STRUCTURAL VARIANTS IN THE BOVINE GENOME

Young Lim Lee

Wageningen University

Propositions

1. Increased bovine *de novo* structural mutation induced by *in vitro* fertilization is a blessing for breeders.
   (this thesis)

2. Claiming the causality of candidate variants should only be done if supported with both tissue-specific transcriptome and epigenome data.
   (this thesis)

3. Dutch scientific funding is too much focused on collaboration with industry instead of scientific excellence.

4. The current academic system relies too much on unpaid labour.

5. Overstating and oversimplifying scientific findings in media harms both society and science.

6. Our society should reward the young generation for starting a family earlier to prevent the pain caused by *de novo* mutation in older couples.

Propositions belonging to the thesis, entitled

Structural variants in the bovine genome

Young-Lim Lee
Wageningen, 26 April 2022

Structural variants in the bovine genome

Young-Lim Lee

Thesis committee

Promotor
Prof. Dr R. F. Veerkamp
Special Professor, Numerical Genetics
Wageningen University & Research

Co-promotors
Dr A. C. Bouwman
Researcher, Animal Breeding and Genomics
Wageningen University & Research

Dr M. Bosse
Researcher, Animal Breeding and Genomics,
Wageningen University & Research

Other members
Dr G. Sahana, Aarhus University, Denmark
Dr B. Harzilius, Topigs Norsvin, Beuningen
Prof. Dr Y. Bai, Wageningen University & Research
Prof. Dr M. G. M. Aarts, Wageningen University & Research

This research was conducted under the auspices of the Graduate School of Wageningen Institute of Animal Sciences (WIAS).

Structural variants in the bovine genome

Young-Lim Lee

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 26 April 2022
at 1:30. p.m. in the Aula.

Young-Lim Lee

Structural variants in the bovine genome

PhD thesis, Wageningen University, the Netherlands (2022)

With references, with summary in English

ISBN 978-94-6447-144-1

DOI https://doi.org/10.18174/566496

Abstract

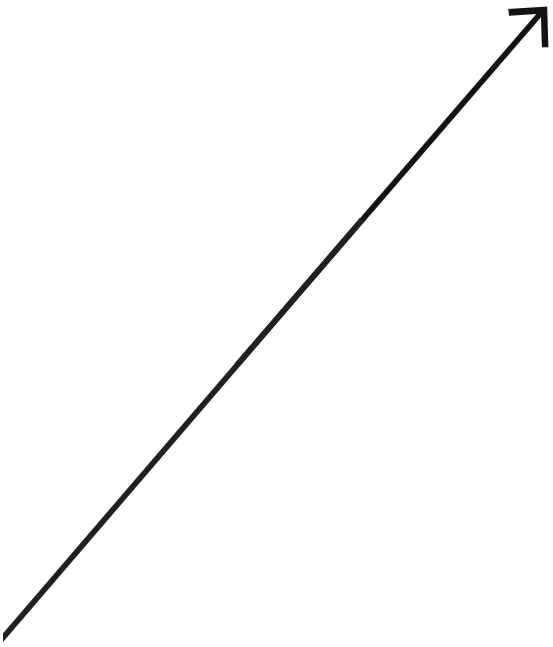
Lee, Y-L. (2022). Structural variants in the bovine genome. PhD thesis, Wageningen University, the Netherlands

Genome and environment together greatly influence how an organism develops and functions. Cattle is a livestock species with high economic significance; thus, deciphering how its genome and genomic information relates to its phenotype is crucial. Nevertheless, our understanding of bovine genomes is incomplete: investigation of bovine genomes thus far has mainly focused on single-nucleotide variants, leaving complex and less tractable variants, such as structural variants, underexplored. In this thesis, I present a comprehensive and in-depth investigation of structural variants in bovine genomes. Two structural variant catalogues, generated from genotyping data and deeply sequenced bovine genomes, revealed ~32 and ~5,000 SVs per genome, respectively. Furthermore, I integrated statistical associations and ~omics data, and delineated a 12-Kb copy number variant (CNV) as the likely causal variant of a major clinical mastitis QTL. This CNV encompasses an enhancer that targets the group-specific component gene, which encodes a vitamin D binding protein. Lastly, I investigated the emergence of *de novo* structural variants by exploiting 127 multi-generational deeply sequenced genomes. The pedigree-based germline mutation rate corresponded to one *de novo* structural variant per 8.5 births. The *de novo* structural variants were strongly biased towards male germlines and *in vitro* produced animals, unravelling sex and reproductive technology effects. Together, this thesis highlights a broad spectrum of bovine structural variants: from extremely rare *de novo* mutations to population variants, highlighting one that exerts strong effects on economically important traits. This thesis contributes to advancing our understanding of bovine structural variation.

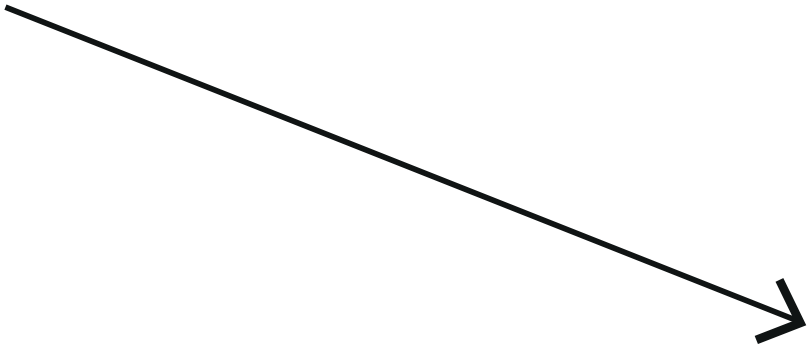
Contents

1

INTRO



DUCTION



1

1.1. The basis of cattle breeding

Humans have been living with, living off, relying on, and domesticating cattle for the last ~10,000 years (Frantz et al. 2020). Through domestication, ancient farmers likely selected and bred animals that suited their demands, such as tame behaviour and unique morphology (Georges 2007). Notably, the foundation of selective breeding is heredity, where genetic material is passed on from one generation to the next, however, the molecular basis was unknown for most of the bovine domestication process.

Modern cattle breeding is greatly enhanced and accelerated by state-of-the-art technologies, progress of which is driven by advancements in multiple scientific fields (Georges et al. 2019). Breeding programmes routinely collect pedigree information and phenotypes of diverse traits. The genetic information of cattle can now be elucidated at a molecular level, via e.g. genotyping arrays and whole genome sequencing (Womack 2012). A large amount of genomic and phenotypic data available for cattle populations provides opportunities for studying the bovine genome and how the genomic information relates to phenotypes. In this thesis, I look into structural variants, a subset of genetic variants present in bovine genomes. To introduce this topic, I provide an explanation of the bovine genome and of genetic variants below.

1.2. The bovine genome

A genome is a blueprint for a living organism containing deoxyribonucleic acid (DNA); this blueprint is used to encode all proteins that work together to form and maintain the organism. Genome and environment together influence greatly how an organism develops and functions. The cattle (bovine) genome spans over six billion nucleotides, which are distributed over 30 chromosome pairs. So far, ~22,000 genes have been discovered in the bovine genome, many of which encode proteins, and the rest serves regulatory functions. A stretch of DNA can be sectioned according to its functions: for instance, a small part of DNA is occupied by genes, many of which have a canonical structure, alternating between exons (coding part, encoding proteins) and introns (non-coding part). Intergenic DNA, which is not covered with genes, is mostly filled with repetitive sequences. Besides the functional nucleotide sequences, the genome also harbours the epigenome, which refers to biochemical and structural modifications that can alter the genome's function, and thus add another layer of information to the genome.

This thesis presents an in-depth investigation of hundreds of bovine genomes, which is not trivial, given that already a single genome contains an extensive amount of information, as elaborated above. A caveat is that when genomes of multiple animals are compared, a large part of the genomes is identical, meaning that they rarely vary. Those varying sites in genomes can potentially explain phenotypic differences observed in animals, and hence genomic variations are the basis for deciphering genome-to-phenome relationships. Variations in genomes come in different shapes and sizes, and hereafter I will call them genetic variants. Below, I introduce different types of genetic variants, including structural variants, the main focus of this thesis.

1.2.1. Genetic variants

Genetic variants come in different shapes and sizes. Single nucleotide variants (SNVs) include one base-pair nucleotide substitutions, insertions, or deletions. Small-scale insertions and deletions ranging between 1 and 50 base-pairs are grouped together as indels. Structural variations (SVs) are variants of 50-bp or larger, and include deletions, duplications, inversions, insertions, mobile elements insertions, translocations, segmental duplications, repeat expansions, and complex rearrangements (Figure 1.1). Among these, copy gain and loss events are referred to as copy number variants (CNVs). Upon the completion of the full sequence of the human genome in the early 2000s, it was expected that SNVs explain the majority of genetic differences between individuals. However, later discoveries showed that SVs are abundant, affect more nucleotide changes in the genome than small variants, and are present in healthy individuals (as opposed to previous textbook views that SVs are associated with diseases (Sudmant et al. 2015)). Below, I elaborate on how views on SVs have evolved over time and their relevance to genetic studies.

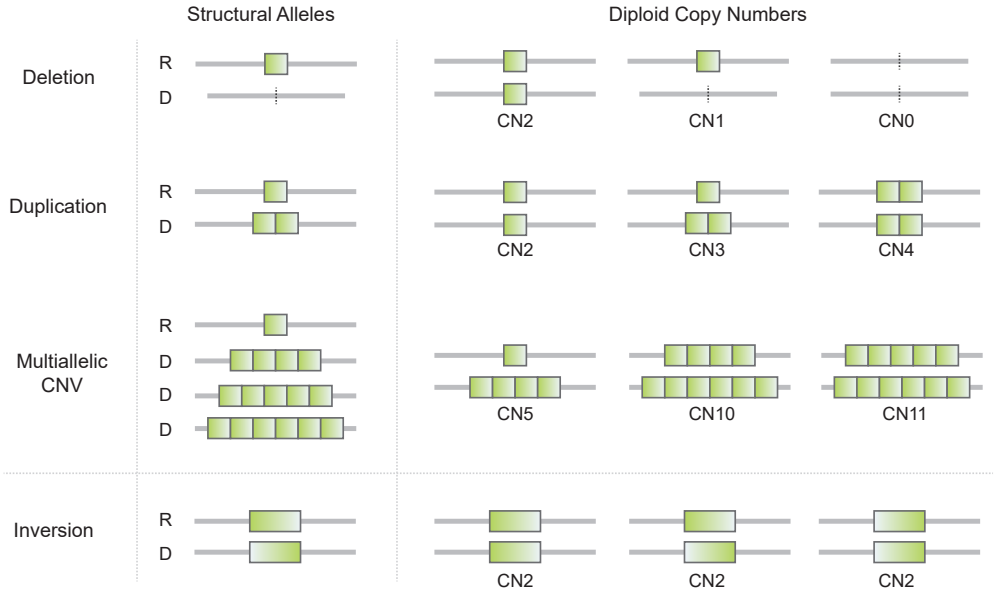


Figure 1.1. Different types of structural variants. A schematic overview of four different SVs (deletion, duplication, multiallelic CNV, and inversion). The structural alleles segregating at each SV are marked with D (derived) and non-affected alleles are marked with R (reference). The panel on the right side shows the diploid copy numbers (CN) that can appear for the four different SVs. Notably, a multiallelic CNV can have high CNs, as it harbours structural alleles with high CNs. The copy number of inversion is always CN 2, as it is balanced SV, which does not involve gain or loss of DNA.

1.2.2. Discovery of SVs

Initial discovery of SVs goes back to the early 20th century. Sturtevant (1913), whilst generating a genetic map of fruit flies (*Drosophila melanogaster*), realized a genetic factor inhibiting crossing over, that is, an inversion, in the third chromosome. Follow-up studies reported the Bar allele, a locus harboring a duplication that alters the shape of eyes in Drosophila (Tice 1914; May 1917; Zeleny 1919). Notably, these findings were achieved without sequencing and PCR, but solely relied on experimental breeding of numerous flies (Wolfer and Miller 2016).

Following these initial reports, our understanding of and the ability to study SVs expanded, along with technological advancements. McClintock (1931) exploited a cytological approach to observe inversions and translocations, and later discovered mobile genetic elements altering kernel color in Maize (*Zea mays*; McClintock 1950). Development of karyotyping techniques enabled the detection of large-scale chromosomal aberrations (Tjio and Levan, 1959), elucidating the etiology of important human diseases. These include the Down syndrome (caused by an additional copy of chromosome 21; Lejeune et al. 1959), Klinefelter syndrome, and Turner syndrome (caused by a missing X chromosome (X0) and an additional X chromosome (XXY), respectively; Ford et al. 1959; Jacobs and Strong 1959).

Following developments in banding technology**,** fluorescence *in situ* hybridization (FISH), and comparative genomic hybridization (CGH) improved the resolution and sensitivity of analyses from chromosomal to microscopic scale events of several million base pairs (Trask 2002). Exploiting these techniques, two landmark papers showed hundreds of large-scale CNVs (>100-Kb), segregating in phenotypically healthy individuals (Sebat et al. 2004; Iafrate et al. 2004). Many of these CNVs are rare, likely reflecting a strong selective pressure in relation to their high impact size. However, some of the CNVs are polymorphic and affect coding genes, suggesting that SVs can contribute to phenotypic differences.

Recent developments in Next Generation Sequencing (NGS) technology and SV detection tools have advanced SV detection further (Sudmant et al. 2015). Current NGS data (short-read sequencing at >30X coverage) and the most up-to-date detection algorithms can discover ~9,000 SVs in the human genome, including diverse types of SVs that are as small as 50-bp in size (Byrska-Bischop et al. 2021). These recent developments suggest that SVs might become a part of routine genetic analysis, which thus far has been focused on SNPs.

1.2.3. Why are SVs important?

Reports on SVs showed that they play a crucial role in evolutionary processes, and also are key determinants for phenotypic variations. In particular, SVs associated with a unique breed defining trait (e.g. color-sidedness) and economically important traits have been delineated in cattle populations, highlighting the importance of SVs in cattle breeding (Kadri et al. 2014; Durkin et al. 2012) Here, I explain some notable examples of SVs in cattle and other animals to demonstrate their significance.

Firstly, SVs can have significant evolutionary consequences at the population level, accelerating evolutionary processes such as differentiation and speciation. A 2.25-kb retrotransposon insertion in a European crow species gives rise to different plumage colors, resulting in a pre-zygotic barrier to other crow species (Weissensteiner et al. 2020). Alternatively, large chromosomal events, such as fusion and fission of chromosomes, result in animals with different karyotypes. Mating of these animals might result in viable but infertile offspring, functioning as a post-zygotic barrier.

Secondly, the roles of SVs are crucial at an organismal level within a population. As noted earlier, human studies underlined that SVs are not necessarily associated with diseases. Many SVs are common and segregate in healthy human cohorts, and they affect more base pairs than small variants. Hence, the genetic basis of various traits can be attributed to SVs. Some of the well-characterized molecular mechanisms of trait-associated SVs include (i) dosage effect, where changes in copy numbers affect gene expression and alter phenotype(s), (ii) gene interruption, (iii) gene fusion, and (iv) disruption of topologically associating domain (see Harel and Lupski (2017) for a detailed explanation for each mechanism).

In this thesis, I aim to characterize SVs in cattle genomes. Accordingly, I describe some unique and noteworthy SVs that are delineated at the molecular level among farm animals (see Clop et al. (2012) and Bickhart and Liu (2014) for review). Breed-defining coat color traits are often associated with well-characterized genes, such as *KIT* and *ASIP*. A few studies showed that SVs disrupting either the coding sequence or the regulatory regions of these genes could give rise to unique coat colors. For instance, serial translocation of *KIT* was associated with color-sidedness in Belgian Blue cattle (Durkin et al. 2012), whereas complex SVs of the same gene in domestic pigs leads to white coat color, as opposed to the dark coat in wild boars (Giuffra et al. 1999; Rubin et al. 2012). Likewise, an SV scan of goat breeds with diverse coat colors revealed that these breeds have unique SVs, disrupting either coding sequences or regulatory elements of *KIT* and *ASIP* genes (Henkel et al. 2019).

Morphologies valued by humans may lead to the formation of new breeds. A comparison between sheep populations exhibiting fleece variations (long and hairy fleece in ancestral breeds vs short and woolly fleece in modern breeds) revealed an *EIF2S2* retrogene insertion into the 3' untranslated region of the *IRF2BP2* gene, resulting in abnormal *IRF2BP2* transcripts and the woolly fleece phenotype (Demars et al. 2017). Furthermore, distinctive morphologies in farm animals can be utilized for herd management. The sex of young layer chicks can be determined based on their feather development, and the underlying mutation was a 176-kb tandem duplication disrupting the *PRLR* and *SPEF2* genes (Elferink et al. 2008).

