABLIM1**: abnormal bone mineralization [MP:0002896], abnormal bone structure [MP:0003795], eye hemorrhage [MP:0006203], decreased circulating potassium level [MP:0005628], in mouse

[1]upstream gene variant; [2]intergenic variant; [3]intron variant. Annotations are based on UMD3.1 and Ensembl release 94. Information in the parenthesis represents the distance between the top marker and the start position of the corresponding gene; MGI: Mouse Genome Informatics (http://www.informatics.jax.org/); MP: Mouse Phenotype (https://www.mousephenotype.org/); ZFIND: The Zebrafish Information Network (http://zfin.org/); OMIA: Online Mendelian Inheritance in Animals; OMIM: Online Mendelian Inheritance in Man (https://www.omim.org/).

5.6 References

Abo-Ismail, M. K., L. F. Brito, S. P. Miller, M. Sargolzaei, D. A. Grossi, S. S. Moore, G. Plastow, P. Stothard, S. Nayeri, and F. S. Schenkel. 2017. Genome-wide association studies and genomic prediction of breeding values for calving performance and body conformation traits in Holstein cattle. Genet Sel Evol 49:82. https://doi.org/10.1186/s12711-017-0356-8.

Aulchenko, Y. S., S. Ripke, A. Isaacs, and C. M. van Duijn. 2007. GenABEL: an R library for genome-wide association analysis. Bioinformatics 23:1294-1296. https://doi.org/10.1093/bioinformatics/btm108.

Bouwman, A. C., H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, H. Pausch, R. F. Brøndum, P. J. Bowman, B. Thomsen, B. Guldbrandtsen, M. S. Lund, B. Servin, D. J. Garrick, J. Reecy, J. Vilkki, A. Bagnato, M. Wang, J. L. Hoff, R. D. Schnabel, J. F. Taylor, A. A. E. Vinkhuyzen, F. Panitz, C. Bendixen, L. E. Holm, B. Gredler, C. Hozé, M. Boussaha, M. P. Sanchez, D. Rocha, A. Capitan, T. Tribout, A. Barbat, P. Croiseau, C. Drögemüller, V. Jagannathan, C. Vander Jagt, J. J. Crowley, A. Bieber, D. C. Purfield, D. P. Berry, R. Emmerling, K. U. Götz, M. Frischknecht, I. Russ, J. Sölkner, C. P. Van Tassell, R. Fries, P. Stothard, R. F. Veerkamp, D. Boichard, M. E. Goddard, and B. J. Hayes. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet 50:362-367. https://doi.org/10.1038/s41588-018-0056-5.

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci 98:4107-4116. https://doi.org/10.3168/jds.2014-9005.

Browning, B. L. and S. R. Browning. 2016. Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 98:116-126. https://doi.org/10.1016/j.ajhg.2015.11.020.

Cai, Z., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2018. Prioritizing candidate genes post-GWAS using multiple sources of data for mastitis resistance in dairy cattle. BMC Genomics 19:656. https://doi.org/10.1186/s12864-018-5050-x.

Das, S., L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P. R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, and C. Fuchsberger. 2016. Next-generation genotype imputation service and methods. Nat Genet 48:1284-1287. https://doi.org/10.1038/ng.3656.

de Roos, A. P., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179:1503-1512. https://doi.org/10.1534/genetics.107.084301.

Delaneau, O., B. Howie, A. J. Cox, J. F. Zagury, and J. Marchini. 2013. Haplotype estimation using sequencing reads. Am J Hum Genet 93:687-696. https://doi.org/10.1016/j.ajhg.2013.09.002.

Delaneau, O., J. Marchini, C. Genomes Project, and C. Genomes Project. 2014. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. Nat Commun 5:3934. https://doi.org/10.1038/ncomms4934.

Elsik, C. G., D. R. Unni, C. M. Diesh, A. Tayal, M. L. Emery, H. N. Nguyen, and D. E. Hagen. 2016. Bovine Genome Database: new tools for gleaning function from the Bos taurus genome. Nucleic Acids Res 44:D834-839. https://doi.org/10.1093/nar/gkv1077.

Gebreyesus, G., A. J. Buitenhuis, N. A. Poulsen, M. Visker, Q. Zhang, H. J. F. van Valenberg, D. Sun, and H. Bovenhuis. 2019. Multi-population GWAS and enrichment analyses reveal novel genomic regions and promising candidate genes underlying bovine milk fatty acid composition. BMC Genomics 20:178. https://doi.org/10.1186/s12864-019-5573-9.

Hoglund, J. K., B. Buitenhuis, B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015a. Genome-wide association study for female fertility in Nordic Red cattle. BMC Genet 16:110. https://doi.org/10.1186/s12863-015-0269-x.

Hoglund, J. K., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015b. Identification of genomic regions associated with female fertility in Danish Jersey using whole genome sequence data. BMC Genet 16:60. https://doi.org/10.1186/s12863-015-0210-3.

Hoglund, J. K., G. Sahana, B. Guldbrandtsen, and M. S. Lund. 2014. Validation of associations for female fertility traits in Nordic Holstein, Nordic Red and Jersey dairy cattle. BMC Genet 15:8. https://doi.org/10.1186/1471-2156-15-8.

Iso-Touru, T., G. Sahana, B. Guldbrandtsen, M. S. Lund, and J. Vilkki. 2016. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genet 17:55. https://doi.org/10.1186/s12863-016-0363-8.

Jardim, J. G., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2018. Association analysis for udder index and milking speed with imputed whole-genome sequence variants in Nordic Holstein cattle. J Dairy Sci 101:2199-2212. https://doi.org/10.3168/jds.2017-12982.

Jiang, L., Z. Zheng, T. Qi, K. E. Kemper, N. R. Wray, P. M. Visscher, and J. Yang. 2019. A resource-efficient tool for mixed model association analysis of large-scale data. bioRxiv:598110. https://doi.org/10.1101/598110.

Kadri, N. K., B. Guldbrandtsen, M. S. Lund, and G. Sahana. 2015. Genetic dissection of milk yield traits and mastitis resistance quantitative trait loci on chromosome 20 in dairy cattle. J Dairy Sci 98:9015-9025. https://doi.org/10.3168/jds.2015-9599.

Kadri, N. K., G. Sahana, C. Charlier, T. Iso-Touru, B. Guldbrandtsen, L. Karim, U. S. Nielsen, F. Panitz, G. P. Aamand, N. Schulman, M. Georges, J. Vilkki, M. S. Lund, and T. Druet. 2014. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. PLoS Genet 10:e1004049. https://doi.org/10.1371/journal.pgen.1004049.

Marete, A., G. Sahana, S. Fritz, R. Lefebvre, A. Barbat, M. S. Lund, B. Guldbrandtsen, and D. Boichard. 2018. Genome-wide association study for milking speed in French Holstein cows. J Dairy Sci 101:6205-6219. https://doi.org/10.3168/jds.2017-14067.

NAV. 2013. NAV routine genetic evaluation of dairy cattle–data and genetic models. date accessed (July 8, 2019). url:http://www.nordicebv.info/wp-content/uploads/2015/04/General-description_from-old-homepage_06052015.pdf.

Sanchez, M. P., Y. Ramayo-Caldas, V. Wolf, C. Laithier, M. El Jabri, A. Michenet, M. Boussaha, S. Taussat, S. Fritz, A. Delacroix-Buchet, M. Brochard, and D. Boichard. 2019. Sequence-based GWAS, network and pathway analyses reveal genes co-associated with milk cheese-making properties and milk composition in Montbeliarde cows. Genet Sel Evol 51:34. https://doi.org/10.1186/s12711-019-0473-7.

van den Berg, I., D. Boichard, B. Guldbrandtsen, and M. S. Lund. 2016. Using Sequence Variants in Linkage Disequilibrium with Causative Mutations to Improve Across-Breed Prediction in Dairy Cattle: A Simulation Study. G3 (Bethesda) 6:2553-2561. https://doi.org/10.1534/g3.116.027730.