Trait associated SVs are not limited to morphologic traits like those discussed above. Various disease-associated deletions have been delineated, including a 3.3-Kb partial deletion of the *FANCI* gene, leading to embryonic loss in homozygous carrier in cattle (Charlier et al. 2012). Furthermore, recent studies discovered a large deletion under balancing selection, where a pleiotropic 660-kb deletion in Nordic red cattle increased milk production, despite resulting in an embryonic loss in a homozygous carrier (Kadri et al. 2014). Similarly, a 212-kb deletion

segregating in a commercial pig population showed pleiotropic effects. This deletion is located within the *BBS9* gene and harbors an enhancer for the *BMPER* gene, which is located downstream of *BBS9*, resulting in suppressed expression of *BMPER*. This deletion results in faster growth of heterozygous carriers, yet leads to an embryonic loss in homozygous carriers (Derks et al. 2018).

1.2.4. How do we detect SVs?

As elaborated earlier, the understanding of SVs has been driven by technological advancements. Among various techniques that enabled SV studies, I will elaborate on two approaches, SNP genotyping array and short-read WGS, both of which were used in the research described in this thesis.

1.2.4.1. SNP genotyping array

A SNP genotyping array contains probes that hybridize to polymorphic sites in the genome. B allele frequency and hybridization intensity values are used to infer underlying copy number states (Alkan et al. 2011). Globally, millions of animals are being genotyped, as genomic prediction has become a routine practice in many cattle breeding programmes (Wiggans et al. 2017). Thus, SV detection, based on the already generated SNP array data, is a cheap option for high throughput SV screening. A caveat of the array-based approach is that the detection resolution is bound to the density of the probes. The most widely used bovine SNP arrays are the 50K density and BovineHD array, harboring 50K and 770K SNPs, respectively. CNVs of a mean size of ~150-Kb and ~50-Kb can be discovered using the 50K and 770K density arrays (Sasaki et al. 2016; Bae et al. 2010). However, despite the economic benefits, this approach has some downsides: (i) detectable events are limited to unbalanced SVs (e.g. CNVs), (ii) repetitive regions are poorly covered by probes (ascertainment bias), and (iii) breakpoints are unresolved.

1.2.4.2. Next generation sequencing (NGS)

DNA sequencing determines the order of nucleotides in a genome. Given the overwhelming length of the bovine genome, sequencing from start to the end of the genome can be arduous and low-throughput. NGS technology revolutionized this sequencing process by shearing the whole genome into short fragments of 500-800-bp length, and then sequencing them in parallel. The sequencing will be done for ~100-250-bp ends of each fragment in paired-end sequencing technology. Then the fragments are aligned to a reference genome. Genomic positions where the reads are abberantly aligned suggest structurally variant site(s) present in the sample, relative to the reference genome. Such aberrant alignments can be screened to infer presence of SVs (Figure 1.2).

Largely, SVs can be inferred from three different signals: read-depth (RD) signals, clusters of discordantly mapped read-pairs (RP), and clusters of split reads (SR). Firstly, RD of copy number variant regions will differ compared to diploid regions. Detection methods screen genome-wide RD signals, using a window-based approach (Abyzov et al. 2011). RD-based SV detection is computationally fast and performs well in detecting large-scale CNVs (Kosugi et

al. 2019). However, detection is limited to CNVs, leaving copy number neutral events (e.g. inversions) uncaptured. Also, the detection resolution is low, and the breakpoints are left unresolved.

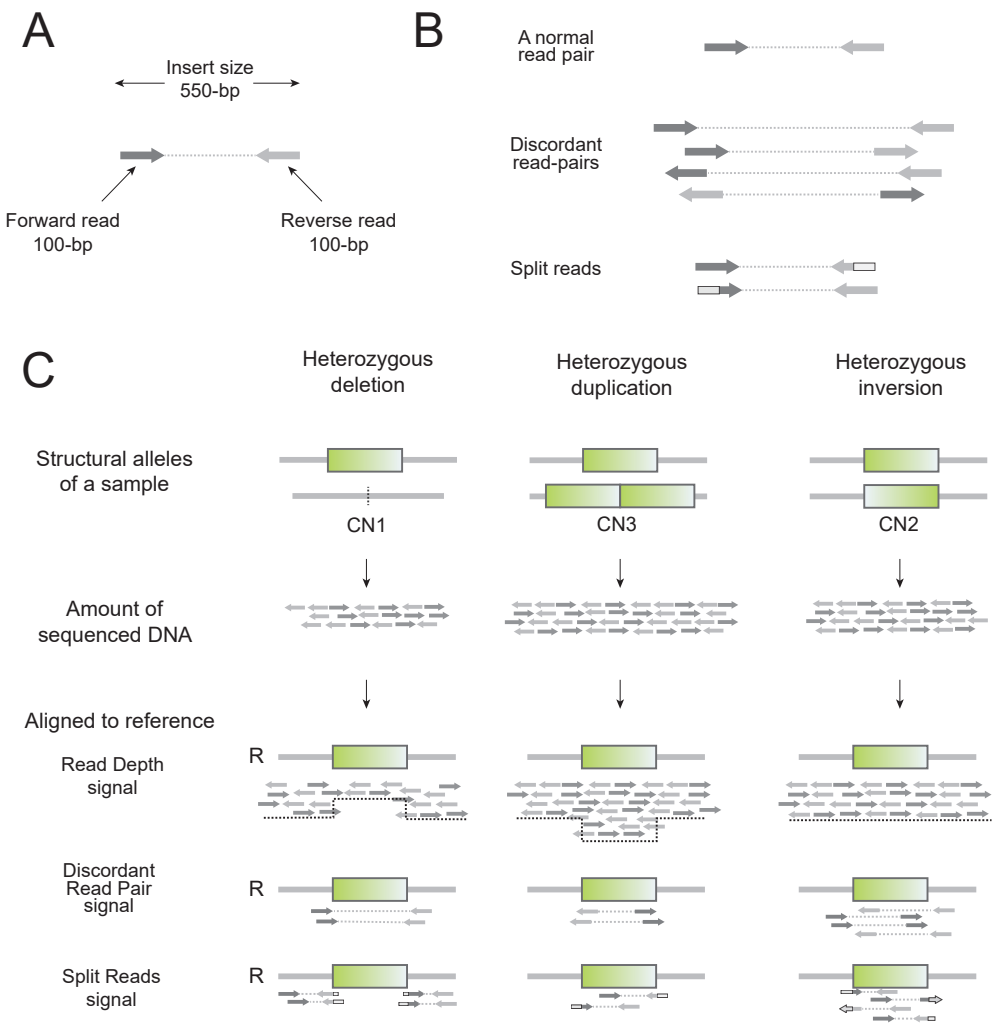


Figure 1.2. Different sequencing reads alignment signals used in SV detection. (**A**) An example of a sequencing read pair based on the WGS data used in this thesis. The insert size was 550-bp, where forward (marked with dark grey) and reverse (marked with light grey) each sequenced 100-bp of both ends of the insert. The inner ~350-bp is not sequenced. (**B**) A schematic representation of discordant and split reads compared to the normally aligned read pairs. Discordant read pairs are aligned further than expected (hence the distance between the reads are larger than 350-bp for the WGS data used in this thesis) and the orientation of the reads (either forward or reverse) may be different than the canonical orientation. Split reads are marked with light grey shade. (**C**) A schematic overview of detection of SVs using NGS data based on read depth, discordant read pairs, and split reads signals.

Secondly, RP may have aberrant insert size (longer than normally mapped RP) or non-canonical orientation(s). In principle, aberrantly mapped RPs can detect many types of SVs (deletions, duplications, inversions, and translocations). However, detection relying solely on RP signals was shown to generate many false positive calls (Cameron et al. 2019). Also, detectable events are larger than the fragment size, leaving smaller events undetected.

Thirdly, SR refers to sequencing fragments where reads are split and aligned to different regions. Clusters of SR appear at breakpoints of SVs, and hence are informative in delineating the underlying structure. SV detection solely relying on SR is computationally expensive and time-consuming, but can achieve finer resolution compared to RD- and RP- based methods (Cameron et al. 2019).

Extensive benchmarking of SV detection performance of numerous tools concluded that (i) software packages exploiting multiple signals perform better than those relying on a single signal and (ii) no single software package detected all SV types of various size ranges with high detection sensitivity (Cameron et al. 2019; Kosugi et al. 2019). Instead, hybrid callers, combining different detection signals, can have better performance compared to single signal callers. Also, a large genetics study ventured into combining multiple SV detection software packages, to build an ensemble SV calling and QC pipeline (Abel et al. 2020; Collins et al. 2020).

1.2.5. How many SVs are present in the bovine genome?

Until now, more than 20 studies characterized SVs in bovine genomes, reporting a variable number of SVs. In Table 1, three representative studies using different datasets are shown, and these resulted in large differences in the number of loci discovered. Such discrepancies are expected, given the inherent tendency to detect different SVs in varying size ranges depending on input data and detection method used.

Table 1.1. Number of SVs discovered in bovine genomes

Study	No. samples*	Data	No. loci#	No. per genome
Bae et al (2010)	265 (1)	50K SNP array	368 CNVRs	~3.2 CNVs
Sasaki et al (2016)	1,481 (1)	BovineHD array	861 CNVRs	~35 CNVs
Mesbah-Uddin et al (2017)	175 (3)	WGS (~10X)	8,480 DELs	~2,000 deletions

*Number in the brackets indicates the number of breeds; #CNVR stands for CNV regions

1.3. Emergence and fate of *de novo* mutations

So far, I discussed SVs as if they are present in the genome by default. In reality, most SVs are inherited from the parents via parental reproductive cells (gametes), whereas only a small fraction of SVs in a given individual are novel mutations that arose spontaneously (*de novo* SVs). Gametes are haploid cells derived from the germlines, serving the specialized function of carrying DNA from one generation to the next. To deliver the DNA intact, fidelity, an ability

to proofread replication errors and to repair DNA damage, is crucial in the germline. An extensive survey on the mutational landscape in different cell types confirmed that the germline maintains high fidelity throughout the lifespan, compared to other somatic cell types (Coorens et al. 2021).

1.3.1. Emergence of *de novo* mutations

From the moment of fertilization, a zygote undergoes rapid cell replications. During each cell division, DNA is replicated. This process may involve replication errors, or DNA damage may occur. These changes to the DNA result in *de novo* mutations (DNMs) if they are not or only erroneously repaired, unless they cause cell death (apoptosis). Although the scope of this thesis is limited to SVs, here I explain about DNMs of all types, not only those limited to *de novo* SVs (dnSVs). DNMs can arise in all cells. Somatic DNMs might induce diseases in the individual carrying the DNM, but not in the next generation. On the contrary, germline DNMs can be transmitted to the next generation. Thus high impact germline DNM forms an important research subject for sporadic disease studies.

Germline DNMs were assumed to occur in parental germlines, mainly due to spontaneous mutations, thus affecting a single, if not a few, gametes. Particularly, the male germline was considered more prone to mutations, even before a direct investigation was possible at a molecular level (Haldane 1947). This view was confirmed in NGS data-driven, direct investigation on human data (Conrad et al. 2011; Kong et al. 2012). However, contrary to the assumption that DNMs mostly arise in germlines, a study on bovine pedigrees showed that the first zygotic cell division is highly mutation-prone (Harland et al. 2017). This study underlines that a significant fraction of detectable DNMs is post-zygotic in origin (not from a parental gamete). Such post-zygotic DNMs can develop into cell lineages that later differentiate into the germline. As a consequence, the germline affected by mosaic DNMs can produce multiple gametes carrying the post-zygotic DNM, showing recurrent transmission of the DNM.

1.3.2. How does a DNM become a population variant?

So far, I have explained DNMs arising in a single animal. However, livestock populations consist of a large number of animals, where evolutionary forces, such as drift, selection, and the effect size of DNMs, orchestrate the fate of DNMs within a population.

1.3.2.1. From a DNM to a population variant

Animals are born carrying millions of SNPs and thousands of SVs. Of these numerous variants, only a fraction consists of DNMs that arose either in the parental gametes or the offspring itself, while the vast majority were transmitted from ancestors over generations. These variants, which have been transmitted through generations, are present and segregate in a population, and hence are referred to as population variants. A population variant also once was a DNM that occurred in an ancestral animal and underwent drift and/or selection.

The neutral theory of molecular evolution states that mutations, if they do not have deleterious effects, will be neutral (thus have no-to-benign effects), hence they are under neutral selec-

tion, and some may be fixed due to genetic drift. In contrast, beneficial DNMs are rare and if present at all, they can reach fixation rapidly (Kimura 1968). Applying this view not only at an organismal but also a cell and zygote level, one can conclude that a DNM (here, assuming germline DNM in parental gametes, for brevity) can segregate as a population variant, if (i) the DNM does not lead to cell death in the germline, (ii) DNM carrying gametes are viable, and hence can be fertilized, (iii) viable offspring can be produced, meaning no embryonic loss, (iv) the liveborn offspring does not suffer from disease(s) that prevent it from reproducing. DNMs that pass these multiple selection steps can segregate as population genetic variants, although many will be lost due to genetic drift.

1.3.2.2. A strong bottleneck in livestock populations: artificial selection and drift

On top of the selection steps discussed above, DNMs in livestock populations face further complications. Livestock animals are under artificial selection, where especially animals with desired phenotypes are allowed to pass on their DNA to the next generation. Selection in modern dairy cattle breeding relies on genomic selection (GS), where the genetic merit of SNP markers that evenly cover the genome is known (Meuwissen et al. 2001). As GS does not take into account the effect of DNMs, the DNMs in non-selected animals are lost due to drift, even if the DNMs have a positive effect on the traits under selection (Mulder et al. 2019). Contrastingly, this also results in segregation of DNMs, even the ones with a negative effect, if they arose in animals selected for mating. The fate of DNMs, which survived the strong bottleneck induced by artificial selection, is then dependent on their effect and mode of inheritance.

1.3.2.3. Fate of a DNM with a positive effect

The fate of a DNM with a positive effect on a desired phenotype may be determined by the effect size and mode of inheritance. A dominant variant with a strongly positive effect will rapidly segregate in a population. It does so more strongly when appearing in elite sires, which tend to have a large offspring pool, ultimately reaching fixation. When it comes to recessive DNMs, they may be lost by drift before reaching a meaningful frequency. Yet, if the phenotype in homozygous carriers is strongly desired, as shown in the double muscling phenotype caused by an 11-bp deletion in *MSTN* gene in beef cattle breeds, it may reach fixation in populations, in which it is heavily selected for (e.g. Belgian Blue cattle breed; Grobet et al. 1997; McPherron and Lee 1997). An exception is balancing selection, where a single variant confers a heterozygous advantage, thus remaining in a population unfixed, despite its undesirable effects on homozygous carriers (Derks and Steensma 2021). To what degree dnSVs can have positive effects on economically important traits of dairy cattle is currently unknown.

1.3.2.4. Fate of a DNM with a negative effect

As with positive effects, the fate of a DNM with a negative effect will be determined depending on the mode of inheritance and effect size. Animals born with a dominant DNM causing severe deleterious effects (e.g. physical abnormalities) can be identified and consequently the DNM will be purged. In contrast, a recessive DNM with negative effects may be unnoticed in a heterozygous state, and hence can segregate in a population. If this mutation occurs in an

elite sire with a large offspring pool, the population frequency will increase rapidly, eventually leading to a recessive defect outbreak (Charlier et al. 2008). Analyses on deeply sequenced human genomes (n>4,000) using *in silico* methods estimated that an individual carries on average ~120 deleterious variants. Of these, 17% were SVs, showing a strong overrepresentation of SVs in deleterious variants (840-fold), given the numerical abundance of SNVs and indels (Abel et al. 2020).

1.4. Aims and scope of this thesis

Modern dairy cattle populations represent a unique opportunity for genetic studies, given (i) the large amount of genomic data generated, (ii) population structure and pedigree records, (iii) routine phenotyping performed by breeding programmes, and (iv) the wide adoption of assisted reproduction technology (ART), thereby providing a model to investigate the impact of ART on *de novo* SVs.

In this thesis, I catalogued genome-wide SVs in dairy cattle populations, using high density genotyping array and whole-genome sequencing (WGS) data. Furthermore, I performed an in-depth investigation on two types of SVs, (i) a common variant associated with economically important traits and (ii) an extreme form of rare structural variants, *de novo* SVs.

In **Chapter 2**, I catalogued CNVs in two dairy cattle populations, Holstein Friesian and Jersey, using high-density SNP genotyping array data. Population genetics tools were applied to CNVs, which provided insight into divergently selected CNVs between the two cattle populations.

I further refined the SV catalogue by exploiting WGS data in **Chapter 3**. I quantified how many SVs affect coding sequences and thus are likely to have functional consequences. By mapping expression QTL of selected SV loci, I showed that some SVs have a functional impact at the expression level.