Yang, J., T. Ferreira, A. P. Morris, S. E. Medland, A. T. C. Genetic Investigation of, D. I. G. Replication, C. Meta-analysis, P. A. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, R. J. Loos, T. M. Frayling, M. I. McCarthy, J. N. Hirschhorn, M. E. Goddard, and P. M. Visscher. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44:369-375, S361-363. https://doi.org/10.1038/ng.2213.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011a. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76-82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Yang, J., M. N. Weedon, S. Purcell, G. Lettre, K. Estrada, C. J. Willer, A. V. Smith, E. Ingelsson, J. R. O'Connell, M. Mangino, R. Magi, P. A. Madden, A. C. Heath, D. R. Nyholt, N. G. Martin, G. W. Montgomery, T. M. Frayling, J. N. Hirschhorn, M. I. McCarthy, M. E. Goddard, P. M. Visscher, and G. Consortium. 2011b. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet 19:807-812. https://doi.org/10.1038/ejhg.2011.39.

Zerbino, D. R., P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C. G. Giron, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O. G. Izuogu, S. H. Janacek, T. Juettemann, J. K. To, M. R. Laird, I. Lavidas, Z. Liu, J. E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D. N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C. K. Ong, A. Parker, M. Patricio, H. S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow, K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S. E. Hunt, M. Kostadima, N. Langridge, F. J. Martin, M. Muffato, E. Perry, M. Ruffier, D. M. Staines, S. J. Trevanion, B. L. Aken, F. Cunningham, A. Yates, and P. Flicek. 2018. Ensembl 2018. Nucleic Acids Res 46:D754-D761. https://doi.org/10.1093/nar/gkx1098.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marcais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol 10:R42. https://doi.org/10.1186/gb-2009-10-4-r42.

Chapter 6.

Genomic prediction for female fertility using imputed whole-genome sequence variants including large chromosomal deletions

Md Mesbah-Uddin,[1,2*] Aurélien Capitan[2,3], Bernt Guldbrandtsen,[1] Mogens Sandø Lund,[1] Goutam Sahana,[1] and Didier Boichard,[2]

[1]Department of Molecular Biology and Genetics, Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark

[2]GABI, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[3]Allice, 75595 Paris, France

[*]Corresponding author: mdmesbah@gmail.com

Manuscript in preparation (for a Short Communication)

6.1 Abstract

We performed within-breed genomic prediction analyses using genomic best linear unbiased prediction (GBLUP) method. Besides Bovine50k BeadChip (50k) markers, in GBLUP, we used one or several (genetic) variance components to investigate the effect of imputed whole-genome sequence (WGS) variants and large chromosomal deletions (deletions) on prediction accuracy of female fertility in Nordic Holstein (HOL), Danish Jersey (JER) and Nordic Red Dairy cattle (RDC) populations of artificial insemination bulls. In HOL, we observed same prediction accuracy (0.639) using either 50k (one-component model) or 50k and deletion markers (two-component model). In JER, we achieved a prediction accuracy of 0.664 with fertility index (FI) associated markers alone, which was ~10% higher than the prediction accuracy with 50k markers only (0.595). In RDC, we observed gains in prediction accuracy when imputed WGS variants as well as deletions were included in addition to 50k markers; we achieved a 2.1% gains in prediction accuracy in RDC with 50k and FI associated markers (in two-component model) vs. that of only 50k markers (one-component model). Interestingly, we also achieved higher gains in prediction accuracy for FI in all three breeds compared with the previous reports. Finally, our results showed that imputed WGS variants and deletions could improve prediction accuracy. However, direct genotyping is necessary to explore the full potential of these markers as predictor in routine genomic evaluation program.

Key words: genomic prediction, large deletion, fertility, GBLUP, dairy cattle

6.2 Introduction

Estimation of genetic merits of animals using genomic markers (known as 'genomic selection') has revolutionized livestock breeding (Georges et al., 2019). Since the introduction of genomic selection in 2008 in US dairy cattle, the rates of annual genetic gain has increased by 50-100% on the breeding objectives (Garcia-Ruiz et al., 2016). This increase is attributed to a strong reduction in generation interval, reduced from 7 y to 2.5 y on the sire-daughter and sire-son pathways. For fitness traits, such as productive life and fertility, this relative increase is much larger, 3 to 4-fold in Holstein cattle, despite low heritability (0.04-0.08) of these traits (Garcia-Ruiz et al., 2016). This performance is due to the higher accuracy of genomic estimated breeding values for these traits. Nevertheless, prediction accuracy for breeding values is markedly lower than unity and can still be increased by various means, such as, larger reference population, improvement of statistical models, use of trait-associated markers from whole genome sequence variants and putative causal variants, use of other class of markers, besides SNPs and indels, like structural variants etc. (as discussed in **Chapter 1** and **Chapter 7** in this thesis).

Several studies investigated the effect of using the Bovine50k BeadChip (50k) markers along with few known quantitative trait loci (QTL), identified from imputed whole-genome sequence (WGS) variants, on prediction accuracy for female fertility (Brøndum et al., 2015, Su et al., 2015, Ma et al., 2019); however, very little gain in accuracy was reported for this trait. Structural variants (SVs) such as large chromosomal deletions (>50 bp, referred to as deletions, hereafter), on the other hand, has the potential to be causal, as

shown previously (Charlier et al., 2012, Kadri et al., 2014, Sahana et al., 2016), and thus when included as predictors could improve prediction accuracy, especially for fitness traits. To our knowledge, only one study investigated the effect of deletions and duplications along with other markers for predicting several growth traits in Nellore cattle (Hay et al., 2018). Liu et al. (2018) also reported SVs along with other WGS variants in their dataset; however, the primary focus of that study was to investigate the effect of training population size on prediction reliability. The objective of this study was to investigate the effect of imputed WGS selected markers, such as markers with suggestive association, and deletions, with or without 50k markers, on accuracy of genomic prediction for female fertility in HOL, JER and RDC.

6.3 Methods

Phenotype

We considered deregressed proofs of fertility index (FI) as the trait of interest for investing the improvement in accuracy of genomic prediction when various classes of DNA markers were used. This study was conducted in three Nordic dairy breeds, namely HOL (5,577), JER (1,211), and RDC (4,471) bulls with FI in our dataset.

Marker selection

We used imputed WGS variants, namely SNPs, indels, and deletions from our joint imputation study presented in **Chapter 4** in this thesis. From this dataset, we selected 50k markers, deletions, and a set of markers associated with FI or its component traits that we identified in a within-breed GWAS (presented in **Chapter 5**). We only included markers with minor allele frequency (MAF) >0.01 and Minimac3 software (Das et al., 2016) imputation accuracy R^2 >0.1 in the training dataset. The GWAS markers were included when the *P-value* of association with FI or any of the five-component traits were $<5\times10^{-5}$ in JER, while $<10^{-7}$ in HOL/RDC. Five FI component traits were number of inseminations per conception in heifers (AISh) or in cows (AISc), interval (number of days) from calving to first insemination in cows (ICF), and interval (number of days) from first to last insemination in heifers (IFLh) or in cows (IFLc). Markers were selected within-breed; each selected marker was allocated to a class corresponding to a trait. When a single marker was associated with multiple traits within a breed, it was assigned to a trait according to the preference of FI >AISc >AISh >IFLc >IFLh >ICF. This trait preference was based on relative importance of a component trait on fertility index calculation as described in (NAV, 2013). A summary of the selected variants is presented in Table 6.1.

Prediction model

We performed within-breed prediction analysis using genomic best linear unbiased prediction (GBLUP) model, implemented in GCTA software (Yang et al., 2011), with a minimum of one-component variance model to a maximum of eight-component variance model. For each breed, we built eight genomic relationship matrices (GRM, or simply, **G**), each corresponding to a class of variants, $\mathbf{G}_{\mathbf{50k}}$, $\mathbf{G}_{\mathbf{Deletions}}$ and one class for each trait (Table 6.1). We used the following GBLUP model:

$$\boldsymbol{y} = 1\mu + \sum_{i=1}^{n} \boldsymbol{g}_i + \boldsymbol{\varepsilon}$$

Here, $\boldsymbol{y}$ is the vector of phenotypes (de-regressed breeding values), μ is the phenotypic mean, n is the number of variant classes included in the model ($n = 1, 2, \ldots, 8$); $\boldsymbol{g}_i$ is a vector of random genetic effects corresponding to i^{th} variant class with $\boldsymbol{g_i} \sim N(\boldsymbol{0}, \boldsymbol{G}_i\sigma_i^2)$, $\boldsymbol{G}_i$ was constructed following VanRaden (2008) method-1 using GCTA software, σ_i^2 was the additive genetic variance explained by the i^{th} variant class; and $\boldsymbol{\varepsilon}$ is a vector of residual effect with $\varepsilon \sim N(\boldsymbol{0}, \boldsymbol{I}\sigma_\varepsilon^2)$ where $\boldsymbol{I}$ is an identity matrix and σ_ε^2 is the residual variance. Differences in accuracy of deregressed proofs were neglected. Variance components were estimated in GTCA software by restricted maximum likelihood (REML) approach with a combination of expectation-maximization (EM-REML) and average-information (AI-REML) algorithms (Yang et al., 2011).

For each breed, the GBLUP models were trained on animals born before the year 2005 (4,275 HOL, 1,031 JER and 3,626 RDC), and tested on animals born in the year 2005 or later (1,302 HOL, 180 JER and 845 RDC).

6.4 Results and Discussion

In a within-breed genomic prediction using the GBLUP method, we investigated the effect of a group of selected markers, each with a separate variance component (Table 6.1), on the accuracy of prediction for female fertility index in populations of HOL, JER, and RDC bulls. The average reliability of FI in our dataset was (train vs. test) 0.729 vs. 0.645 in HOL, 0.635 vs. 0.587 in JER, and 0.811 vs. 0.654 in RDC. Additive genetic variance and the proportion of phenotypic variance explained by different markers set is presented in Table 6.2. We present the prediction accuracy in Figure 6.1 and prediction bias in Table 6.3. In Table 6.4, we report reliability of predicted breeding values (without scaling) for the ease of comparisons with previous results on the same population.

In HOL, the prediction accuracy was highest (Figure 6.1a) and prediction bias (Table 6.3) was lowest when 50k markers in a one-component or 50k and deletion markers in a two-component GBLUP model were used. However, when 50k and FI associated markers were used in a two-component GBLUP model prediction accuracy decreased by 0.023 compared with that of 50k, or 50k and deletion markers. Furthermore, using FI and its component traits associated markers only (without 50k) resulted in a ~31% relative reduction in prediction accuracy (Figure 6.1a) and ~61% increase in prediction bias where bulls' breeding values were overestimated (Table 6.3). Higher reliability of prediction with 50k markers was also reported previously for this breed (Brøndum et al., 2015). However, using 50k (or 50k with deletion) markers we had a gain of 36% ($\frac{0.408}{0.645}$ in Table 6.4 vs. 0.402 in Brøndum et al. (2015) Table A1) in reliability of prediction compared to that of Brøndum et al. (2015), with similar prediction bias (0.939 in this study vs. 0.943 in Brøndum et al. (2015) Table A4). Interestingly, contrary to Brøndum et al. (2015), prediction reliability of 50k with GWAS selected markers was relatively lower in our study compared with that of 50k markers only.

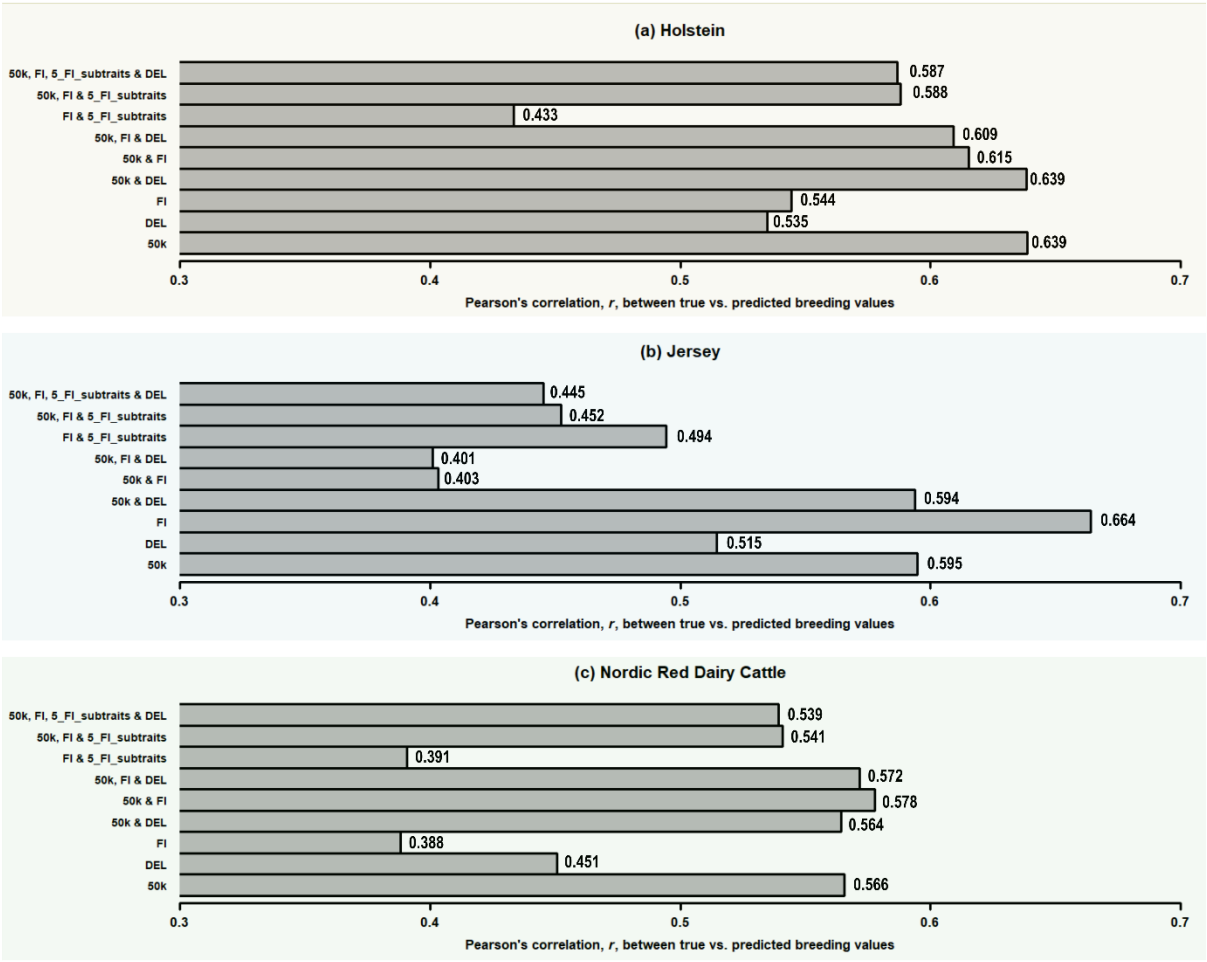


Figure 6.1. Accuracy of genomic prediction, r, of fertility index in Holstein, Jersey and Nordic Red Dairy Cattle. *Here, X-axis was started at 0.30. Within-breed genomic prediction was performed using genomic best linear unbiased prediction (GBLUP) method with single or multiple genomic relationship matrices. Markers associated with fertility index (or any of the 5-component traits) were selected as GWAS markers. Five FI component traits are number of inseminations per conception in heifers (AISh) and cows (AISc), interval (number of days) from calving to first insemination in cows (ICF), interval (number of days) from first to last insemination in heifers (IFLh) and cows (IFLc). 50k: Illumina BovineSNP50 BeadChip markers; FI: markers associated with fertility index; 5_FI_subtraits: markers associated with FI component traits; DEL: large chromosomal deletions (size>50 bp) that are not in the GWAS marker list.*

In JER, using only FI associated markers we observed a substantial gain in prediction accuracy (and reliability), with a regression slope (true vs. predicted) close to unity (Figure 6.1b and Table 6.3 and 6.4). Using only 50k markers or 50k and deletion markers reduced prediction accuracy by 0.069 and 0.070, respectively, compared with that of FI associated markers only (Figure 6.1b). Although predictions with both one-component model of 50k markers and two-component model of 50k and deletion markers were biased upwards (breeding value were underestimated), the latter model was relatively less biased (Table 6.3); although deletion component had nearly zero variance in this model (Table 6.2). Importantly, using 50k or FI associated markers only we had a relative gain of ~12% (0.354 in this study vs. 0.311 previously reported) and ~29% (0.441 in this study vs. 0.311 previously reported) in prediction reliability of FI compared with previous reports on the same population (Thomasen et al., 2012, Su et al., 2015, Liu et al., 2018).