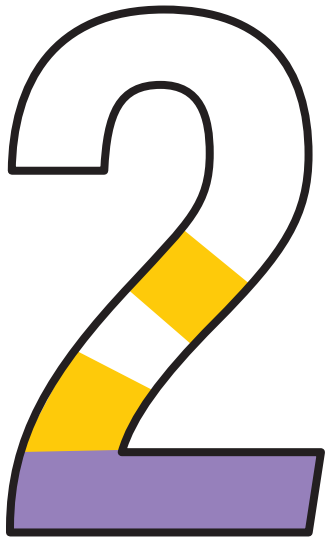
In **Chapter 4**, I performed an in-depth analysis of a 12-kb CNV overlapping the *Group-specific Component* (*GC*) gene, a candidate causal gene for a major QTL for clinical mastitis (CM) resistance. I dissected this QTL by integrating association analyses, expression QTL mapping, and the bovine epigenome map and showed that the 12-kb CNV is likely the causal variant underlying the major CM resistance QTL.

In **Chapter 5**, I screened *de novo* SVs from the WGS SV catalogue obtained from a healthy bovine family cohort. The germline *de novo* SVs unraveled an extreme paternal bias and a strong ART effect.

Finally, in **Chapter 6,** I discuss the key findings of the current thesis and discuss trends in SV research.

Frequently used abbreviation can be found in the glossary section in appendices.

CHAPTER



Functional and population genetic features of copy number variations in two dairy cattle populations

Young-Lim Lee[1*], Mirte Bosse[1], Erik Mullaart[2], Martien A. M. Groenen[1], Roel F. Veerkamp[1], Aniek C. Bouwman[1]

[1] Wageningen University & Research, Animal Breeding and Genomics , P.O. Box 338, 6700 AH Wageningen, the Netherlands [2] CRV, Arnhem, the Netherlands

Published in BMC Genomics, 2020, 21:89,

2.1. Abstract

Background Copy Number Variations (CNVs) are gain or loss of DNA segments that are known to play a role in shaping a wide range of phenotypes. In this study, we used two dairy cattle populations, Holstein Friesian and Jersey, to discover CNVs using the Illumina BovineHD Genotyping BeadChip aligned to the ARS-UCD1.2 assembly. The discovered CNVs were investigated for their functional impact and their population genetics features.

Results We discovered 14,272 autosomal CNVs, which were aggregated into 1,755 CNV regions (CNVR) from 451 animals. These CNVRs together cover 2.8% of the bovine autosomes. The assessment of the functional impact of CNVRs showed that rare CNVRs (MAF<0.01) are more likely to overlap with genes, than common CNVRs (MAF>0.05). The Population differentiation index (Fst) based on CNVRs revealed multiple highly diverged CNVRs between the two breeds. Some of these CNVRs overlapped with candidate genes such as *MGAM* and *ADAMTS17* genes, which are related to starch digestion and body size, respectively. Lastly, linkage disequilibrium (LD) between CNVRs and BovineHD BeadChip SNPs was generally low, close to 0, although common deletions (MAF>0.05) showed slightly higher LD (r^2=~0.1 at 10kb distance) than the rest. Nevertheless, this LD is still lower than SNP-SNP LD (r^2=~0.5 at 10kb distance).

Conclusions Our analyses showed that CNVRs detected using BovineHD BeadChip arrays are likely to be functional. This finding indicates that CNVs can potentially disrupt the function of genes and thus might alter phenotypes. Also, the population differentiation index revealed two candidate genes, *MGAM* and *ADAMTS17*, which hint at adaptive evolution between the two populations. Lastly, low CNVR-SNP LD implies that genetic variation from CNVs might not be fully captured in routine animal genetic evaluation, which relies solely on SNP markers.

Keywords: Copy number variations, *Bos taurus*, Linkage disequilibrium, population genetics

2.2. Background

Genetic variations exist in various forms in genomes. Although single nucleotide polymorphisms (SNPs) have been the choice of variants in numerous studies, there is a growing body of evidence that copy number variations (CNVs) can have functional impact. Copy number variations are DNA segments of 1 kb or larger, and are present in varying copy numbers, compared to a reference genome (Feuk et al. 2006). Since the initial discovery of large sub-microscopic CNVs (some hundred kb; Iafrate et al. 2004; Sebat et al. 2004), rapid developments in detection platforms and algorithms have advanced knowledge about CNVs, mainly in humans (Alkan et al. 2011; Sudmant et al. 2015).

In the early phase of their discovery, CNVs were expected to resolve the missing heritability (significant SNPs identified from genome-wide association studies (GWAS) together account small part of the heritability; Manolio et al. 2009; Eichler et al. 2010). It was because, as in terms of base pairs, they cover a larger proportion of the genome, compared to SNPs. With the accumulation of data and analyses, the occurrence in the genome of CNVs was shown to be biased outside of functional elements (Sudmant et al. 2015). Nevertheless, numerous studies have shown that CNVs play a role in determining a wide range of human health conditions, from obesity to neurodevelopmental diseases (Bochukova et al. 2010; Coe et al. 2014; Macé et al. 2017; Marshall et al. 2017). For instance, high copy numbers of the *CCL3L1* and *CYP2D6* genes confer reduced susceptibility to infection with HIV and the development of AIDS (Gonzalez et al. 2005). Also, the role of CNVs in adaptive evolution is further exemplified by mean copy numbers of the *AMY1* gene (which codes for amylase alpha1, an essential enzyme for starch digestion). The mean copy number of *AMY1* gene was shown to differ in human populations depending on dietary starch composition (Perry et al. 2007). These findings demonstrate that CNVs may contribute to adaptive potential, and thus contain information about population history.

Studies in livestock species also highlighted the role of CNVs in affecting various phenotypes. For example, several genes affected by CNVs determine coat colours of specific breeds. Duplications of the *KIT* gene in pigs are related to white coat, which is only shown in domestic pigs (Rubin et al. 2012; Giuffra et al. 2002). In cattle, serial translocation of the *KIT* gene was related to a colour-sidedness phenotype (Durkin et al. 2012). Moreover, CNVs were shown to be associated with quantitative traits that are economically important in livestock breeding, in various cattle populations (Xu et al. 2014; Zhou et al. 2018; Prinsen et al. 2017). One study investigated whether trait associated CNVs are in linkage disequilibrium (LD) with, and thus are tagged by, SNP markers, and revealed that ~25% of CNVs were not in LD with SNP markers (Xu et al. 2014). However, this study was based on Illumina BovineSNP50 array data, in which SNP density and CNV resolution were low.

Holstein Friesian (HOL) and Jersey (JER) are the two main commercial dairy cattle breeds that have been bred under different breeding schemes. Although there have been studies investigating the link between CNVs and individual production traits (Zhou et al. 2018; Durán

Aguilar et al. 2017; Prinsen et al. 2017; Ben Sassi et al. 2016; Xu et al. 2014), in-depth assessment of functional impacts of CNVs in cattle genomes has been limited. Also, whether CNVs that have an impact on phenotypes are captured in genomic evaluation, in other words, whether CNVs are in sufficient LD with SNPs, is largely unexplored. Furthermore, CNVs have been shown to be useful in disentangling population history and provide valuable insights in understanding how populations have evolved over time (Xu et al. 2016; Bickhart et al. 2016; Upadhyay et al. 2017; Pierce et al. 2018). However, population genetics analyses exploring CNVs, with their main focus on HOL and JER, have been sparse.

Here, we aimed at discovering CNVs in bovine genomes based on genome assembly ARS-UCD1.2 (USDA ARS 2018) using high density SNP array data, in two dairy cattle populations. Subsequently, we performed in-depth analyses on the functional impact of CNVs and further explored the population genetic features of CNVs by analysing population differentiation index (Fst) and LD.

2.3. Results

2.3.1. CNV discovery in the genome build ARS-UCD1.2

The data consisted of Illumina BovineHD BeadChip (Illumina, San Diego, CA, USA) genotypes from two distinct dairy breeds (Holstein Friesian – HOL (n=331), Jersey – JER (n=115)) and their crossbreds (n=29). A previous study using PennCNV on BovineHD data, of which 47 HOL animals overlapped with our study, showed high rate of CNV confirmation based on qPCR validation (91.7% for CNVs found in multiple animals, 40% for singleton CNVs; Upadhyay et al. 2017). Therefore, we chose to perform CNV detection on bovine autosomes using the PennCNV software (Wang et al. 2007).The Bovine HD SNPs were aligned to genome assembly ARS-UCD1.2.

We discovered 14,272 CNV calls from 451 individuals that passed the quality control criteria (31.6 calls/individual). Deletion calls were 1.8 times more frequent but 40% shorter (n=9,171, mean length=44.2 kb) than duplication calls (n=5,101, mean length=74.6 kb; Supplementary Table 1 and Supplementary Fig. 1). The mean probe density (number of supporting SNPs per Mb CNV) was 403 SNPs/Mb. The 14,272 CNV calls were aggregated into 1,755 CNV regions (CNVRs), based on at least 1 bp overlap, following Redon *et al.*(2006a). These CNVRs cover 2.8% of the autosomal genome sequence (69.6/2,489.4 Mb; Figure 2.1; A full list of CNVR is in Supplementary Table 2.). These CNVRs consist of 1,125 deletion CNVRs (mean length=29.2 kb), 513 duplication CNVRs (mean length=36.8 kb), and 117 complex CNVRs (mean length=152.7 kb). The distribution of CNVR length is exponential, where the majority CNV are short to medium length (<100 kb, 93%), while only a few observations were made for long CNVRs (>100 kb, 7%).

The CNVRs are non-randomly distributed over the chromosomes: chromosome-wide CNVR coverage varies from 0.6% on BTA24 to 4.9% on BTA12 (Table S3). BTA12 is most densely covered with CNVR in terms of bp (4.2 Mb), and especially enriched for complex type CNVRs

(2.2 Mb). Allele frequency of CNVRs ranges between 0.001 and 0.21.

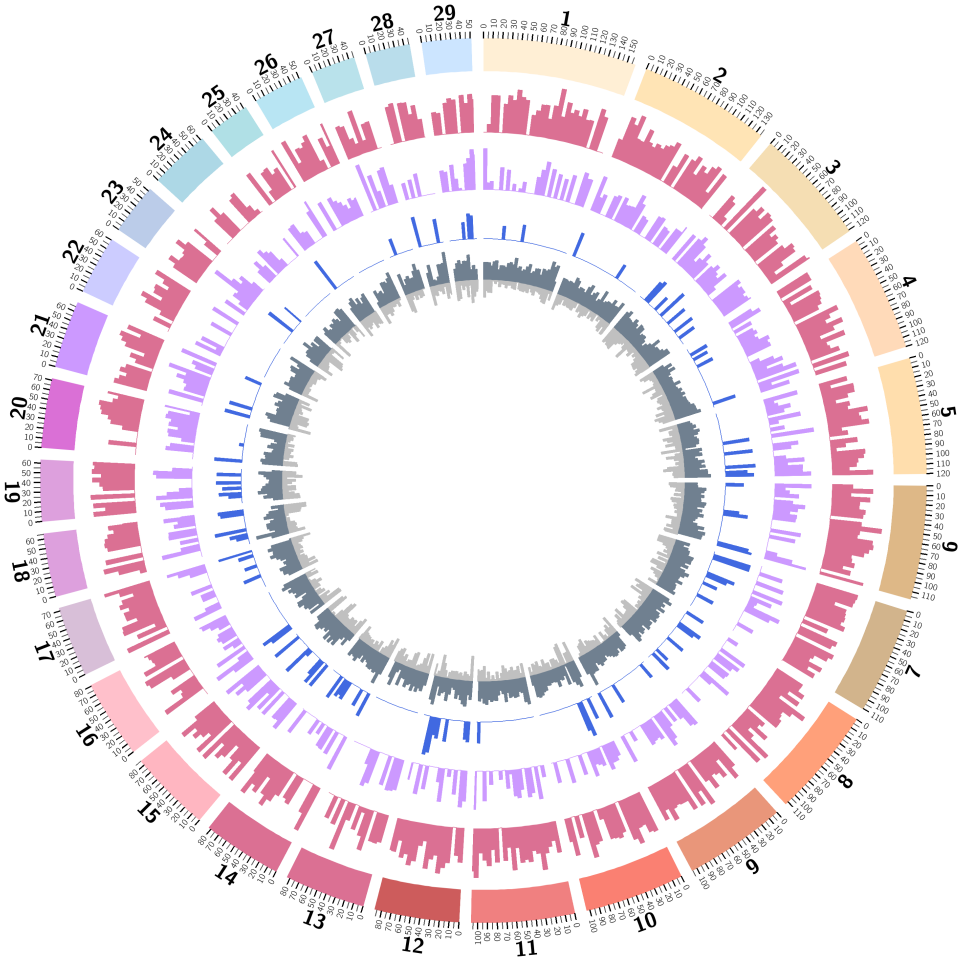


Figure 2.1. Circular map of autosomal copy number variant regions and their population genetics features. From the outside to the inside of the external circle: chromosome name; genomic location (in Mb); histogram representing density of deletion CNVRs in 5 Mb bin (pink); histogram representing density of duplication CNVRs in 5 Mb bin (purple); histogram representing density of complex CNVRs in 5 Mb bin (blue); number of BovineHD BeadChip array SNPs in 5 Mb bin (dark grey); histogram representing density of segmental duplications in 5 Mb bin (light grey).

Since most cattle CNV studies used genome assembly UMD3.1, we also repeated the CNV detection procedures, using UMD3.1. Subsequently, we used these calls to assess our CNV discovery results with other cattle CNV papers. From the 447 individuals that passed the QC criteria, 24,264 CNVs were called (54.3 calls/individual) and the mean probe density was 326 SNPs/Mb. These CNVs were aggregated into 1,866 CNVRs (1,130 deletions, 593 duplications,

and 143 complex CNVRs). The mean length of deletion, duplication, and complex CNVRs was 29, 36, and 193 kb, respectively (Supplementary Table 1). These CNVRs together cover 82 Mb (3.3%) of bovine autosomes. The chromosome-wide coverage varies between 1% on BTA24 and 10% on BTA12 (Table S4 and Supplementary Fig. 2). Compared to other cattle CNV studies conducted using the same SNP array and the genome assembly UMD3.1 (Jiang et al. 2013; Sasaki et al. 2016; Xu et al. 2016; Prinsen et al. 2016; Upadhyay et al. 2017; Nandolo et al. 2018), our CNV discovery results are in a similar range (Table S5).

When we compared to our CNVs discovered based on UMD3.1 and ARS-UCD1.2, we observed several differences. Firstly, the number of CNVs called per individual based on ARS-UCD1.2 is 42% lower than what was obtained using UMD3.1. Also, the mean probe density increased from 326 SNPs/Mb in UMD3.1 to 404 SNPs/Mb in ARS-UCD1.2, indicating that with ARS-UCD1.2, CNVs are supported by more SNPs. Lastly, the mean length of complex CNVRs decreased by 40kb, from 193 kb in UMD3.1 to 152.7 kb in ARS-UCD1.2. We further inspected BTA12:70-77 MB region where a large change between UMD3.1 and ARS-UCD1.2 was observed. This region was reported to have a large number deletion and duplication calls by other cattle CNV studies based on UMD3.1, regardless of the studied breeds (Hou et al. 2012; Jiang et al. 2013; Sasaki et al. 2016; Prinsen et al. 2016; Upadhyay et al. 2017; Nandolo et al. 2018). In our CNV discovery, we identified 7 CNVRs (total length of ~6.2 Mb) in this region based on UMD3.1, whereas ARS-UCD1.2 based results revealed 9 CNVRs that covered ~1 Mb. We compared the positions of BovineHD SNPs in UMD3.1 and ARS-UCD1.2 to see whether the changes in genome assemblies caused this discrepancy. The results showed that 43% of the SNPs located in BTA12:70-77Mb based on UMD3.1 were either moved to unmapped contigs or reference and alternative SNPs were undefined. The genome-wide ratio of SNPs that were moved to different chromosomes or contigs was much lower (2.3%) than 43%. This indeed indicates that the two genome assemblies differ in this regions, and thus led to different CNV discovery results.

2.3.2. Functional impact of CNVRs

The expression of genes can be altered by CNVs. Deletions and duplications of a part of and/or complete gene can disrupt the gene expression and can potentially lead to changes in various phenotypes (Lupski and Stankiewicz 2005). Therefore, identification CNVRs that coincide with genes can be a primary step to assess their functional impact. To achieve this, we explored CNVRs found based on ARS-UCD1.2 further. The overlap of CNVRs with Ensembl annotated genes were analysed, and among the 1,755 CNVRs, 912 (52%) are genic and 843 (48%) are intergenic. Genic CNVRs overlap with 1,739 genes out of 27,570 Ensembl annotated genes (6.3%) and 2,936 out of 43,949 gene transcripts (6.7%). Among the 1,739 genes that overlap with CNVRs, 957 (55%) are completely within the CNVRs and the rest (45%) are partially affected (genic features were inside the CNVRs).

The following functional impact categories were assigned to each CNVR depending on types of overlap between CNVRs and genes (numbers in the brackets indicate number of CNVRs and genes respectively for each category; see materials and methods for detailed explanation for

the classification): 1) intergenic (843 CNVRs; 0 genes), 2) intronic (214 CNVRs; 234 genes), 3) whole gene (253 CNVRs; 957 genes), 4) stop codon (147 CNVRs; 203 genes), 5) promoter regions (124 CNVRs; 187 genes), and 6) exonic (174 CNVRs; 165 genes). Then, these functional categories were intersected with other features of CNVRs such as types (deletion, duplication, complex), MAF (common, intermediate, and rare; see methods for detailed explanation), and the populations (HOL and JER; Figure 2.2).