In RDC, we achieved highest prediction accuracy with a two-component model of 50k and FI associated markers (r = 0.578) followed by a three-component model comprising 50k, FI and deletion markers (r = 0.572). Prediction accuracy with 50k markers was the third best from the nine scenarios we considered in this study (Figure 6.1c). Our results are consistent with the previous report (Brøndum et al., 2015). Interestingly, we observed a relative gain of 10% in prediction reliability compared with that of Brøndum et al. (2015), although we overestimated bulls' genetic merit (Table 6.3).

Overall, predictability of imputed deletions alone was rather impressive considering the relative size of this marker-set (Table 6.1). For example, in RDC, even with ~4 times more markers in FI dataset, prediction accuracy was ~14% lower compared with that of using deletion markers only. Furthermore, we achieved ~80% of 50k markers' accuracy using 10 times fewer markers in the deletion dataset (0.451 with deletion vs. 0.566 with 50k; ~4k deletion markers vs. ~41k in 50k dataset).

6.5 Conclusion

In this study, we illustrated the potential of large chromosomal deletions as well as imputed WGS markers in predicting genomic breeding values in dairy cattle. For female fertility, we reported a substantial gain in prediction accuracy/reliability compared with previous reports in the same population. We showed that in a two-component model, including deletions along with 50k markers did not improve prediction accuracy compared with that of one-component 50k model. However, we also showed that they did not reduce accuracy, and this could mean that these imputed deletions did not add noise to our prediction model and could be useful for predicting fitness related traits. However, direct genotyping of these imputed deletions and WGS markers is necessary to explore the full potential of these markers as predictor.

Acknowledgement

We are grateful to the Nordic Cattle Genetic Evaluation (NAV), Aarhus, Denmark for providing the phenotypic data used in this study and Viking Genetics, Randers, Denmark for proving blood/semen samples for genotyping. This research was supported by the Center for Genomic Selection in Animals and Plants (GenSAP) funded by Innovation Fund Denmark (grant 0603-00519B). Md Mesbah-Uddin benefited from a joint grant from the European Commission within the framework of the Erasmus-Mundus joint doctorate 'EGS-ABG'. Authors also acknowledge the 1000 Bull Genomes Project for sharing the VCF files. Authors are also grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing and storage resources.

Table 6.1. Summary of selected markers in the training dataset, used in constructing genomic-relationship matrices for genomic prediction

Type of variants[1]		HOL	JER	RDC	GRM
50k		40,917	36,400	41,394	G_{50k}
DEL		3,683	3,101	4,225	G_{DEL}
GWAS	**FI**	12,990	3,465	16,541	G_{FI}
	AISc	2,186	3,521	3,402	G_{AISc}
	AISh	673	3,679	1,901	G_{AISh}
	IFLc	929	1,507	1,408	G_{IFLc}
	IFLh	461	2,710	128	G_{IFLh}
	ICF	16,694	3,517	398	G_{ICF}

[1]Here, 50k: Illumina BovineSNP50 BeadChip markers. DEL: large chromosomal deletions (size >50 bp) that are not in the GWAS marker list. GWAS: markers selected from single-marker GWAS of fertility index (FI) and five component traits, e.g. number of inseminations per conception in heifers (AISh) or cows (AISc), interval (number of days) from calving to first insemination in cows (ICF), interval (number of days) from first to last insemination in heifers (IFLh) or cows (IFLc). HOL: Holstein; JER: Jersey; RDC: Nordic Red Dairy cattle; GRM/G: genomic relationship matrix.

Table 6.2. Variance explained by different classes of variants

Variance Components [1]	HOL		JER		RDC	
	V(G)	**V(G)/Vp**	**V(G)**	**V(G)/Vp**	**V(G)**	**V(G)/Vp**
G_{50k}	65.721	72.77%	82.491	67.38%	69.016	71.49%
G_{DEL}	50.615	55.14%	60.951	50.99%	43.708	47.34%
G_{FI}	313.260	84.69%	329.756	83.92%	129.174	65.68%
$G_{50k} + G_{DEL}$	63.096 + 2.752	69.90% + 3.05%	88.532 + 0.000	72.56% + 0.00%	66.336 + 3.061	68.64% + 3.17%
$G_{50k} + G_{FI}$	45.233 + 38.300	41.49% + 35.13%	38.747 + 249.261	12.14% + 78.10%	54.704 + 25.971	50.41% + 23.93%
$G_{50k} + G_{FI} + G_{DEL}$	43.749 + 38.057 + 1.613	40.21% + 34.98% + 1.48%	37.555 + 249.350 + 1.209	11.76% + 78.11% + 0.38%	54.345 + 25.957 + 0.413	50.08% + 23.92% + 0.38%

[1]Here, the order of the variance components are same as in corresponding values. **V(G)**: Genetic variance at marker set **G**; **V(G)/Vp**: proportion of phenotypic variance explained by marker-set **G**; **G** corresponds to G_{50k}, G_{FI} and G_{DEL}.

Table 6.3. Slope of the regression of phenotype on the genomic predicted breeding value for fertility index using the 50k, selected GWAS loci, large deletions (>50 bp) and combinations of these markers

Breed	Regression slope of true vs. predicted breeding values								
	50k	**DEL**	**FI**	**50k & DEL**	**50k & FI**	**B50k, FI & DEL**	**FI & 5-subtraits**	**B50k, FI & 5-subtraits**	**B50k, FI, 5-subtraits & DEL**
HOL	**0.94**	0.88	0.81	**0.94**	0.78	0.77	0.36	0.67	0.67
JER	1.18	1.22	**0.99**	1.12	0.18	0.18	0.42	0.33	0.32
RDC	**0.91**	0.90	0.75	**0.91**	0.88	0.87	0.49	0.72	0.73

Table 6.4. Reliability of genomic prediction (without scaling[1]) for fertility index

Breed	Reliability of prediction								
	50k	**DEL**	**FI**	**50k & DEL**	**50k & FI**	**B50k, FI & DEL**	**FI & 5-subtraits**	**B50k, FI & 5-subtraits**	**B50k, FI, 5-subtraits & DEL**
HOL	**0.408**	0.286	0.296	**0.408**	0.379	0.371	0.188	0.346	0.344
JER	0.354	0.265	**0.441**	0.353	0.163	0.161	0.244	0.205	0.198
RDC	0.320	0.203	0.151	0.318	**0.334**	0.327	0.153	0.293	0.291

[1]Reliability of prediction could be scaled using: $\frac{squared\ correlation\ between\ EBVs\ and\ GEBVs}{reliability\ of\ EBVs\ in\ testing\ dataset}$. Here, we reported squared prediction accuracy, i.e. prediction reliability, which could be scaled using average reliability of FI in the test population (HOL: 0.645, JER: 0.587, and RDC: 0.654).

6.6 References

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci 98:4107-4116. https://doi.org/10.3168/jds.2014-9005.

Charlier, C., J. S. Agerholm, W. Coppieters, P. Karlskov-Mortensen, W. Li, G. de Jong, C. Fasquelle, L. Karim, S. Cirera, N. Cambisano, N. Ahariz, E. Mullaart, M. Georges, and M. Fredholm. 2012. A deletion in the bovine FANCI gene compromises fertility by causing fetal death and brachyspina. PLoS One 7:e43085. https://doi.org/10.1371/journal.pone.0043085.

Das, S., L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P. R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, and C. Fuchsberger. 2016. Next-generation genotype imputation service and methods. Nat Genet 48:1284-1287. https://doi.org/10.1038/ng.3656.

Garcia-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-Lopez, and C. P. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci U S A 113:E3995-4004. https://doi.org/10.1073/pnas.1519061113.

Georges, M., C. Charlier, and B. Hayes. 2019. Harnessing genomic information for livestock improvement. Nat Rev Genet 20:135-156. https://doi.org/10.1038/s41576-018-0082-2.

Hay, E. H. A., Y. T. Utsunomiya, L. Xu, Y. Zhou, H. H. R. Neves, R. Carvalheiro, D. M. Bickhart, L. Ma, J. F. Garcia, and G. E. Liu. 2018. Genomic predictions combining SNP markers and copy number variations in Nellore cattle. BMC Genomics 19:441. https://doi.org/10.1186/s12864-018-4787-6.

Kadri, N. K., G. Sahana, C. Charlier, T. Iso-Touru, B. Guldbrandtsen, L. Karim, U. S. Nielsen, F. Panitz, G. P. Aamand, N. Schulman, M. Georges, J. Vilkki, M. S. Lund, and T. Druet. 2014. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. PLoS Genet 10:e1004049. https://doi.org/10.1371/journal.pgen.1004049.

Liu, A., M. S. Lund, D. Boichard, S. Fritz, E. Karaman, Y. Wang, and G. Su. 2018. Using additional single nucleotide polymorphisms selected from whole genome sequence data for genomic prediction in Danish Jersey. Proceedings of the World Congress on Genetics Applied to Livestock Production 11:586.

Ma, P., M. S. Lund, G. P. Aamand, and G. Su. 2019. Use of a Bayesian model including QTL markers increases prediction reliability when test animals are distant from the reference population. J Dairy Sci. https://doi.org/10.3168/jds.2018-15815.

NAV. 2013. NAV routine genetic evaluation of dairy cattle–data and genetic models. date accessed (July 8, 2019). url:http://www.nordicebv.info/wp-content/uploads/2015/04/General-description_from-old-homepage_06052015.pdf.

Sahana, G., T. Iso-Touru, X. Wu, U. S. Nielsen, D. J. de Koning, M. S. Lund, J. Vilkki, and B. Guldbrandtsen. 2016. A 0.5-Mbp deletion on bovine chromosome 23 is a strong candidate for stillbirth in Nordic Red cattle. Genet Sel Evol 48:35. https://doi.org/10.1186/s12711-016-0215-z.

Su, G., P. Ma, U. S. Nielsen, G. P. Aamand, G. Wiggans, B. Guldbrandtsen, and M. S. Lund. 2015. Sharing reference data and including cows in the reference population improve genomic predictions in Danish Jersey. Animal:1-9. https://doi.org/10.1017/S1751731115001792.

Thomasen, J. R., B. Guldbrandtsen, G. Su, R. F. Brøndum, and M. S. Lund. 2012. Reliabilities of genomic estimated breeding values in Danish Jersey. Animal 6:789-796. https://doi.org/10.1017/S1751731111002035.

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

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher. 2011. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88:76-82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Chapter 7.
General Discussion

7.1 Recessive Lethals

Poor fertility is generally related to embryonic mortality. Embryonic mortality can be caused by any lethal mutation homozygous in an embryo and affecting any essential gene, not necessarily always related to reproduction. Early embryonic mortalities are not observed and therefore, usual mapping methods cannot be used because biological material from dead foetus is not available. Nevertheless, when the number of genotypes is high enough in a population, such defects can be indirectly observed by a deficit in some genotypes. This approach is much more powerful with haplotypes that are more informative than single SNP and better reflect identical by descent (IBD) chromosomal regions. This method proposed by VanRaden et al. (2011) has been successfully applied in several breeds by different authors. In Normande breed, we have been able to map such a haplotype with a strong deficit in homozygous state and have identified the causal variant in *CAD* gene (**Chapter 2**). This study was rich in conclusions:

1) several previously described haplotypes in this breed were not confirmed, emphasizing the importance of a high quality imputation and phasing process and of a carefully adjusted statistical threshold accounting for the huge number of comparisons performed;
2) like other recessive lethals (Michot et al., 2017, Fritz et al., 2018), this mutation affected a very basic metabolic pathway (DNA base synthesis) and induced rapid mortality, when the compensatory metabolic stocks of maternal origin are depleted;
3) because of their strong effects, such mutations are always recent and therefore usually breed specific, but they can reach moderate frequencies (3-10%) because of the limited effective population size of our modern dairy breeds and strong genetic drift, as well as less efficient counter selection due to their recessive determinism;
4) because these defects are recessive, they are difficult to map with conventional GWAS, except when the frequency is already high and when the proportion of at risk mating becomes important. From this point of view, haplotype deficiency and GWAS are two complementary approaches to identify deleterious recessive mutations.
5) This haplotype deficiency study is a by-product of the large-scale genotyping for genomic selection. It requires very large numbers of genotypes and one can anticipate that the number of detected loci will increase over time, simply because detection power increases. It illustrates one out of many applications that can be derived from this large genotype data accumulating over time. Once discovered, such a mutation can be readily

added to SNP genotyping chip, disseminating this information to the whole population at nearly no additional cost.

6) Finally, from a practical point of view, because these defects are recessive and frequencies are usually below 10%, their economic impact is limited and does not justify a drastic elimination of the carriers. Instead, including this information in mating plans to avoid mating between carriers (or potentially carriers) is recommended. In addition, a careful monitoring of its frequency is important to avoid any further dissemination of the mutation and ensure a gradual elimination.

7.2 Use of Whole-Genome Sequence Variants

The markers used in majority of the genomic prediction studies were SNPs and (in fewer studies) indels. Furthermore, to our knowledge, no studies systematically explored the use of structural variants in the cattle genome for mapping or prediction purposes. This situation does not reflect the relative importance of SNP and SVs. SVs are less numerous than SNPs but the number of affected bases in the genome is of the same magnitude and each SV is much more likely to modify biological processes than any single SNP, especially when they affect one or even several genes. This situation of limited study is explained by the relative ease of access to the information for these two classes of variants. SNP data are readily available from SNP chips, without additional bioinformatics analyses. In contrast, with the noticeable but limited exception of SV breakpoints designed to be tested like SNPs by custom array (Boichard et al., 2018), most SVs are difficult to call.

Since our interest was to identify lethal mutations in dairy cattle, we choose to focus on large (>50 bp) chromosomal deletions besides SNPs and indels (**Chapter 3-6**). This is because deletion of an essential gene or its gene-regulatory region will certainly lead to loss-of-function (LOF) and can lead to lethality or a substantial reduction in fitness for animals homozygous for the deletion. Technological advances over the past decades for whole genome sequencing and availability of large WGS data in dairy cattle offer the possibility of testing different class of genomic variants for severe recessive genetic disorders. SVs can be called from sequence data with different algorithms (variation in read depth, discordant mapping of single reads or read pairs) but this remains tedious on a large scale and with a strong proportion of false positives. They can also be called from genotyping fluorescence signals and sometimes from Mendelian incompatibilities, but only for large SVs. However, SV genotype calling as well as imputation are not trivial and many challenges remain regarding the use of other type of

variants, besides SNPs and small indels, like SVs for large-scale genome-wide screening. Consequently, to our knowledge, these procedures are not yet included in any routine evaluation system because SVs lack flexibility and fully automatic pipelines.

In this PhD, we devoted two studies in mapping and imputing large deletions using WGS:

- In **Chapter 3**, we showed a systematic approach for high-resolution mapping of large deletions from WGS data of 175 animals, and subsequent genotyping and validation of the discoveries. Using deletion genotypes, we illustrated the population-genetic properties such as population diversity, stratification and differentiation. For essential gene deletion, we could only observe heterozygotes in our dataset. While for nonessential genes, homozygotes were observed among live animals. Using QTL enrichment analysis, we showed that deletion could be better predictor for health and fertility related traits; this was our hypothesis for performing subsequent imputation, GWAS and prediction studies. This study has made a catalogue of more than 8,000 large deletions segregating in three Nordic dairy cattle breeds. This catalogue of deletions will facilitate further research for mapping gene, improving accuracy in genomic selection and population genetics study.

- In **Chapter 4**, we described a Gaussian mixture model-based approach to estimate deletion-genotype likelihood using SNP read-depth data from VCF file. This was needed for enlarging the haplotype reference panel for imputing deletions into our bull population. There was a possibility of including ~600 additional animals (from HOL, JER and Norwegian RDC) from *Run-6* of the 1000 Bull Genome Project, only if we could have deletion genotypes for these animals. However, access to the raw sequence (BAM) file is limited in this case. Nevertheless, read-depth data supporting each SNP genotype call present in VCF file is an indirect representation of number of sequencing reads mapped to the SNPs neighbouring region. Furthermore, number of reads aligned to a position is proportional to the copy-number, and there is a linear relationship between number of reads and copy-number of a region. We used sequencing depth from VCF as proxy and applied a similar genotyping approach that is used for SNP and SVs with raw sequencing data. We validated our approached using the known carrier status for Brachyspina deletion among the sequenced bulls. This technique has several limitations, to name few, it can only genotype copy-number loss or gain, will perform purely in regions with shallow coverage, fully depends on the presence of SNPs.

Nevertheless, using this deletion genotyping approach, we were able to include additional WGS animals. Next, we built haplotype reference panel and subsequently imputed our array-typed population for SNPs, indels and deletions. Imputation accuracy for SNPs and indels were higher than in previous reports. And imputation accuracy for deletions were also high. Our results showed that common deletions could be imputed jointly with SNPs and indel with comparable accuracy. We used this imputed genotypes to perform GWAS and genomic prediction analyses.