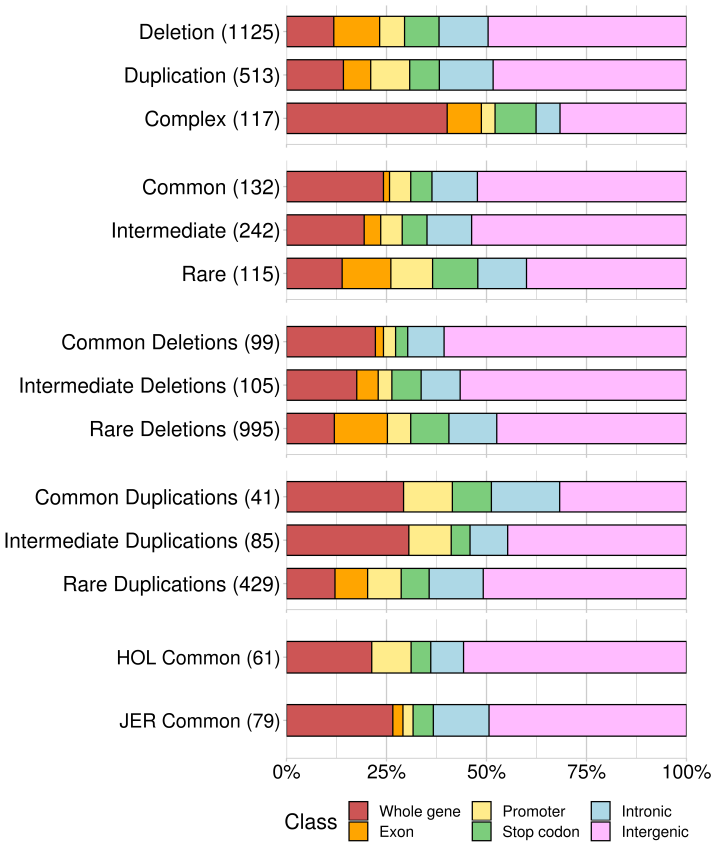


Figure 2.2. Functional impact of CNVRs by type, frequency, and population. Functional impact of CNVRs were investigated by type, frequency, and population. CNVRs were categorized into different types (deletion, duplication, and complex) and frequency (common: 0.05≤MAF in any population, intermediate: 0.01≤MAF<0.05, rare: MAF<0.01 in all populations). The numbers in the brackets indicate the number of CNVRs in each category.

The functional consequences of CNVRs differ depending on the type of CNVRs: Complex CNVRs were skewed towards genic regions (68% are genic), whereas deletions and duplication CNVRs were biased away from genic regions (51-52% are genic), and the difference is significant (chi-square test P $<10^{-13}$). Also, we observed that MAF have impact on different types of overlap between genes and CNVRs. Rare CNVRs tend to be genic more often (60%), whereas common CNVRs have less overlap compared to it (48%; chi-square test P < 0.002). However,

when seen it separately for deletion CNVRs and duplication CNVRs, we saw a different pattern. Common deletion CNVRs are more often intergenic (61%), yet the common duplication CNVRs are often genic (68%). When CNVRs between HOL and JER are compared, common JER CNVRs are more often genic (51%), than common HOL CNVRs (44%). Subsequently, we performed permutation tests on overlaps between CNVRs and autosomal genes, to test whether the overlap is significantly higher than expected under a neutral scenario. The results show that CNVRs overlap with autosomal genes more often than what is expected from permutation tests with random genomic regions ($P<0.001$). Nextly, gene ontology analyses were performed to understand the functions of the genes that overlap with CNVRs. Genes overlapping deletions, duplications, and complex CNVRs were tested for GO enrichment as separate classes (Table 1). Among the findings, genes overlapping with the complex CNVRs (n=407) show a pronounced enrichment in response to stimulus (GO:0050896; FDR=$1.8x10^{-6}$), immune response (GO:0006955; FDR=$1.9x10^{-3}$), and detection of stimulus involved in sensory perception (GO:0050906; FDR=$1.1x10^{-2}$). These findings are similar to the findings from earlier cattle CNV studies (Hou et al. 2012; Sasaki et al. 2016).

Table 2.1. GO enrichment results for different types of CNVR

Type of CNVRs	GO ID	GO Term	Size	Count	EXP *	Enrichment	FDR P-val
Deletion	GO:0007268	chemical synaptic transmission	278	22	8.3	2.65	0.126
Deletion	GO:0098916	anterograde trans-synaptic signaling	278	22	8.3	2.65	0.063
Deletion	GO:0099537	trans-synaptic signaling	279	22	8.33	2.64	0.044
Deletion	GO:0099536	synaptic signaling	279	22	8.33	2.64	0.033
Duplication	GO:0002821	positive regulation of adaptive immune response	32	6	0.44	13.76	0.019
Duplication	GO:0050778	positive regulation of immune response	57	7	0.78	9.01	0.021
Duplication	GO:0048584	positive regulation of response to stimulus	75	7	1.02	6.85	0.053
Duplication	GO:0002250	adaptive immune response	108	9	1.47	6.11	0.018
Duplication	GO:0002252	immune effector process	104	8	1.42	5.64	0.049
Complex	GO:0050896	response to stimulus	1,718	45	16.63	2.71	0.000
Complex	GO:0006955	immune response	298	14	2.88	4.85	0.002
Complex	GO:0050906	detection of stimulus involved in sensory perception	477	16	4.62	3.47	0.011
Complex	GO:0042113	B cell activation	17	4	0.16	24.31	0.013
Complex	GO:0050907	detection of chemical stimulus involved in sensory perception	477	16	4.62	3.47	0.014
Complex	GO:0051606	detection of stimulus	501	16	4.85	3.3	0.015
Complex	GO:0002376	immune system process	322	12	3.12	3.85	0.025
Complex	GO:0050853	B cell receptor signaling pathway	23	4	0.22	17.97	0.027

2.3.3. Population genetics of CNVRs

Population genetics analyses provide a framework to understand genetic variation seen in specific (cattle) populations. Understanding general properties of genetic variants is important, but further characterization of specific variants of interest can bring insights in recent adaptation and genome biology (Conrad and Hurles 2007). Although SNPs have been extensively used in characterizing various cattle populations (The Bovine Hapmap Consortium 2009), we explored the population genetic properties of CNVRs.

We focused our analyses on HOL (n=315) and JER (n=107) animals, derived from distinct origins and with a different breed formation history (Welch 1940). First, we coded the genotypes of our bi-allelic CNVRs (n=1,154 for HOL; n=700 for JER) as "+/+", "+/-", and "-/-". The CNVR allele frequency was classified as rare (MAF<0.01), intermediate (0.01≤MAF<0.05) and common (0.05≤MAF). In HOL, the allele frequency ranged from 0.002 to 0.29, and 5%, 13%, and 82% of the 1,154 CNVRs were categorized as common, intermediate, and rare CNVRs, respectively. For the JER population, allele frequency ranged from 0.005 to 0.37, and 11%, 20%, and 69% of the 700 CNVRs were categorized as common, intermediate, and rare CNVRs, respectively.

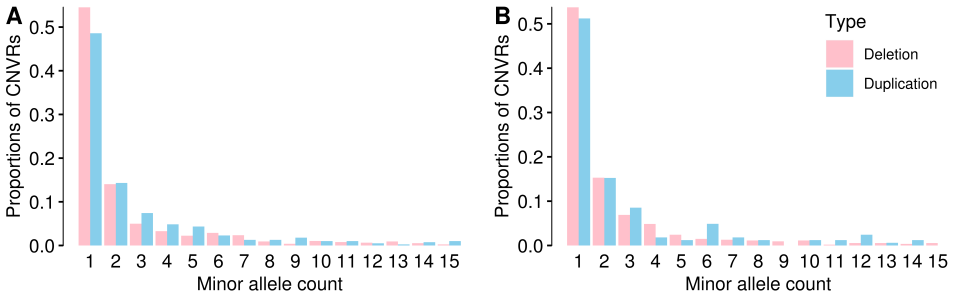


Figure 2.3. Site frequency spectrum of CNVRs. Site frequency spectra of CNVRs in HOL (**A**) and JER (**B**) population. Deletion CNVRs (pink) and duplication CNVRs (blue) are shown separately. Deletions tend to be enriched for rare CNVRs, whereas duplications tend to be enriched in common variants.

We constructed site frequency spectra of CNVRs for HOL and JER separately (Figure 2.3). For both populations, we observed that deletions and duplications have slightly different spectra, where deletions were more skewed towards rare CNVs, whereas duplications were observed relatively more frequent than deletions in each MAF class. We further explored the allele frequencies by applying Wright's fixation index (Fst; Wright 1950) to characterize population structure (Jakobsson et al. 2008) and detect loci that underwent selection(The International HapMap Consortium 2005), as done in Xue *et al.* (2008). Given that HOL and JER have distinctive origins and breed formation history (Welch 1940), we hypothesized that Fst on their CNVRs can reveal regions that underwent recent population differentiation.

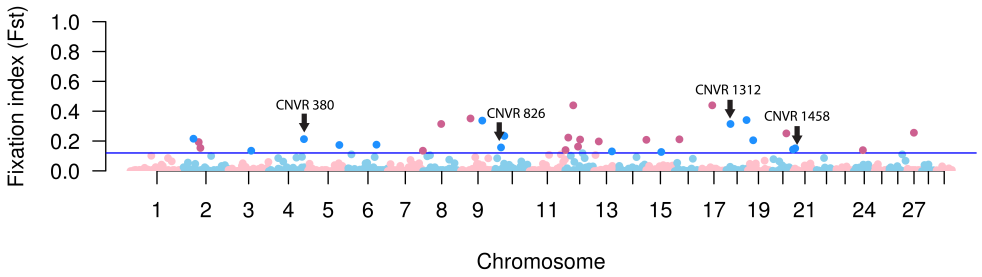


Figure 2.4. Manhattan plot for population fixation index (Fst) of CNVRs between HOL and JER. Population fixation index (Fst) of bi-allelic CNVRs between HOL and JER is shown in a Manhattan plot. 17 intergenic CNVRs (magenta) and 15 genic CNVRs (dark blue) were above the suggestive threshold (0.12; Fst>mean + 3 S.D.). CNVRs containing candidate genes are marked with arrows.

The Fst distribution followed an exponential decay pattern, as expected, underlining that majority of CNVRs have values close to 0, whereas only a few outliers (~3%) that are potentially under positive selection reached high Fst values (Supplementary Fig. 3). We identified 32 highly diverged CNVRs (Fst>mean + 3 S.D.) of which 15 are genic and 17 are intergenic (Figure 2.4 and Table S6). Among the 17 intergenic CNVRs with high population differentiation (Fst=0.12-0.44), 7 CNVRs had regulatory elements such as lncRNA and snoRNA within ~300 kb from the CNVRs. Among the genic CNVRs, CNVR 380 (Fst=0.21; duplication), which is more frequent in JER (MAF = 0.24) than in HOL (MAF=0.04), contains three genes, *CLEC5A* (Wade et al. 2014), *TAR2R38* (Destito et al. 2014), and *MGAM*. The known functions of these genes include abnormal eating behaviour, bitter taste perception, and the synthesis of maltase glucoamylase, a starch digestive enzyme. Furthermore, CNVR 826, 1312, and 1458 overlap with genes that are known to regulate body size: *LRRC49* (Dickinson et al. 2016), *CA5A* (Smith et al. 2018), and *ADAMTS17* (Bouwman et al. 2018; Frischknecht et al. 2016; Allen et al. 2010), respectively. Interestingly, these CNVRs are duplications and have a high allele frequency in JER (MAF=0.08–0.37), and a low allele frequency in HOL (MAF=0–0.06).

Subsequently, we calculated Vst statistic, which is a widely used statistic in CNV studies (Strillacci et al. 2018; Bickhart et al. 2016). This statistic is analogous to Fst, but using LRR values instead of allele frequencies (Redon et al. 2006). The Vst statistic ranges between 0 and 1, where 1 indicates population differentiation. To strengthen our confidence in the high Fst outlier regions we compared Fst and Vst statistics. Firstly, we calculated Vst for 1,464 CNVRs where Fst values are available. The Pearson correlation coefficient between Fst and Vst was low (0.22), and many selection candidate CNVRs that were found privately in Vst were either driven by rare CNVRs (less than 5 copies), or with a small number of SNPs (the numbers of average SNPs for top 20 Vst CNVRs and Fst CNVRs was 3.7 and 20.7 respectively; Supplementary Fig. 4 A-C). To correct for this, we removed CNVRs with less than 5 CNVs are called from either HOL or JER population (n=1,154 CNVRs). We observed that this filtering removed outlier CNVRs that were private to Vst, that were consisting of a small number of SNPs. After this filter, the 32 high Fst CNVRs were kept and the correlation coefficient was 0.52 (n=310 CNVRs; Supplementary Fig. 4 D-F). Also, CNVR 1458 which overlaps with

ADAMTS17, showed a high Vst of 0.17 (mean Vst mean=0.03, Vst S.D.=0.04). Furthermore, when the copy number filter was applied to both populations, and therefore both HOL and JER had more than five copies of CNVs at each CNVRs (n=44), the correlation coefficient increased to 0.81 (Supplementary Fig. 5).

2.3.4. Linkage disequilibrium of CNVRs

There has been a large number of genome-wide associations (GWAS) performed using SNPs in livestock species, aiming to unravel genomic regions related to phenotypes of interest (Sharma et al. 2015). This approach exploits a large number of tagging SNPs that are in sufficient LD with causal variants. Under this framework, genetic variation caused by the causal variants is captured by the tagging SNPs, without knowing the exact causal variants. Thus, the genome-wide level of LD between SNP markers and causal variants is an important foundation of GWAS (Visscher et al. 2017).

We showed that CNVRs overlap with genes more often than would be expected by chance, and that CNVs are thus likely to have an influence on phenotypes. The important follow-up question is whether the variations from CNVs are already captured by SNPs typed on commercial arrays, which are commonly used in livestock breeding programmes. We, therefore investigated pairwise LD between bi-allelic CNVRs and neighbouring SNPs on the BovineHD SNP chip. We observed generally low r^2, close to zero, regardless of the distance between CNVRs and SNPs (results not shown). Subsequently, we categorized CNVRs by their allele frequency and type to investigate whether these factors influence the degree of LD. Common CNVRs have markedly higher LD (r^2=~0.1 for deletion CNVRs at ~10-kb distance), compared to other CNVR categories (Supplementary Fig. 6). As common CNVRs had higher LD than the rest, we compared the LD of common CNVRs with the LD of SNPs in the same MAF range ($0.05 \leq MAF < 0.29$ for HOL and $0.05 \leq MAF < 0.37$ for JER).

We observed distinct LD decay patterns between the CNVR-SNP pairs and SNP-SNP pairs (Figure 5A and 5B). SNP-SNP LD follows a typical LD decay pattern where strong LD is observed with SNPs in vicinity and gradual decline as the distance increases, whereas CNVR-SNP LD does not follow this pattern. Also, compared to the CNVR-SNP LD (r^2=~0.1 at ~10 kb distance), the frequency matching SNP-SNP LD was stronger (r^2=~0.5 at ~10 kb distance). Afterwards, we used another metric to assess LD: taggability. Taggability is the maximum r^2 among the r^2 values that are obtained from a variant of interest and SNP pairs. We calculated taggability for SNP-SNP pairs and CNVR-SNP pairs. For the CNVR-SNP pairs, we considered common deletion CNVRs only, as they showed the highest LD in the previous analyses. Then, mean taggability for each MAF class (bin size=0.05) was plotted (Figure 5C and 5D).The mean taggability of common deletion CNVRs is low (<0.1) when MAF is below 0.05, and it increases as MAF increases. The SNP mean taggability follows the same pattern as shown in common deletion CNVRs. However, in spite of the similar pattern, common deletion CNVRs taggability is below the level of the SNP taggability. This shows that there is a gap in SNP taggability and CNVR taggability.

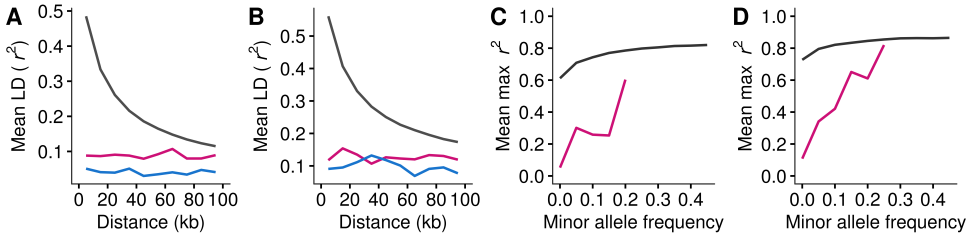


Figure 2.5. Linkage disequilibrium properties of CNVRs. Average strength of linkage disequilibrium (mean r^2) as a function of distance from a SNP is shown for HOL (**A**) and JER (**B**). Common CNVRs (0.05≤MAF) were used for the calculation; common deletion CNVRs (magenta) and common duplication CNVRs (blue) are shown together with common SNPs (black) for comparison. Taggability for HOL (**C**) and JER (**D**) was expressed as ratio of variants in high LD (r^2>0.8) with SNPs within 100-kb distance. Common deletion CNVRs (magenta) and common SNPs (black) are shown in the figure. Illumina BovineHD Genotyping BeadChip SNP set was used for the LD calculation.

2.3.5. Interesting CNVR

A large number of QTLs has been identified from various GWAS on a wide range of traits. As most GWAS have been done using SNP markers, chances are that genetic variation caused by a CNV could have been captured by QTLs that are in a high-to-perfect LD (r^2=~1) with the CNV. Hence, inspecting CNVRs that are in high LD with QTLs is a preliminary step to identify potentially causal CNVs. To identify candidate causal CNVs, we subset the CNVR-QTL pairs, from the total CNVR-SNP pairs, based on the QTL information from the animal QTLdb (Hu et al. 2016). We then subset the CNVR-QTL pairs further based on r^2, and kept high LD CNVR-QTL pairs only.

In total ~100,000 bovine QTLs for various traits have been reported in the animal QTL database, and we identified 2,519 QTLs to be paired with 679 CNVRs within a distance of 100 kb in the HOL population. Among these, CNVR 547 (BTA6:84,395,081-84,428,819, deletion, MAF=0.24) had the highest LD with 13 QTLs (average r^2=0.59; max r^2=0.74). The 13 QTLs were associated with casein proteins, which constitute four out of six bovine milk proteins. The four genes coding for the casein proteins are located in the so called casein cluster, which is ~1 Mb distant region from CNVR 547 (BTA6:85.4-85.6 Mb). Given the degree of LD for CNVR 547 and the QTLs that is lower than perfect linkage, it is unlikely that the CNVR 547 is the causal variant for the casein protein traits. Nevertheless, CNVR 547 was an interesting variant as it was private to in HOL population with high MAF (0.24), and was close to the casein cluster that are highly relevant for dairy production.

Assuming that CNVR 547 is not the causal variant for the casein traits, a possible explanation for the high MAF can be selective sweeps. Selective sweeps increase allele frequencies of neutral variants that are in LD with the selection target variant, which in this case probably is the casein cluster. Two studies of Holstein populations support this hypothesis. Firstly, one selective sweep study in a German Holstein population revealed an extended range of LD in haplotypes that contain the casein cluster (Qanbari et al. 2010). Secondly, GWA study on casein traits in

a Danish Holstein population identified a broad GWAS peak (BTA6:60-100 Mb) that contains the casein cluster (Buitenhuis et al. 2016). The broad GWAS peak also indicate high LD in this regions, that matched with the findings from Qanbari *et al.* (Qanbari et al. 2010)

Another explanation for the high MAF of CNVR 547 might be the direct selection on the variant itself. For instance, CNVR 547 overlaps with the *UGT2B4* gene, which is involved in the detoxification pathway of exogenous compounds (Barre et al. 2007). To see whether CNVR 547 overlaps with regulatory elements, besides overlapping with the upstream region of the *UGT2B4* gene directly, we called promoters and enhancers from ChipSeq data from Villar *et al.* (Villar et al. 2015). CNVR 547 overlaps not only with the upstream (a start codon and the first two exons), but also with the enhancer of *UGT2B4* (BTA6: 84,413,246-84,413,740), and is thus likely to disrupt the function of the *UGT2B4* gene. To summarize, our analyses imply that a high MAF of CNVR 547 might be due the selective sweep in the casein cluster or the consequence of direct selection on CNVR 547 itself due to the functional impact of the overlap with *UGT2B4* and its enhancer. Nonetheless, we cannot exclude drift as a possible driver for the high allele frequency of the CNVR 547.

2.4. Discussion

In this study, we discovered CNVs using bovine high density SNP array. Using CNVRs that are constructed using the CNVs, we reported the functional impact and population genetic features of the CNVRs. They are further discussed below.

2.4.1. CNV discovery in the genome build ARS-UCD1.2

We observed different CNV discovery results between UMD3.1 and ARD-UCD1.2. The different results were to be expected, given the different sequencing platforms used for the assemblies. Long-read sequencing platforms are shown to perform better in retrieving repeat regions, which is considered to be challenging in short-read sequencing (Nakano et al. 2017). Among others, the most intriguing difference was observed for the BTA12:70-77 MB region. Based on the changes in BovineHD SNPs between UMD3.1 and ARS-UCD1.2, we postulated that the two genomes assemblies differ in this regions largely. Subsequently, the changes in the genome assemblies led to different CNV discovery results. We, then, further postulated that this region (BTA12:70-77Mb in UMD 3.1) might contain repeated sequences, rather than the reported CNV, for two reasons. Firstly, the SNP density in this region is a quarter of the genome-wide average SNP density in UMD3.1 (71 SNPs/Mb and 292 SNPs/Mb, respectively; Supplementary Fig. 2). SNP probes in repeat regions can reduce specificity of hybridization, and hence are often filtered out during SNP probe selection (ArrayGen; Koltai and Weingarten-baror 2008; Lemoine et al. 2009), which can explain why some regions show a sharp decrease in SNP density. Secondly, SNP probes in segmental duplications (sequence identity $> 90\%$) can induce confounded deletion calls due to cross-hybridization of paralogous sequences (Cooper et al. 2008). Our data set based on UMD3.1 was indeed enriched for a large number of deletion calls in this region. We regard this large difference as evidence underlining the importance of the quality of the reference genomes and the impact this has on CNV calling results.

2.4.2. Functional impact of CNVRs

In our functional impact analyses, we showed that the overlap between genes and CNVRs is higher than the overlap in a neutral scenario. This finding is in line with human and rat CNV studies, which showed that the overlap between CNVs and genes is significantly higher than expected by chance (Nguyen et al. 2006; Guryev et al. 2008; Cooper et al. 2007). These studies were based on medium-to-large size human CNVs, and rat CNVs were found from exome arrays (CNV length ranged between 5-kb and 256-kb). However, more recent studies, based on a finer resolution of CNVs, concluded that CNVs are biased away from genes and functional elements (Conrad et al. 2010; Sudmant et al. 2015; Mesbah-Uddin et al. 2017; Boussaha et al. 2015).

Also, we observed that MAF have impact on different types of overlap between genes and CNVRs. In our findings, common deletion CNVRs were biased away from the genic regions, yet the common duplication CNVRs were skewed toward the genic part. This was contradicting with findings from another study, which showed both common deletions and duplications are skewed away from genic part (Conrad et al. 2010).

We assume that these conflicting findings might arise from a curation of SNP array based CNVs in our dataset, which is affected by an ascertainment bias. An ascertainment bias of SNPs in commercial arrays can introduce a two-fold bias in CNV discovery. Firstly, the SNP density of a given array will constrain the size of CNVs that can be discovered. Secondly, SNP probes are designed such that complex regions, such as segmental duplications (SD), are under-represented (McCarroll et al. 2008; Cooper et al. 2008). The SNP density of BovineHD BeadChip array in unique regions is 292 probes/Mb, whereas it drops to 95 probes/Mb in SD regions, showing a 67.5% reduction. Based on this, we speculate that the uneven genome-wide SNP coverage might introduce a systematic bias in CNV discovery. Taken together, the studies that focused on mid-sized CNVs (Guryev et al. 2008; Cooper et al. 2007; Nguyen et al. 2006) are in line with our findings, whereas studies based on tiling oligonucleotide microarrays (Conrad et al. 2010) and whole genome sequencing data (Mesbah-Uddin et al. 2017; Sudmant et al. 2015), which can provide rather complete genome-wide coverage with a various size range of CNVs (Alkan et al. 2011), show different results.

Furthermore, another layer of bias in CNV discovery using SNP array is that discovery of duplication is less sensitive than that of deletions. Consequently, most small CNVs are overwhelmingly deletions, whereas duplications usually are discovered based on relatively large number of SNPs than deletion, which makes duplications longer than deletions (Alkan et al. 2011). Indeed, our CNV discovery we found two folds more deletions than duplications (9,171 vs. 4,101), and the mean length of duplication was longer (44.2-kb vs 74.6-kb). This deletion-duplication bias might explain why common duplication CNVRs in our dataset are more likely to affect genic region compared to the rare duplication CNVRs, whereas it was the opposite in a study mentioned above (Conrad et al. 2010).

The need to re-evaluate the functional impact of CNVs, as CNV detection resolution became

finer, along with the advancement in assay technologies and detection algorithms, was already pointed out (McCarroll et al. 2008). Moreover, a recent study exploiting long-read sequencing data detected 237% and 34% more insertions and deletions, respectively, compared to known variants detected from short-read sequencing data (Audano et al. 2019). Taken together, the CNVs discovered in our dataset (>1kb) were shown to be biased towards genic regions. However, we stress the need of re-visiting CNVs with finer resolution and unbiased genome-wide coverage, to fully comprehend their functional consequences in cattle genomes.

2.4.3. Population genetics of CNVRs

We explored the population genetics of CNVRs by examining the site frequency spectra and Fst. The frequency spectra differed for deletion CNVRs and duplication CNVRs. Given the skewed number of rare deletions and common duplications, we corroborate that deletions might be under stronger purifying selection. Nevertheless, as explained earlier, inherent bias in CNVs from SNP array (deletion discovery is more sensitive than duplication discovery), we cannot entirely exclude a possibility that the differed frequency spectra might be an artefact.

Furthermore, we used Fst to identify CNVRs that are highly diverged. Among the 32 CNVRs that pass the threshold, of which 7 intergenic CNVRs had regulatory elements in neighbouring regions. This finding underlines that potential recent positive selection probably acted on regulatory elements. Among the 17 genic CNVRs, we identified CNVRs that overlap with interesting candidate genes. The CNVR 380 overlaps with *CLEC5A*, *TAR2R38* and *MGAM* gene that are related to taste perception and a digestion enzyme, maltase. One selective sweep study revealed that a region containing *TAR2R38* and *MGAM* is highly diverged between dogs and wolves. Dogs produce a longer form of maltase than wolves, due to a 2bp deletion that disrupts the stop codon, and the same mutation was also seen in herbivore species (rabbits and cows) (Axelsson et al. 2013). The longer form of maltase might be the consequence of adaptive evolution in response to a starch-rich diet during dog domestication. Given that the partial duplication of *MGAM* can lead to increased length of maltase, a high duplication frequency seen in the JER population (MAF=0.24) might be a hint that feed related adaptive evolution occurred in the JER population. Also, we identified genes related to body size (*LRRC49* in CNVR 826, *CA5A* in CNVR 1312, and *ADAMTS17* in CNVR 1458). Among these genes, *ADAMTS17* has been reported as one of the height determining genes in various species, such as cattle, horse, and human (Allen et al. 2010; Frischknecht et al. 2016; Bouwman et al. 2018). Also, a deletion variant overlapping with *ADAMTS17* was shown to be highly diverged between HOL and JER in a previous study (Mesbah-Uddin et al. 2017). Given that CNVR 1452 we found is a duplication locus, it might be a different mutation than the one found by Mesbah-Uddin *et al.* (Mesbah-Uddin et al. 2017). Nonetheless, our and the previous findings revealed that CNVs overlapping with *ADAMTS17* gene to be diverged between HOL and JER. This supports *ADAMTS17* gene as a candidate gene that can explain the phenotypic differences (i.e. body size) between the two breeds.

Additionally, we used Vst analyses to confirm the selection candidate CNVRs based on Fst analyses. The preliminary results from Vst statistic from 1,464 bialleleic CNVRs showed that

extreme Vst could be obtained from very rare CNVs (less than 5 CNVs observed) and short-sized variants. We consider correcting for these factors in analysing Vst statistic is crucial, as it could reduce falsely derived selection signal from false positive singletons (Upadhyay et al. 2017). We have seen that overall concordance between Vst and Fst was 0.52, when rare CNVRs (number of CNVs <5) were filtered out in either of the populations. Furthermore, when rare CNVs were filtered for both of the populations, which means CNVRs were present in both populations with more than 5 copies, the correlation coefficient was 0.81. This number is slightly lower than 0.9, which was shown in human CNV study. These findings underline high concordance of Fst and Vst when CNVRs are present in both populations with sufficient MAF. Thus, although we could obtain Vst confirmation for CNVR 1458, which overlaps with *ADAMTA17*, we could not obtain such confirmation for CNVRs that are at low MAF in either of the two populations.

2.4.4. Linkage disequilibrium of CNVRs

To summarize our findings on LD properties of CNVRs, CNVRs are generally in low LD with SNPs, and CNVR taggability is lower than SNP taggability, which indicates a taggability gap. However, findings on the taggability are conflicting. Although some studies reported high CNV taggability (Conrad et al. 2010; Mills et al. 2011; Mccarroll et al. 2006; McCarroll et al. 2008; Hinds et al. 2006), as high as SNP-SNP taggability, some studies reported low CNV taggability (Kato et al. 2010; Redon et al. 2006; Locke et al. 2006; McCarroll et al. 2008; Cooper et al. 2008) as shown in our results.

The taggability gap can be explained by three factors. Firstly, LD is affected by allele frequency. High LD can be obtained when the allele frequencies of the two loci match (Wray 2005). Van Binsbergen *et al.* (2014) empirically showed that SNP-SNP pairs with small MAF difference (<0.05) had high predicted LD (r^2 >0.8) using WGS data (Binsbergen et al. 2014). In our dataset, the majority of CNVRs is at low allele frequency (88% and 95% of CNVRs in JER and HOL are at MAF≤0.05), whereas BovineHD SNPs are biased away from rare MAF (10% of SNPs are at MAF≤0.05). Thus, the allele frequencies of CNVRs and SNPs were largely unmatched, which can be explain the low LD. Secondly, deletions are tagged better than duplications. Even studies that found high taggability for common deletions, only found relatively poor taggability for duplications (Conrad et al. 2010; Mills et al. 2011; Sudmant et al. 2015). This might be due to dispersal duplications (Gondo et al. 1993), which relocate the duplicated segment of DNA in a different haplotype background than the "parental locus" (Schrider and Hahn 2010). Thus, the LD of duplications might be lower than that of deletions. Lastly, local SNP density can influence the level of LD.

Redon *et al.* (Redon et al. 2006) and Locke *et al.* (Locke et al. 2006) suggested that a paucity of SNPs in repeat-rich regions to serve as potential tags can be an explanation for the taggability gap. Indeed, Cooper *et al.* (Cooper et al. 2008) and McCarroll *et al.* (McCarroll et al. 2008) used different SNP sets in their CNV LD analyses. The first SNP set was HapMap phase 2 SNP set, which is known to cover the whole genome uniformly (~3.1M probes). Next to this, they used SNP sets obtained from commercial SNP arrays, which have uneven SNP density

along the genome (550K ~ 1M probes). They found that ~80 % of CNVs are in high LD (r^2> 0.8) when HapMap phase 2 SNP set was used, whereas ~50 % of CNVs were in high LD with the commercial array SNP sets.

Based on our and previous findings, we postulate that LD between common deletion CNVRs and SNPs is not necessarily low. However, we could not obtain high LD with our CNVRs, because our CNVRs were skewed towards rare MAF. The MAF difference between CNVRs and SNPs can explain lower LD shown in rare CNVRs, compared to common CNVRs. However, as shown in another study (Upadhyay et al. 2017), singletons found from PennCNV software could be false positives, which could lead to low LD as well. Thus, we could not exclude the possibility that the low LD in rare CNVRs was partly caused by false positive singleton CNVs driving low LD. Also, BovineHD SNPs were underrepresented in SD regions, where SNP probe design is difficult due to high sequence identity. Deprivation of SNPs in these regions probably led to lack of markers that can serve as tagging markers. Follow-up research using a SNP set that uniformly covers the whole bovine genome might unravel more complete LD properties of CNVs.

2.4.5. Interesting CNVR

Furthermore, in search of CNVRs that are causal variants of traits, we investigated CNVRs that are in high LD with known QTLs. CNVR 547 was shown to be in high LD with casein QTLs, although it was below perfect linkage, thus unlikely to be the causal variant. However, this opened up an interesting avenue to see the CNVR 547 in light of LD and selection. We proposed three possible explanation for CNVR 547 to reach high MAF:1) selective sweeps, 2) direct selection on CNVR 547 that affects the enhancer of *UGT2B4*, and 3) drift. Although we could not unravel how CNVR 547 has reached high MAF in the current study, we deem it as an interesting case, which a CNVR can be understood in standard population genetics theories, such as selective sweeps and drift. Also, we had a limited number of CNVRs obtaining high LD with QTLs. This was partially due to because most CNVRs were rare, thus predisposed to have low LD. Therefore, re-visiting CNVR-QTL pairs, based on CNVs that are detected from a different platform (i.e. WGS) might reveal more candidate CNVs that might be the underlying causal variants of traits.