7.3 Genomic Prediction

In dairy cattle, genomic selection provided substantial genetic gains, primarily by reducing the generation interval, for traits with low, moderate to high heritability (Garcia-Ruiz et al., 2016, Georges et al., 2019). Indeed, in contrast with phenotypic selection, high accuracy of genomic breeding values can be achieved with an appropriate reference population: many individuals with both phenotypic and genomic information can compensate for a low heritability. In addition, in its early stage in dairy cattle, genomic selection has been able to take advantage of a tremendous already existing resource, i.e. the large number of progeny tested bulls with very accurate breeding values even for low heritability traits. In this particular situation, the parameter of interest is not the heritability but the reliability of the evaluation, and the estimated breeding value based on progeny information is equivalent to an own performance of the bull with a heritability equal to the reliability.

However, the accuracy of genomic prediction has not reached to the label that was usually obtained through progeny testing, especially for populations with limited reference population size. Further increase in the accuracy of genomic prediction is necessary to achieve higher genetic gain. Several options have been studied in the literature in order to increase accuracy of genomic evaluation:

- **Increase size of the reference population**. This may be achieved (1) by genotyping more individuals with phenotype. However, this option is valid only if genotyping is cheap (or not directly supported by the breeding program) and if the population is large enough. Because the most informative individuals were genotyped first (i.e. the progeny tested bulls), the additional individuals are less informative and many more need to be genotyped to observe a significant gain. In addition, if the heritability of the trait is low such as for fertility, this option requires many individuals and may be unrealistic. Increasing size of the reference population may be also achieved by (2) building large consortia within breed such as EuroGenomics in Holstein

(Lund et al., 2011, Boichard et al., 2018) or Intergenomics in Brown Swiss (https://www.brown-swiss.org/genetics), or by exchanging information across countries in a bilateral way, such as between Denmark and USA in Jersey (Wiggans et al., 2015). This is an easy and cost-effective way to proceed, but it is possible only for international breeds and when $G \times E$ interactions are limited. (3) Many efforts have been dedicated to the use of a multi-breed reference population (Calus et al., 2018, van den Berg et al., 2019). Theoretically, this approach is very appealing because it reduces the fixed cost of the reference population in each breed. Its success relies on different assumptions, e.g. a strong common basis of the genetic determinism (i.e. many common QTL across breeds) and a strong conserved linkage disequilibrium phase across breeds. In spite of all these efforts, results are not fully convincing and significant gains in accuracy have been obtained only for closely related breeds and for traits with high heritability. Due to the major associated stakes, these efforts are continued based on sequence data and methods able to extract the appropriate information.

- **Feature Selection.** Ideally, inclusion of causal variants as predictor will provide maximum prediction accuracy, as seen in simulation study (Meuwissen and Goddard, 2010). Inclusion of markers in high LD with the causal variants will also improve prediction accuracy (van den Berg et al., 2016a, van den Berg et al., 2016b). One extreme example of this approach is the prediction of stature in cattle using only 163 highly predictive markers: multi-breed meta-analysis to map QTL and subsequent selection of markers (Bouwman et al., 2018). Here, homogeneity of genetic architecture for the trait across breeds is vital, i.e. if different sets of QTL affect the trait of interest in different breeds, this will not work as expected. This point remains an open question mark. Indeed, many QTL, especially those with moderate effects, are not found to be shared across populations. But most mapping designs have limited detection power for these small QTLs. Therefore, it is difficult to conclude whether the genetic determinism is different or if it is the same but we are not able to show that.

Nevertheless, the reported gains in prediction accuracy for fertility traits were negligible at best, although the gains were substantial for production traits (Brøndum et al., 2011, Brøndum et al., 2015). This could be a circular-problem: it is hard to map fertility associated QTL with the available mapping population, and thus inclusion of less informative markers rather adds noise to prediction models and often resulted in reduction in accuracy. Noteworthy, prediction

performance with the same mapping population is relatively higher for traits with high heritability.

However, in this PhD thesis, we have demonstrated that improving imputation accuracy of WGS variants (**Chapter 4**) and subsequently, selecting fewer predictive markers from GWAS (**Chapter 5**) can enhance accuracy in genomic prediction for female fertility (**Chapter 6**):

- In **Chapter 5**, we described single-marker association analyses for eight fertility traits using imputed SNPs, indels and deletions from **Chapter 4**. After the initial GWAS scan, we performed conditional and joint analyses. We detected several QTL regions, some were previously reported, and some new discoveries. Besides SNPs and indels, several deletions passed the GWAS significant threshold of $<5\times10^{-8}$. However, only BTA12 deletion, which was previously reported by Kadri et al. (2014), was among the top signals. Nevertheless, association *p-values* for majority of the significant loci in our results were stronger (small) compared with previous report, which could be due to the higher accuracy of our imputed dataset. Overall, our results highlights the importance of including SVs for association mapping in livestock species.

- In **Chapter 6**, we presented the results of genomic prediction analysis. Here, we investigated the impact of imputed WGS variants on accuracy of genomic prediction. We investigated several scenarios, such as, use of 50k data only, imputed deletion, fertility index associated markers, and a combination of these. We used GBLUP method with one or several genomic-relationship matrices derived from these marker sets. This method allows considering several classes of markers with different variance priors, while keeping a good simplicity of implementation. We observed substantial gain in prediction accuracy for fertility compared with previous reports. Our results demonstrated that deletions in our imputed dataset were also good predictors for fertility.

7.4 Evolutionary Conservation as a Tool for Identifying Lethal Genes

An interesting observation from three of the mapping studies (**Chapter 2, 3** and **5**) is that higher organisms share a common set of essential genes. These genes are indispensable for the organism's survival. On the other hand, higher organisms also share a variable set of genes that are dispensable due to redundancies in the genome. In fact, this is the basis of evolutionary

genetics (Meadows and Lindblad-Toh, 2017, Van de Peer et al., 2017). While looking for candidate mutations or candidate genes, this evolutionary conservation across species could be a powerful tool. Genes that are essential for survival and fitness are usually under strong purifying selection compared with nonessential genes (Hart et al., 2015, Wang et al., 2015), have fewer paralogs and involve in more protein-protein interactions (Blomen et al., 2015, Hart et al., 2015). These genes contain more singletons and very few deleterious variants at rare or low frequency (Blomen et al., 2015, Hart et al., 2015). Including *CAD* gene, these features were common for majority of the previously detected lethal genes in dairy cattle (OMIA, 2018).

7.5 Conclusions

- Causal factors for recessive lethals could be detected using a combination of HHD screen and subsequent validation
- Positional mapping and genotyping of large chromosomal deletions using WGS data provides opportunity for incorporating deletions in genomic studies
- Copy-number status, such as copy-number loss or gain, could be inferred using SNP read depth data from the VCF file
- Large deletions could be imputed with high accuracy along with SNPs and indels
- Better imputation could provide stronger GWAS signal
- Imputed WGS selected markers could improve within-breed prediction accuracy

7.6 Perspectives

- Detection of lethal haplotype, as well as lethal mutation will facilitate management of lethals in the population by avoiding at-risk mating. A first step is to include these markers on the SNP chip used for genomic selection and genotypes will be available to the whole genotyped population, making it possible to use it in selection and mating decisions.
- GWAS at the sequence level is supposed to include the causal variants. However, many association studies are limited to SNPs. Because SV are more likely to be causal than a SNP, it is important to include them in the studies, even if it increases genotype calling complexity.
- A balance between genetic gains and inbreeding should be maintained for sustainability in dairy cattle breeding. Rapid inbreeding trend favours both genetic drift and homozygotes proportions—two risk factors for fertility.

- Breeding objectives should include both production and fitness traits to maintain a steady progress for all traits
- Direct genotyping of imputed deletions and potential QTL is necessary to explore the full potential of these markers as predictor. It is possible to simplify SV genotyping process by SNP-like tests based on the SV breakpoints. Then large-scale information is readily available.
- Use of fewer predictive markers should be given preference for genomic prediction in small breeds, because there is less information to estimate each marker effect.
- Lastly, automated selection of predictive markers using deep learning should be explored

7.7 References

Blomen, V. A., P. Majek, L. T. Jae, J. W. Bigenzahn, J. Nieuwenhuis, J. Staring, R. Sacco, F. R. van Diemen, N. Olk, A. Stukalov, C. Marceau, H. Janssen, J. E. Carette, K. L. Bennett, J. Colinge, G. Superti-Furga, and T. R. Brummelkamp. 2015. Gene essentiality and synthetic lethality in haploid human cells. Science 350:1092-1096. https://doi.org/10.1126/science.aac7557.