2.5. Conclusions

In this study, we discovered CNVs in bovine genomes and explored their functional impact and population genetics features. Using commercial high-density SNP arrays, we identified 14,272 CNVs, that built 1,755 CNVRs (cover ~2.8% of the bovine autosomes), and the CNVRs were further used as genetic loci this study. In the functional impact analyses, we showed that CNVRs are likely to have functional impact based on their overlap with genes. Also, we investigated CNVRs in light of population genetics. We identified 32 highly differentiated CNVRs between HOL and JER based on Fst values. Two of the highly diverged CNVRs overlapped with the *ADAMTS17* gene and *MGAM* gene, which are involved in body size and starch digestion enzyme, respectively. In the LD analyses, CNVR-SNP LD was lower than SNP-SNP LD,

mainly due to low MAF in CNVRs and uneven SNP density.

These findings together impose several implications for future CNV studies. The first implication is about the functional impact of CNVs. SNP based GWAS is a commonly used design to find functional SNPs that are associated with traits. Given the low CNVR-SNP LD, SNP based GWAS are unlikely to detect CNVRs with functional impact. Consequently, GWA studies that associate CNVRs and traits directly can add valuable insights into understanding economically important traits. Secondly, the low CNVR-SNP LD implies that the majority of CNVRs in our study is probably not captured in the current genomic prediction, where SNP markers are used. Thus, we underline the importance of follow-up studies on investigating methods to include CNVs in genomic prediction and evaluating the usefulness of CNVs in improving the accuracy of genomic prediction.

2.6. Methods

2.6.1. Animal samples and ethics

The study population consisted of two dairy cattle breeds, 331 Holstein Friesian (HOL), 115 Jersey (JER), as well as 29 crossbreds of HOL and JER. Among these, 18 HOL and 17 JER animals were cows and the rest were bulls. All samples were genotyped using an Illumina BovineHD Genotyping BeadChip (Illumina, San Diego, CA, USA), which contains 777,692 SNPs. All of these genotypes are owned by commercial dairy breeding company CRV (Arnhem, the Netherlands). The Genotype data was provided by CRV.

2.6.2. Identification of CNVs

We identified CNVs using PennCNV software (Wang et al. 2007) which exploits a Hidden Markov Model algorithm. For each individual, log R ratio (LRR) and B allele frequency (BAF) per SNP were inferred using the Illumina Genome Studio software package (Illumina, San Diego, CA, USA). Autosomal SNPs of BovineHD Genotyping BeadChip (n=735,965; Illumina, San Diego, CA, USA) were used, and their positions were based on the genome assembly ARS-UCD1.2. We called CNVs in 29 Bovine autosomes. The waviness in LRR values caused by GC contents were adjusted afterwards. We chose PennCNV software, together with BovineHD Genotyping BeadChip, as this method showed high confirmation based on qPCR validation in a previous study (91.7% for CNVs found in multiple animals and 40% for singleton CNVs) (Upadhyay et al. 2017).

After the initial CNV detection, poor quality individuals (n=13) were filtered out with the default criteria suggested by the developer of the PennCNV software (LRR standard deviation> 0.30, BAF standard deviation>0.001 and Waviness factor>0.05). Afterwards, the distribution of the number of CNVs per individual was inspected using QQ plots (Supplementary Fig. 7). The distribution was continuous until 100, and individuals with more than 100 CNVs largely deviated from the distribution (n=10). The same filter on the distribution of the total length of CNVs per individual was applied and identified outlier samples (n=11). These two filter steps identified 11 outlier individuals (among the 11 outlier animals identified in the second

filter, 10 were identified as outliers in the first filter), and subsequently these individuals were removed to prevent the introduction of a large number of possible false positive CNVs. Lastly, we merged two adjacent CNVs that have the same copy number state, when the gap between the two CNVs was less than 10% of the total length, using the clean_cnv.pl script provided by PennCNV software, which resulted in 451 individuals with 14,272 CNVs in the combined dataset of the two breeds.

2.6.3. Constructing CNVRs

The CNVs were aggregated into CNV regions (CNVR) based on 1 bp overlap, following Redon *et al.* (2006) (Redon et al. 2006). CNV regions that exclusively contain deletions or duplications were classified as deletion CNVRs and duplication CNVRs and treated as bi-allelic loci. In case of CNVRs that consisted of both deletions and duplications, we defined them as complex CNVRs. The CNVRs were compared together with SD and SNP density in . The SDs detected by Feng *et al.*(2017) based on UMD3.1 were remapped to ARS-UCD1.2 using NCBI Genome Remapping Service. Afterwards, the density of SDs and SNPs were calculate for 5 Mb bin using BEDtools (Quinlan and Hall 2010). Circos software (Krzywinski et al. 2009) was used to visualize CNVRs, SD density, and SNP density.

2.6.4. Assessment of CNV discovery results

We repeated the same CNV calling steps using 735,293 autosomal SNPs based on the genome assembly UMD3.1. After the initial CNV detection, the same quality control filters were applied as explained above. The default criteria filtered out 18 individuals, and another 11 outliers detected from QQ plots of the number of CNV per individual and the total length of CNV per individual were removed (Supplementary Fig. 7). Subsequently, split CNVs that have small gaps in between were merged as described for ARS-UCD1.2. From the 447 individuals that passed the quality control criteria, 24,264 CNVs were called, and 1,866 CNVRs were constructed as explained above. Finally, we compared the CNVs and CNVRs between the two different genome assemblies, UMD3.1 and ARS-UCD1.2, in terms of number and length.

2.6.5. Functional impact of CNVRs

The CNVRs were overlapped with gene annotations using Ensembl Variant Effect Predictor (McLaren et al. 2016) (Cow release 95) to explore their functional impact. Subsequently, CNVRs were classified depending on their functional impact, as done in Conrad *et al.*(2010). First, we identified intergenic CNVRs, which did not overlap with genes, and genic CNVRs which overlapped with genes. Among the genic CNVRs, ones containing a complete gene or genes are classified as "whole gene". Genic CNVRs that overlapped with some part of genes were further classified as "intronic", when CNVRs overlapped with introns exclusively; as "stop codon", when CNVRs overlapped with stop codon; as "promoter region", when CNVRs included promoter region (500-bp from transcription start site). The remaining CNVRs that overlap with an exon or exon(s) and intron(s) were considered as "exonic". In the case of CNVRs overlapping with more than one gene, and thus having more than one category assigned, (i.e. that contains a complete gene and also a promoter region of another gene), we assigned one

unique category in the following order: 1) whole gene, 2) stop codon, 3) promoter region. With the steps explained above, each CNVR was assigned a unique category. Then, we investigated whether the functional impact classes were influenced by type of CNVRs (1,125 deletion, 513 duplication, 117 complex CNVRs). Also, the influence of allele frequency on the functional impact classes were analysed and the allele frequency classes were defined as common (MAF ≥0.05 in any population, 56 CNVRs), intermediate (0.1>MAF≥0.01, 267 CNVRs), and rare (MAF<0.01 in HOL and JER, 115 CNVRs). To see whether the functional impact category differs significantly depending on type of CNVRs and MAF classes, Pearson's Chi-square tests were performed. Afterwards, CNVRs were classified depending and type and allele frequency in HOL and JER separately and the overlap with functional classes were analysed. Afterwards, we performed permutation tests to understand whether the observed overlap between CNVRs and a genomic feature is high or low, compared to random genomic regions. The permutation tests were performed with the R package "regioneR" (Gel et al. 2016). We generated a random set of regions in the genome, with the same number and length of genomic features, and did this 1,000 times. For each permutation, the number of overlaps between random CNVRs and the genomic features was recorded and then used to estimate the expected number of observations. The observed and the expected numbers of overlaps were then tested for significance (z-test). Subsequently, the PANTHER classification system (Thomas et al. 2003) was used to perform gene ontology enrichment tests for the genes that overlapped with CNVRs. All known bovine genes (Ensembl release 95) were used as a reference set to test whether the CNVR overlapping genes were enriched for or deprived of a specific biological process, cellular composition, and molecular function, with False Discovery Rate correction (α<0.05) for multiple tests.

2.6.6. Population genetics of CNVRs

We explored bi-allelic CNVRs in HOL and JER in light of population genetics. We genotyped bi-allelic CNVRs in HOL and JER into "+/+", "+/-", and "-/-", following McCarroll *et al.*(2006). These genotypes were used to calculate the allele frequency of each CNVR locus. Subsequently, we constructed site frequency spectra in HOL and JER to understand selection pressure on CNVRs. Wright's population differentiation index (Fst; Wright 1950) was used to investigate recent divergent selection in HOL and JER populations. Fst was calculated for 1,471 bi-allelic CNVRs, using PLINK (version 1.9.; Purcell et al. 2007).

2.6.7. Linkage disequilibrium of CNV

We estimated the degree of LD between bi-allelic CNVRs and SNPs by calculating r^2 in the JER and HOL populations, respectively. To have a reference, we also estimated SNP-SNP LD, limited to SNPs with the same MAF range as common CNVRs (0.05<MAF<0.30 for JER and 0.05<MAF<0.24 for HOL). The SNPs inside the CNVRs were masked to prevent a bias introduced during the phasing step, as done in Conrad *et al.*(2010). SNPs with low minor allele frequency (MAF <0.001), with low call rates (<90%), or with deviations from the Hardy–Weinberg equilibrium ($P<1e^{-9}$) were removed. For CNVRs, the same filters were applied, except the call rate criteria. Phasing was done with Shapeit (Delaneau et al. 2012) and the r^2 values of CNVR-SNP pairs within a 100 kb distance were calculated in PLINK (version 1.9.; Purcell

et al. 2007). Afterward, QTLs that were shown to be significant in association studies were downloaded from Animal QTLdb (Hu et al. 2016) (release 37) and intersected with the CNVR-SNP pairs to see whether CNVRs are in high LD with known QTLs. To overlap the CNVR 547 and functional elements in bovine genomes, we used the data from Villar *et al.*(2015). We downloaded the ChipSeq data and aligned them to ARS-UCD1.2 using BWA-MEM(0.7.15) (Li 2013), and called the enhancers and promoters as explained in the original paper.

2.7. List of abbreviations

CNV: copy number variation;

CNVR: CNV region;

Fst: Fixation index;

HOL: Friesian Holstein;

JER: Jersey;

LD: Linkage disequilibrium;

MAF: Minor allele frequency;

QTL: quantitative trait loci;

SD: segmental duplications;

SNP: single nucleotide polymorphism

2.8. Declarations

2.8.1. Ethics approval and consent to participate

The data used for this study were collected as part of routine data recording for a commercial breeding program. Samples collected for DNA extraction were only used for the breeding program. Data recording and sample collection were conducted strictly in line with Dutch law on the protection of animals (Gezondheids en welzijnswet voor dieren).

2.8.2. Availability of data and material

The data that support the findings of this study are available from CRV B.V. (Arnhem, the Netherlands) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of CRV B.V.

2.8.3. Competing interests

The authors declare that this study received funding from CRV B.V., Cobb, Hendrix Genetics, and TopigsNorsvin. All funders were involved in the study design, CRV B.V. performed data collection and was involved in preparation of the manuscript. MG is a member of the editorial board (Associate Editor) of BMC Genomics journal. All authors declare that the results are presented in full and as such present no conflict of interest.

2.8.4. Funding

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. MB was financially supported by NWO grant 016.Veni.181.050.

2.8.5. Authors' contributions

All authors designed the study. YLL performed the statistical analyses and drafted the manuscript under supervision of AB and MB. YL, AB, MB, MG, and RV interpreted the results. EM contributed to data collection, conception of the study, and manuscript revisions. All authors participated in discussions. All authors read and approved the final manuscript.

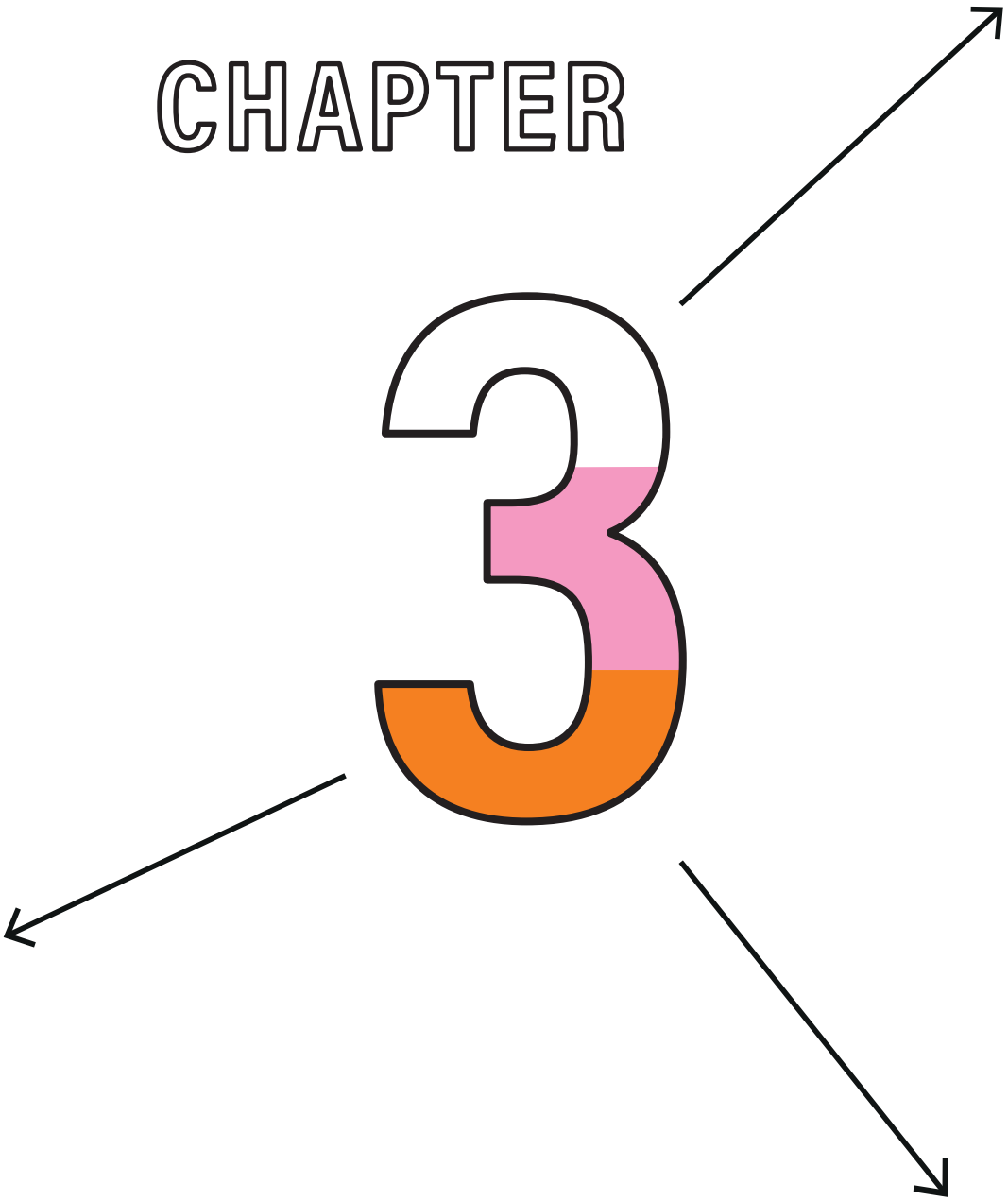
2.8.6. Acknowledgements

The authors acknowledge CRV B.V. (Arnhem, the Netherlands) for providing the genotype data. The use of the HPC cluster has been made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR)

2.9. Supplementary figures and tables

Supplementary figures and tables are available at the online version of the article (doi: 10.1186/s12864-020-6496-1).

CHAPTER



High-resolution structural variation catalogue in deeply sequenced cattle genomes

Young-Lim Lee[1*], Mirte Bosse[1], Haruko Takeda[2], Gabriel Costa Monteiro Moreira[2], Latifa Karim[3], Tom Druet[2], Claire Oget-Ebrad[2], Wouter Coppieters[2,3], Roel F. Veerkamp[1], Martien A. M. Groenen[1], Michel Georges[2], Aniek C. Bouwman[1], Carole Charlier[2]

[1]Wageningen University & Research, Animal Breeding and Genomics, Wageningen, the Netherlands
[2] Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Liège, Belgium [3] GIGA Genomics Platform, GIGA Institute, University of Liège, Liège, Belgium

Submitted

3.1. Abstract

Structural variations (SVs) are large segments that differ between genomes, such as deletions, duplications, insertions, inversions and translocations. Genomic revolution enabled the discovery of sub-microscopic SVs via array and whole-genome sequencing (WGS) data, paving the way to unravelling the functional impact of SVs. Recent large-scale human expression QTL (eQTL) mapping study underlined that SVs play a disproportionally large role in altering gene expression, underlining the importance of including SVs in genetic analyses. However, incorporating SVs in routine genetic analyses has been hindered due to inaccurate genotyping, which is particularly compounded in low depth WGS data. This study generated a high-quality bovine SV catalogue, exploiting 266 deeply sequenced genomes obtained from a dairy cattle family cohort, including 127 trios (mean sequencing coverage=26X). We curated 13,925 SVs segregating in the population, consisting of 12,200 deletions, 1,531 duplications, 22 multi-allelic CNVs, and 172 retrogene insertions. Of these, we validated a subset of copy number variants (CNVs; n=372) utilising a direct genotyping approach in an independent cohort, underlining that at least 80% of the CNVs are true variants, segregating in the population. Among genic sequence disrupting SVs, we prioritised two likely high impact duplications, encompassing *ORM1* and *POPDC3*, respectively. Our liver eQTL mapping results on these genes revealed that the duplications are likely causal variants for the altered gene expression, confirming the functional importance of SVs. Although most of the accurately genotyped CNVs are tagged by SNPs ascertained in WGS data, most CNVs were not captured by SNPs obtained from a 50K genotyping array. Accordingly, many SVs may not be exploited in large scale genomic analyses relying on sparse density SNP data (e.g. 50K). The current SV catalogue is an invaluable resource for future genetics studies for SVs. Lastly, we stress that the effort to capture SVs accurately will be beneficial in exploiting the full spectrum of genetic variants.

3.2. Introduction

Structural variations (SVs) are genomic segments (>50-bp) for which the structure between genomes differs, and may include deletions, duplications, insertions, inversions, and translocations (Sudmant et al. 2015). Due to a larger affected portion of the genome, their phenotypic impact are assumed to be relatively larger compared to single nucleotide polymorphisms (SNPs) or small insertions and deletions (indels; Campbell and Eichler 2013; Alkan et al. 2011; Conrad et al. 2010). Following this idea, large-scale expression QTL (eQTL) studies in humans showed that SVs have a disproportionately larger contribution to altering gene expression than SNPs and indels (Chiang et al. 2017; Scott et al. 2021; Handsaker et al. 2015). Many functional SVs associated with various traits have been elucidated in humans (Weischenfeldt et al. 2013). Likewise, identifying functional SVs associated with economically important trait(s) has been a prime interest for animal breeders. Until now, catalogue of functional SVs reported in farm animals contain many deletions that often are associated with disease traits. In contrast, duplications often are associated with distinguishable coat colours and morphologies (e.g. breed defining traits), with few exceptions (Bickhart and Liu 2014; Clop et al. 2012).

Discovery and genotyping of genetic variants provide a foundation for performing genetic analyses. In recent decades, the genomics revolution enabled accurate detection and genotyping of millions of SNPs through arrays and whole-genome sequencing (WGS) technologies. However, unlike SNPs, detection and genotyping methodologies for structural variants (SVs) have been lagging behind (Huddleston and Eichler 2016). Array data is widely used in animal breeding and can detect unbalanced SVs, such as copy number variants (CNVs, a subset of SVs including deletions and duplications). Still, low resolutions and undefined breakpoints are considered major drawbacks of array-based methodologies (Alkan et al. 2011). Alternatively, short-read WGS data can be used to detect SVs, including CNVs and balanced SVs (e.g. inversions) at a finer resolution (Sudmant et al. 2015). Despite such advancement, WGS data with low sequencing depth (e.g. <10X) suffers from low sensitivity, unresolved breakpoints, and low genotyping accuracy (Huddleston and Eichler 2016; Sudmant et al. 2015; Alkan et al. 2011). These issues can be alleviated by (i) exploiting WGS data with higher sequencing depth (e.g. >30X), (ii) including family samples, and (iii) confirming the discovery results using orthogonal validation (e.g. long-read sequencing data; Huddleston and Eichler 2016). Furthermore, the choice of SV detection methods can affect the discovery results. SV detection tools scan WGS data for split reads (SR) and discordant read pairs (RP) clusters. In contrast, some detection tools measure read-depth (RD) changes relative to the depth of genome-wide diploid regions to determine the copy number variable regions. Recent benchmark studies showed that combining these principles outperforms detection methods solely relying on a single principle (e.g. generating less false calls) (Cameron et al. 2019).

A high-quality catalogue of SVs with high detection sensitivity, including a diverse event size, base-pair resolved breakpoints, and accurate genotyping can benefit genetic studies and accelerate the discovery of functional variants. Yet, until now, lack of suitable data sets hindered obtaining a high-quality SV catalogue in the Holstein Friesian (HF), a major dairy cattle breed (Britt et al. 2021). Absence of a high-quality SV catalogue has left some questions unanswered. Firstly, the potential for SVs for animal breeding is unknown. Whether a widely used 50K SNP genotyping array captures, genome-wide SVs remains to be investigated. Secondly, current SV catalogues based on genotyping arrays consist of large, breakpoint unresolved CNVs with spurious genotyping (Lee et al. 2020), and hence, estimating their functional and phenotypic impact has been limited.

This study aimed to generate a high-quality SV catalogue using deeply sequenced genomes obtained from a cattle family cohort (including 127 trios). We detected four different classes of SVs (deletions, duplications, multi-allelic copy number variants (mCNVs), retrogene insertions) based on a methodology exploiting both SR and RP signals in WGS data, with post hoc RD-based filtering. Furthermore, SVs were validated in an independent cohort of animals using a direct genotyping approach. Using a high-quality call set, we explored population genetics features of SVs and finally, we performed in-depth characterisation of putative high impact duplication events.

3.3. Results

3.3.1. CNV discovery

3.3.1.1. Initial discovery

We discovered SVs using short-read WGS data from 266 Dutch dairy cattle samples (mean coverage of 26X, min=15X, max=47X). Our pipeline found SVs in the individual samples based on SR and RP evidence. The number of discovered SVs per sample increased in relation to the sequencing depth, suggesting the absence of spurious technical bias and high quality underlying WGS data (Supplementary Figure 1). Aggregating the SVs found across all samples, we obtained 38,094 non-redundant SVs (17,826 deletions, 4,652 duplications, 1,811 inversions, 13,805 non-canonical SVs), for which the entire cohort was genotyped. Further analyses were focused on CNVs (simple biallelic deletions and duplications), mCNVs, and retrogene insertions. We applied preliminary filters removing spurious calls based on fold-coverage change in RD for CNVs (e.g. a 2-Mb duplication with no coverage increase was removed). After excluding spurious CNVs, 12,200 deletions and 1,531 duplications were retained. Hereafter, a total of 13,731 CNVs that passed the preliminary filters is referred to as "clean call set" (Figure 3.1). In addition, we catalogued 22 duplications as mCNVs based on their multi-modal RD distributions (Supplementary Figure 2). Together, these CNVs were in a size range between 50-bp and 424-kb (Supplementary Figure 3). During a manual inspection on false calls, we found clusters of false deletions occurring in intronic regions exclusively (Supplementary Figure 4). This indicates processed pseudogenes, which occur when processed mRNAs are reverse transcribed into DNA inserted into a different location in the genome (Ewing et al. 2013). By systematically re-evaluating such false deletions corresponding to the intronic region, we identified 172 source genes, where 169 retro-transposed sites were retrieved. Overall, a median number of 5,252 CNVs per genome was discovered (4,865 deletions and 387 duplications).

3.3.1.2. Differentiating Stringent vs. Lenient call sets

The 13,731 CNVs belonging to the clean call set were further scrutinized. Our pipeline assigned genotypes (GT) based on SR and RP evidence, and corresponding RD (Figure 3.1 A). When all samples are accurately genotyped and the RD reflects the true underlying GT, we expect to see a cloud of dots per GT and the distribution of RD per GT do not overlap. Alternatively, a dispersed cloud of dots within a GT indicates inaccurate genotyping, which do not reflect the true underlying GT (Figure 3.1 D). Hence, we divided the clean CNV call set into a "stringent" call set, with CNVs of which their RD corresponds unambiguously to the reads-based genotypes, and a "lenient" call set, with CNVs of which their RD does not always match with the reads-based genotypes.

The stringent call set consisted of 3,827 deletions and 184 duplications which contained accurately genotyped biallelic CNVs, mostly larger than 500-bp (Figure 3.2 A). On the contrary, CNVs in the lenient call set were often (i) small (<500-bp), hence the RD did not manifest a clear change depending on genotype, or (ii) incorrectly genotyped due to a complex local

genomic context (e.g. discordant RPs in repeat-rich regions lead to low mapping quality, thus were not taken into account in genotyping, see Figure 3.2 A), or (iii) multiplication events, where genotyping relying on biallelic duplications cannot capture the accurate allelic state (e.g. where allelic copy numbers are 2, 5, and 8, instead of 2, 3, and 4).

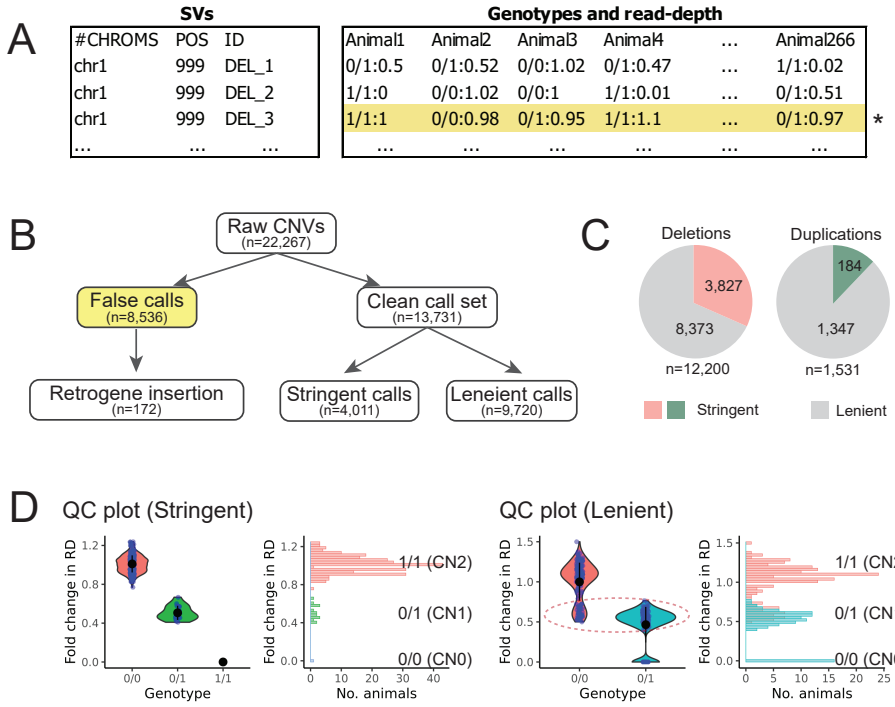


Figure 3.1. Discovery and quality control on SVs in the bovine genome. (**A**) An example of population-side SV detection results. Animals are genotyped for each site, and for CNVs, the fold-coverage change in read-depth is annotated. Marked with yellow is a spurious call where RD do not change according to genotypes. (B) An overview on filtering steps and number of calls in different call sets. (**C**) The overall CNVs calls were divided into stringent and lenient call sets, exploiting the post hoc filter based on RD. The former stands for accurately genotyped biallelic sites. (**D**) Quality control (QC) plots were generated for all CNVs exploiting the genotype and read-depth information. The panel on the left side shows an example of a stringent site where animals GT and RD are unambiguously assigned. Each blue dot represents a sample. The black dots and vertical bars in the violin plot represent the mean and one standard deviation. The right panel represents the RD distribution for each GT group. The QC plot for a lenient site is shown on the right side. In such a case, RD distribution of animals genotyped as 0/0 and 0/1 are overlapping (marked with a red dotted circle), indicating inaccurate genotyping results.

Subsequently, the quality of each CNV call set was evaluated using the family structure in the current data set (127 trios). A quality metric was coined expressing a fraction of trios having Mendelian errors at each site (e.g. with 15 out of 127 trios manifesting Mendelian errors, the fraction corresponds to 0.12). As expected, the stringent call set showed lower Mendelian errors overall than the lenient call set (Figure 3.2 B). Notably, duplications showed higher error rates than deletions in both call sets, suggesting that duplications are prone to have more

genotyping errors even when strict filters are applied. We inspected the site frequency spectra limiting to the stringent call set, which contains accurately genotyped CNVs. Both deletions and duplications showed similar frequency spectra in a sense that they showed many rare variants. However, some deletions were common or even fixed. These deletions suggest assembly issues or mobile elements present in the reference genome, but not in the samples (such cases will be classified as deletions by SV detection tools). Notably, the majority of the stringent duplications were rare, where a handful of them reached allele frequency of ~0.25 (Figure 3.2 C). Finally, we inspected the breakpoints of CNVs. 68% of CNVs had single-base resolved breakpoints, both in stringent and lenient call sets. The high number of single-base resolved breakpoints in the lenient call set, gives confidence that the CNV are correctly called, despite the low genotyping accuracy.

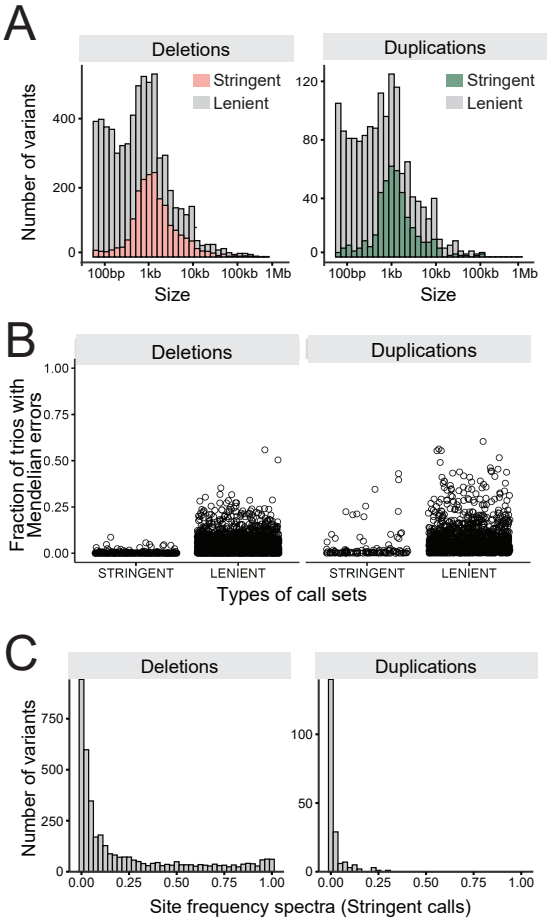


Figure 3.2. Summary of CNV call sets and quality indicator metrics. (**A**) Length distribution of the stringent and lenient call sets. (**B**) Mendelian error fractions obtained for each CNV site are shown for stringent and lenient call sets. (**C**) Site frequency spectra of stringent CNVs.

3.3.1.3. Direct genotyping approach

SVs discovered in the WGS data set, if discovered in unrelated animals in the same population, would confirm the presence of the SVs segregating in the population. To this end, we aimed at validating a subset of the WGS CNVs by directly genotyping the breakpoints of CNVs in unrelated animals. Among the CNVs in the catalogue, breakpoints of 9,642 CNVs had a single-base resolution; thus, genotyping probes could be designed. Of these, we designed probes for 372 CNVs (342 deletions and 30 duplications) appeared in non-repetitive regions and added them to the custom part of the EuroGenomics custom SNP chip genotyping array, which include ~50K SNPs (hereafter referred to as 50K SNP array for brevity; Boichard et al. 2018). Genotyping was done in 815 HF animals, not overlapping with the WGS animals. Of the 284 CNVs (262 deletions and 22 duplications) that passed the quality control (QC) criteria, 210 deletions and 19 duplications were segregating in the genotyped population. The allele frequency of CNVs was skewed towards rare alleles, compared to that of the 50K SNPs obtained from the same array (Figure 3.3). The CNV genotyping results independently validate at least 80% of the 284 CNVs selected from the WGS CNV catalogue. Given that the discovery data is a family cohort, we cannot rule out that the undetected 20% CNVs might be variants private to small families. Furthermore, the 229 CNVs observed in non-related animals confirm that these are variants segregating in the population, which may be exploited in selection.

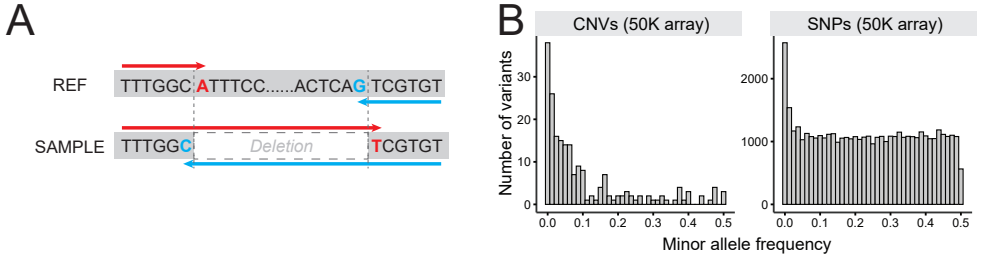


Figure 3.3. Direct genotyping approach and results. (**A**) A schematic overview on primer design. To genotype a deletion, a forward assay can target A (marked with red) in the reference, whereas T will be targeted in deletion carriers. A reverse assay can target G (marked with blue), whereas C will be targeted in deletion carriers. (**B**) Site frequency spectra of CNVs and SNPs obtained from the validation data (50K SNP genotyping array).