Boichard, D., M. Boussaha, A. Capitan, D. Rocha, C. Hozé, M. P. Sanchez, T. Tribout, R. Letaief, P. Croiseau, C. Grohs, W. Li, C. Harland, C. Charlier, M. S. Lund, G. Sahana, M. Georges, S. Barbier, W. Coppieters, S. Fritz, and B. Guldbrandtsen. 2018. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. Page 675 in Proc. Proc. World Congr. Genet. Appl. Livest. Prod., Auckland, New Zealand. AL Rae Centre for Genetics and Breeding,

Massey University, Palmerston North, New Zealand. http://www.wcgalp.org/system/files/proceedings/2018/experience-large-scale-use-eurogenomics-custom-snp-chip-cattle.pdf.

Bouwman, A. C., H. D. Daetwyler, A. J. Chamberlain, C. H. Ponce, M. Sargolzaei, F. S. Schenkel, G. Sahana, A. Govignon-Gion, S. Boitard, M. Dolezal, H. Pausch, R. F. Brøndum, P. J. Bowman, B. Thomsen, B. Guldbrandtsen, M. S. Lund, B. Servin, D. J. Garrick, J. Reecy, J. Vilkki, A. Bagnato, M. Wang, J. L. Hoff, R. D. Schnabel, J. F. Taylor, A. A. E. Vinkhuyzen, F. Panitz, C. Bendixen, L. E. Holm, B. Gredler, C. Hozé, M. Boussaha, M. P. Sanchez, D. Rocha, A. Capitan, T. Tribout, A. Barbat, P. Croiseau, C. Drögemüller, V. Jagannathan, C. Vander Jagt, J. J. Crowley, A. Bieber, D. C. Purfield, D. P. Berry, R. Emmerling, K. U. Götz, M. Frischknecht, I. Russ, J. Sölkner, C. P. Van Tassell, R. Fries, P. Stothard, R. F. Veerkamp, D. Boichard, M. E. Goddard, and B. J. Hayes. 2018. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. Nat Genet 50:362-367. https://doi.org/10.1038/s41588-018-0056-5.

Brøndum, R. F., E. Rius-Vilarrasa, I. Stranden, G. Su, B. Guldbrandtsen, W. F. Fikse, and M. S. Lund. 2011. Reliabilities of genomic prediction using combined reference data of the Nordic Red dairy cattle populations. J Dairy Sci 94:4700-4707. https://doi.org/10.3168/jds.2010-3765.

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci 98:4107-4116. https://doi.org/10.3168/jds.2014-9005.

Calus, M. P. L., M. E. Goddard, Y. C. J. Wientjes, P. J. Bowman, and B. J. Hayes. 2018. Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. J Dairy Sci 101:4279-4294. https://doi.org/10.3168/jds.2017-13366.

Fritz, S., C. Hoze, E. Rebours, A. Barbat, M. Bizard, A. Chamberlain, C. Escouflaire, C. Vander Jagt, M. Boussaha, C. Grohs, A. Allais-Bonnet, M. Philippe, A. Vallee, Y. Amigues, B. J. Hayes, D. Boichard, and A. Capitan. 2018. An initiator codon mutation in SDE2 causes recessive embryonic lethality in Holstein cattle. J Dairy Sci. https://doi.org/10.3168/jds.2017-14119.

Garcia-Ruiz, A., J. B. Cole, P. M. VanRaden, G. R. Wiggans, F. J. Ruiz-Lopez, and C. P. Van Tassell. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci U S A 113:E3995-4004. https://doi.org/10.1073/pnas.1519061113.

Georges, M., C. Charlier, and B. Hayes. 2019. Harnessing genomic information for livestock improvement. Nat Rev Genet 20:135-156. https://doi.org/10.1038/s41576-018-0082-2.

Hart, T., M. Chandrashekhar, M. Aregger, Z. Steinhart, K. R. Brown, G. MacLeod, M. Mis, M. Zimmermann, A. Fradet-Turcotte, S. Sun, P. Mero, P. Dirks, S. Sidhu, F. P. Roth, O. S. Rissland, D. Durocher, S. Angers, and J. Moffat. 2015. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell 163:1515-1526. https://doi.org/10.1016/j.cell.2015.11.015.

Kadri, N. K., G. Sahana, C. Charlier, T. Iso-Touru, B. Guldbrandtsen, L. Karim, U. S. Nielsen, F. Panitz, G. P. Aamand, N. Schulman, M. Georges, J. Vilkki, M. S. Lund, and T. Druet. 2014. A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock. PLoS Genet 10:e1004049. https://doi.org/10.1371/journal.pgen.1004049.

Lund, M. S., A. P. Roos, A. G. Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol 43:43. https://doi.org/10.1186/1297-9686-43-43.

Meadows, J. R. S. and K. Lindblad-Toh. 2017. Dissecting evolution and disease using comparative vertebrate genomics. Nat Rev Genet. https://doi.org/10.1038/nrg.2017.51.

Meuwissen, T. and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623-631. https://doi.org/10.1534/genetics.110.116590.

Michot, P., S. Fritz, A. Barbat, M. Boussaha, M. C. Deloche, C. Grohs, C. Hoze, L. Le Berre, D. Le Bourhis, O. Desnoes, P. Salvetti, L. Schibler, D. Boichard, and A. Capitan. 2017. A missense mutation in PFAS (phosphoribosylformylglycinamidine

synthase) is likely causal for embryonic lethality associated with the MH1 haplotype in Montbeliarde dairy cattle. J Dairy Sci 100:8176-8187. https://doi.org/10.3168/jds.2017-12579.

OMIA. 2018. Online Mendelian Inheritance in Animals. date accessed (12/05/ 2018). url:http://omia.angis.org.au/.

Van de Peer, Y., E. Mizrachi, and K. Marchal. 2017. The evolutionary significance of polyploidy. Nat Rev Genet 18:411-424. https://doi.org/10.1038/nrg.2017.26.

van den Berg, I., D. Boichard, B. Guldbrandtsen, and M. S. Lund. 2016a. Using Sequence Variants in Linkage Disequilibrium with Causative Mutations to Improve Across-Breed Prediction in Dairy Cattle: A Simulation Study. G3 (Bethesda) 6:2553-2561. https://doi.org/10.1534/g3.116.027730.

van den Berg, I., D. Boichard, and M. S. Lund. 2016b. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genet Sel Evol 48:83. https://doi.org/10.1186/s12711-016-0259-0.

van den Berg, I., T. H. E. Meuwissen, I. M. MacLeod, and M. E. Goddard. 2019. Predicting the effect of reference population on the accuracy of within, across, and multibreed genomic prediction. J Dairy Sci. https://doi.org/10.3168/jds.2018-15231.

VanRaden, P. M., K. M. Olson, D. J. Null, and J. L. Hutchison. 2011. Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J Dairy Sci 94:6153-6161. https://doi.org/10.3168/jds.2011-4624.

Wang, T., K. Birsoy, N. W. Hughes, K. M. Krupczak, Y. Post, J. J. Wei, E. S. Lander, and D. M. Sabatini. 2015. Identification and characterization of essential genes in the human genome. Science 350:1096-1101. https://doi.org/10.1126/science.aac7041.

Wiggans, G. R., G. Su, T. A. Cooper, U. S. Nielsen, G. P. Aamand, B. Guldbrandtsen, M. S. Lund, and P. M. VanRaden. 2015. Short communication: Improving accuracy of Jersey genomic evaluations in the United States and Denmark by sharing reference population bulls. J Dairy Sci 98:3508-3513. https://doi.org/10.3168/jds.2014-8874.

Individual Training Plan

European Graduate School in Animal Breeding and Genetics

Mandatory Courses	Where/when	ECTS
Welcome to EGS-ABS	SLU, Sweden, 2015	2
EGS-ABG Research School	SLU, Sweden, 2015	2
EGS-ABG Research School	APT, France, 2018	2
Research Integrity in Scientific Professions	MOOC by University of Bordeaux, France, 2018	0
Advanced scientific courses		
Quantitative Genetics	AU, Denmark, 2015	5
Gene Mapping	AU, Denmark, 2016	5
Quantitative Genomics	AU, Denmark, 2016	5
Next-Generation Sequencing Analysis	AU, Denmark, 2016	5
Quantitative Genetics in Animal Breeding	NOVA PhD course, Finland, 2016	3
An Introduction to Computer Intensive Methods for Genetic Analysis	AU, Denmark, 2016	4
Transferable skills		
Science Teaching	AU, Denmark, 2016	5
Hands-on short course on NGS analysis – from FASTQ to genotypes	AU, Denmark, 2016	2
Techniques for Writing and Presenting a Scientific Paper	AU, Denmark, 2017	2
Programming Fundamentals	MOOC by Duke University on Coursera, 2019	0
Writing, Running, and Fixing Code in C	MOOC by Duke University on Coursera, 2019	0
Python for Genomic Data Science	MOOC by Johns Hopkins University on Coursera, 2019	0
Total ECTS		**42**

Dissemination of knowledge

Teaching	Contributed as teaching assistant (TØ) in the course “Genetics” [STADS UVA code 215151U013], Aarhus University, 2017

International Conferences

PhD Conference, MBG, AU, Denmark, 2016
GenSAP Annual Meeting, Denmark, 2015
The 36th International Society for Animal Genetics Conference, Ireland, 2017
World Congress on Genetics Applied to Livestock Production, New Zealand, 2018
Animal Genetics PhD Seminar, INRA/AgroParisTech, France, 2019

Seminars and workshop
GenSAP Annual Meeting, Denmark, 2016
Oral presentation, GenSAP Annual Meeting, Denmark, 2017
GenSAP Annual Meeting, Denmark, 2018
Oral presentation, Animal Genetics PhD Seminar, INRA/AgroParisTech, France, 2018

Title : Identification of causal factors for recessive lethals in dairy cattle with special focus on large chromosomal deletions

Keywords : recessive lethal, structural variants, dairy cattle, genomic prediction

Abstract :

The overall aim of this PhD thesis is to identify causal variants for recessive lethal mutations and select a set of predictive markers that are in high linkage-disequilibrium with the causal variants for female fertility in dairy cattle. We addressed this broad aim under five articles: (i) describes a systematic approach of mapping recessive lethals in French Normande cattle using homozygous haplotype deficiency (HHD). This study shows the influence of sample size, quality of genotypes, quality of (genotype) phasing and imputation, age of haplotype (of interest), and last but not the least, multiple testing corrections, on discovery and replicability of HHD results. It also illustrates the importance of fine-mapping with pedigree and whole-genome sequence (WGS) data, (cross-species) integrative annotation to prioritize candidate mutation, and finally, large-scale genotyping of the candidate mutation, to validate or invalidate initial results. (ii) describes a high-resolution population-scale mapping of large chromosomal deletions from whole-genome sequences of 175 animals from three Nordic dairy breeds. This study employs three different approaches to validate identified deletions. Next, it describes population genetic properties and functional importance of these deletions. (iii) deals with three main issues related to imputation of structural variants, in this case, large chromosomal deletions, e.g. availability of deletion genotypes, size of haplotype reference panel, and finally, imputation itself. To address the first two issues, this study describes a Gaussian mixture model-based approach where read-depth data from the variant call format (VCF) file is used to genotype a known deletion locus, without the need for raw sequence (BAM) file. Finally, it presents a pipeline for joint imputation of WGS variants along with large chromosomal deletions. (iv) describes genome-wide association studies for female fertility in three Nordic dairy cattle breeds using imputed WGS variants including large chromosomal deletions. This study is based on the analyses of eight fertility related traits using single-marker association, conditional and joint analyses. This study illustrates that inflation in association test-statistics could be seen even after correcting for population stratification using (genomic) principal components, and relatedness among the samples using genomic relationship matrices; however, this was known for traits with strong polygenic effects, among other factors. Finally, mapping of several new quantitative trait loci (QTL), along with the previously known ones, are reported in this study. This study also highlights the importance of including (imputed) large deletions for association mapping of fertility traits. (v) describes prediction of genomic breeding values for fertility using SNP array-chip genotypes, selected QTL and large chromosomal deletion. Using genomic best linear unbiased prediction (GBLUP) method with one or several genomic-relationship matrices derived from a set of selected markers, this study reports higher prediction accuracy compared with previous report. This study also highlights the influence of selecting markers with best predictability, especially for a breed with small training population, in accuracy of genomic prediction. The results demonstrate that large deletions in general have a high predictive performance.

Titre : Etude de délétions chromosomiques et de variants génétiques responsables de mortalité embryonnaire chez les bovins laitiers

Mots-clés : mortalité embryonnaire, variations structurales, bovins laitiers, sélection génomique

Résumé :

L'objectif général de cette thèse est d'identifier les variants causaux ou, à défaut, un ensemble de marqueurs prédictifs - qui présentent un déséquilibre de liaison élevé avec les variants causaux - pour la fertilité des vaches laitières. Nous avons abordé cet objectif général dans cinq articles: (i) décrit une approche systématique de cartographie des variants létaux récessifs chez les bovins Normands français basée sur la recherche de déficit en haplotypes homozygotes (HHD). Cette étude montre l'influence de la taille de l'échantillon, de la qualité des génotypes, de la qualité du phasage des génotypes en haplotypes et de l'imputation, de l'âge de l'haplotype et enfin, de la définition des seuils de signification prenant en compte les tests multiples, sur la découverte et la reproductibilité des résultats de HHD. Elle illustre également l'importance de la cartographie fine avec les données de généalogie et de séquence de génome entier (WGS), l'annotation intégrative (entre espèces) pour hiérarchiser les mutations candidates et, enfin, le génotypage à grande échelle de la mutation candidate, pour valider ou invalider les mutations initiales. (ii) décrit une cartographie à haute résolution de grandes délétions chromosomiques de séquences du génome dans une population de 175 animaux appartenant à trois races laitières nordiques. Cette étude utilise trois approches différentes pour valider les résultats de la cartographie. Le chapitre décrit les propriétés génétiques des populations et l'importance fonctionnelle des délétions identifiées. (iii) traite de trois questions liées à l'imputation de variants structuraux, ici de délétions chromosomiques importantes: la disponibilité des génotypes de délétion, la taille du panel de référence d'haplotypes et, enfin, l'imputation elle-même. Pour aborder les deux premières questions, cette étude décrit une approche basée sur un modèle de mélange gaussien dans laquelle les données de profondeur de lecture provenant de fichiers au format VCF (variant call format) sont utilisées pour génotyper un locus de délétion connu, en l'absence d'information sur la séquence brute. Enfin, il présente un pipeline pour l'imputation conjointe de variants WGS et de grandes délétions chromosomiques. (iv) décrit des études d'association pangénomiques de la fertilité femelle dans trois races de bovins laitiers nordiques à l'aide de variants WGS imputés et de grandes délétions chromosomiques. Cette étude concerne huit caractères de fertilité et utilise des analyses d'association mono-marqueur, conditionnelles et conjointes. Cette étude montre qu'une surestimation, ou « inflation », des statistiques de test peut être observée même après correction pour la stratification de la population à l'aide de composantes principales génomiques et pour les structures familiales à l'aide de matrices de relations génomiques. Ce biais était connu pour les caractères très polygéniques. Enfin, cette étude présente plusieurs locus de traits quantitatifs (QTL) nouveaux et confirme plusieurs autres déjà connus. Elle souligne également l'importance d'inclure les grandes délétions (imputées) pour la cartographie par association des caractères de fertilité. (v) décrit la prédiction des valeurs génomiques de fertilité (ou indice de fertilité) à l'aide de génotypes à puces SNP, de QTL sélectionnés et de délétions chromosomiques importantes. En utilisant la méthode de meilleure prédiction linéaire sans biais génomique (GBLUP) avec une ou plusieurs matrices de relations génomiques dérivées d'un ensemble de marqueurs sélectionnés, cette étude rapporte une précision de prédiction améliorée. Cette étude met également en évidence l'influence de la sélection des marqueurs les plus prédictifs, en particulier pour une race ayant une population d'apprentissage réduite, sur la précision des prédictions génomiques. Enfin, les résultats démontrent que les grandes délétions ont en général un pouvoir prédictif élevé.