3.3.1.4. Comparison with other call sets

We compared the current WGS-based SV catalogue with an array-based catalogue generated from 315 animals from the same HF population (Lee et al. 2020). The concordant CNVs between the WGS- and array- catalogues were mostly large CNVs (231 concordant CNVs, mean size of 33-Kb, min size=1.2-kb, max. size=402-Kb). Given that the size of WGS-based CNVs were discovered mostly around or smaller than 1-kb size (Figure 3.2 A), the WGS-based catalogue seems to contain a large number of finer scale variants undiscovered based on array data.

All things considered, the current SV call set represents major advancements in terms of (i) improved detection sensitivity compared to the array-based catalogue, (ii) diverse types of SVs

included, (iii) high resolution, where 68% of CNVs have single-base resolved breakpoints, and (iv) two CNV subsets (stringent and lenient) representing different levels of genotyping accuracy. Thus, this high-quality call set can be a powerful resource for population and functional analyses.

3.3.2. Population genetics features of detected CNVs

3.3.2.1. CNV-SNP LD in the WGS data set

Although some CNVs associated with complex traits have been delineated (Derks et al. 2018; Kadri et al. 2014), large scale genomic analyses are often centred around utilising SNPs, leaving CNVs unexplored. In theory, if a CNV is in high linkage disequilibrium (LD) with SNPs, those SNPs should sufficiently capture the CNV (e.g. r^2>0.8), functioning as a tagging marker. Hence, we calculated pairwise LD (r^2) between CNVs and SNPs obtained from WGS data to evaluate whether SNPs tag the CNVs. First, we focused on the stringent CNV call set, as it contains accurately genotyped biallelic CNVs. In this call set, 97% and 93% of the deletions and duplications, respectively, were captured by sequence level SNPs, and the CNV-SNP LD broke down as the inter-marker distance increases (Figure 3.4 A and 3.4 B). Our results showed that even rare CNVs were well tagged by SNPs, likely due to rare haplotypes private to particular families (MAF<0.05). Next, we investigated the LD in the lenient CNV call set, and the fraction of tagged CNVs reduced to 83% and 61% for deletions and duplications, respectively (Supplementary Figure 5). The discrepancy between stringent and lenient call sets suggests that the lower degree of LD in the lenient call set arises from inaccurately genotyped CNVs, instead of actual lack of tagging SNPs. We expected that the mCNV-SNP LD would be generally low given that biallelic SNPs would be in partial LD with a multi-allelic variant. Indeed, of 22 mCNVs, only 40% were in LD with SNPs (r^2>0.8), suggesting that some mCNVs are likely missed out in biallelic sequence level SNP based analyses.

3.3.2.2. CNV-SNP LD in the array data set

Our results from the stringent call set showed that sequence level SNPs could capture most of the biallelic CNVs as long as they are accurately genotyped. However, in animal breeding, large-scale genomic analyses (e.g. genomic prediction) rely on 50K, or lower density SNP data. To assess whether array level SNPs capture CNVs, we investigated CNV-SNP LD based on genotypes of 50K SNPs and 284 CNVs directly obtained from our custom 50K SNP array, explained above. In the 50K genotyping array data set, both CNV-SNP and SNP-SNP pairs showed LD decay where the degree of LD declines as a function of inter-marker distance. Intriguingly, CNV-SNP pairs showed lower LD than SNP-SNP pairs, regardless of the allele frequency (Figure 3.4 C and 3.4 D). Only 15.4% of deletions and 4.5% of duplications had tagging SNPs (r^2>0.8), suggesting that the current set of SNPs on the 50K may not capture the variation from CNVs.

3.3.3. Functional impact of SVs

The functional consequence of SVs varies depending on many factors, including SV types, event sizes, the overlap with coding sequences (CDS), and the overlap with non-coding regula-

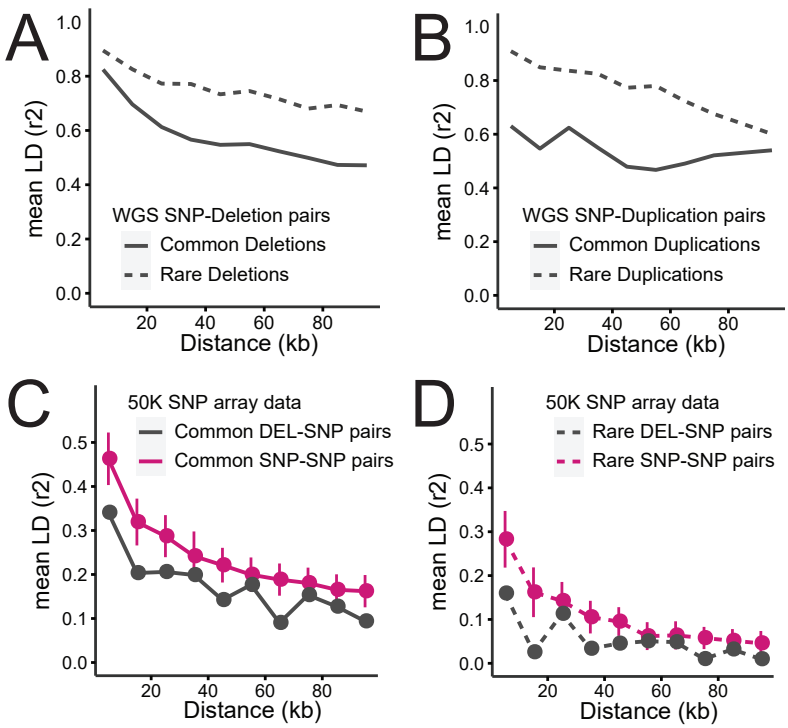


Figure 3.4. Linkage disequilibrium between SNPs and CNVs. (**A**) Mean r^2 obtained from deletion-SNP pairs discovered in WGS data is displayed as a function of inter-marker distances. SNPs paired with common deletions (MAF≥0.05) are marked with a solid line, whereas SNPs paired with rare deletions (MAF<0.05) are marked with a dotted line. (**B**) Mean r^2 obtained from duplication-SNP pairs discovered in WGS data is shown. The legend is the same as the panel (A). (**C**) Mean r^2 obtained from 50K SNP genotyping array is displayed for common variants only (MAF≥0.05). SNP-SNP pairs are marked with a solid magenta line, and DEL-SNP pairs are marked with a solid grey line. The SNP-SNP pairs outnumbered DEL-SNP pairs. To keep the comparison not influenced by the difference in the number of pairs, a subset of SNP-SNP pairs, equivalent to the number of DEL-SNP pairs, was made 1,000 times, and the mean and the standard deviation are displayed in the figure. (**D**) Mean r^2 obtained from 50K SNP genotyping array is displayed for rare variants only (MAF<0.05). Legends are identical to the panel (**C**).

tory elements. In case of deletions and retrogene insertions, they may overlap or occur within CDS of a gene, thus leading to loss-of-function of the gene. In contrast, duplications may have different consequences depending on overlap with CDS. For instance, if a duplication partially overlaps with a coding gene (e.g. overlapping with a subset of exons), it may alter the transcript(s). However, if a duplication harbours an entire gene, it could lead to increased gene expression in theory. Therefore, we categorised CDS overlapping SVs into predicted loss-of-function (pLoF) for CDS disrupting deletions and retrogene insertions, intragenic exonic duplication (IED) for duplications with partial genic overlap, and copy gain (CG) for duplication encompassing entire gene(s), following (Collins et al. 2020) (Figure 3.5 A).

3.3.3.1. CDS affecting SVs

Our SV catalogue contained 426 genes where their CDS were affected (342 pLoFs, 41 CGs, 50 IEDs), and each individual had on average 88 pLoFs, 7.8 CGs, and 8.1 IEDs events. The list of CDS disrupting SVs contained two known causal CNVs: (i) a pLoF event by a 3.3-Kb deletion ablating *FANCI* gene, causing foetal death and brachyspina (Charlier et al. 2012), (ii) a pLoF event by a 138-Kb deletion ablating *TFB1M* gene, which was associated with lethal haplotype mapped in HF population (Schütz et al. 2016), and (iii) an IED event by a 12-Kb mCNV overlapping with the last exon of *GC* gene, associated with mastitis resistance (Lee et al. 2021). As expected, common CDS disrupting SVs (MAF>0.05) were often affecting genes belonging to large gene families (e.g. olfactory receptors), whereas rare SVs often disrupted essential genes without paralogue. For example, we discovered a singleton 50-kb deletion ablating *Centromere Protein C* gene (*CENPC*), which was shown to be recessive lethal in a mouse knock-out study (Kalitsis et al. 1998). Moreover, we identified a 16-kb IED event in the BTA5:27.4Mb region that harbours a large repertoire of keratin genes (Figure 3.5 B). This IED event was classified as mCNV based on the multi-modal RD distribution, which indicated diploid CNs between 2 and 6 (Figure 3.5 C). Furthermore, the QC plot implied inaccurate genotyping (e.g. mCNV carriers with high RD were genotyped as 0/0). Close inspection of carrier animals supported the presence of the mCNV (elevated sequencing coverage; Figure 3.5 D), however the reads spanning over breakpoints had low mapping quality likely resulting in inaccurate genotyping (Supplementary Figure 6). This 16-kb mCNV disrupts two keratin genes that are in the same orientation (*KRT6B* and *KRT6C*), and thus can give rise to a novel fusion gene (Figure 3.5 E). In such case, a diploid CN6 animal is expected to have intact *KRT6B* and *KRT6C* genes and 4 copies of *KRT6B-KRT6C* fusion genes.

3.3.3.2. Molecular characterisation of SV-eQTL

Recent human SV catalogue showed that most SVs are under purifying selection, thus segregating at low allele frequency, except duplications encompassing entire gene(s) (Collins et al. 2020). Therefore, we focused on the 41 CG events aiming at identifying functional duplications. The underlying assumption of functional CG events is that an extra copy of a gene can increase gene expression. Thus, mapping SV expression QTL (SV-eQTL) for CG events seemed a plausible approach to elucidate SVs' molecular contribution. Before SV-eQTL mapping, we prioritised CG events harbouring genes that previous studies reported associations with economically important traits in cattle. Based on the literature, we found two promising CG loci. The first is an 85-kb duplication harbouring *Orosomucoid 1* (*ORM1*; chr8:103,486,032-103,571,582) (Supplementary Figure 7 A-D). *ORM1* is predominantly expressed in liver and encodes acute-phase plasma protein, and has been shown to be upregulated in response to acute inflammation (Sun et al. 2016). In dairy cattle, an increased *ORM1* expression in post-partum cows was associated with decreased feed intake (Brown et al. 2021; McGuckin et al. 2020). The second is a 150-kb duplication harbouring *Popeye Domain Containing 3* (*POPDC3*; chr9:44,725,475-44,875,600) (Supplementary Figure 7 E-H). *POPDC3* is involved in skeletal

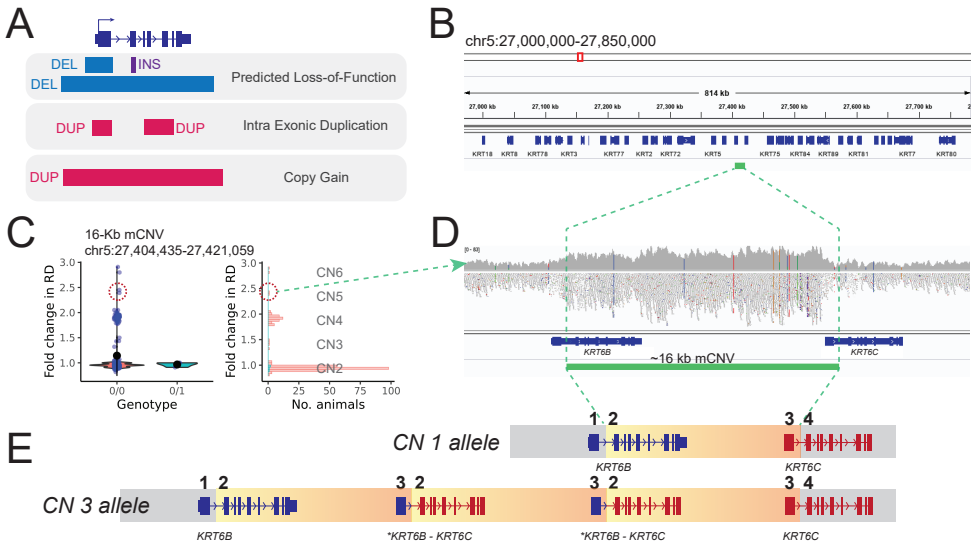


Figure 3.5. CDS disrupting SVs. (**A**) Three different categories of CDS disrupting SVs. CDS disrupting deletions and insertions can lead to loss-of-function variants. If affecting an entire gene, Duplications are equivalent to obtaining an extra copy of a gene (copy gain). However, partial duplication of a gene may have different consequences depending on the context. Figure adapted from(Collins et al. 2020). (**B**) A 16-kb mCNV was found in the Keratin gene-rich region, harbouring more than 20 keratin genes, in the chr5:27Mb region (marked with green). This mCNV affects two keratin genes, *KRT6B* and *KRT6C*. (**C**) The QC plot of the 16-Kb implied that diploid CNs range between 2 and 6, yet reads-based GT indicated inaccurate genotyping. (**D**) WGS data of one of the mCNV carriers was inspected (diploid CN 5). Increased sequencing coverage supports the presence of multiple copies of the 16-kb segment. (**E**) The Tandem arrangement of the 16-kb segment can give rise to a novel fusion gene made of part of *KRT6B* and *KRT6C* (marked with an asterisk; shown in blue and red). In this panel, we depicted a putative tandem arrangement of the haploid CN3 allele.

muscle tissue development and is broadly expressed in multiple tissues (Fang et al. 2020). This 150-kb duplication was associated with hoof health traits in Canadian HF population; however the effect direction was not reported (Butty et al. 2021).

We proceeded with SV-eQTL mapping exploiting BovineHD genotype and liver RNA-seq data obtained from postpartum (d14) dairy cows (n=175). To associate that gene expression with sequence level variants, we generated an imputation panel consisting of SNPs and SVs discovered in 266 WGS animals (materials and method; Supplementary Figure 7I). The BovineHD genotype was imputed to sequence level variants, and then associated with gene expression. The ORM1 duplication was well imputed and ranked as the top variants for *ORM1* eQTL. Same procedure was applied to *POPDC3* duplications and likewise, the imputed POPDC3 duplication was shown to be ranked among top eQTL variants for *POPDC3* (Figure 3.6, Supplementary Figure 7I). Furthermore, bovine liver ChIP-seq data (H3K27ac and H3K4me3; Villar et al. 2015) confirmed the presence of promoters for these genes, providing a mechanistic explanation on these liver SV-eQTL (Supplementary Figure 7 C, G).

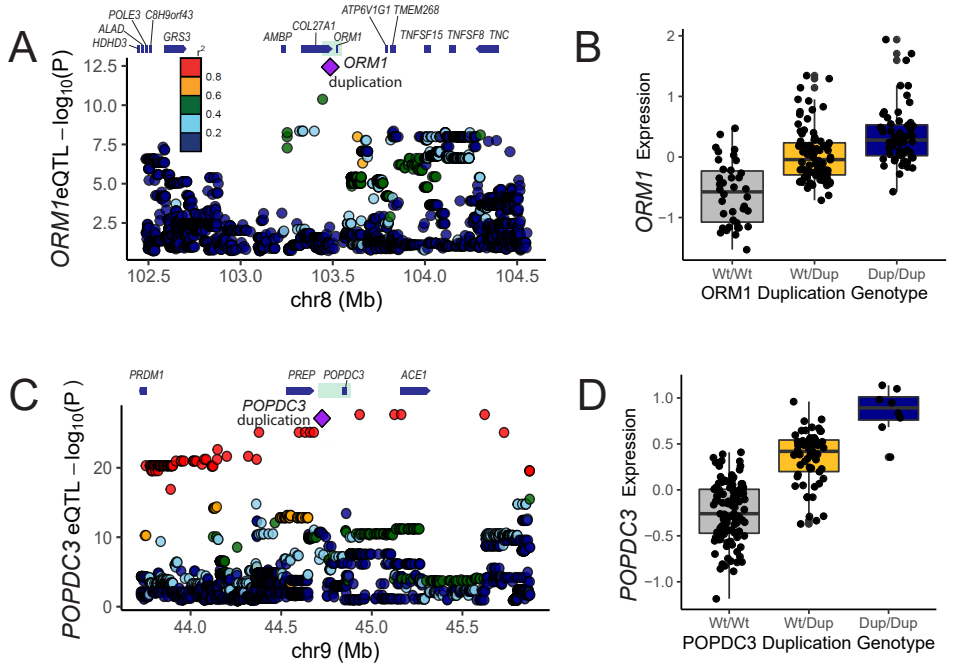


Figure 3.6. CDS disrupting SVs. (**A**) eQTL mapping result for *ORM1*. The ORM1 duplication is marked with purple diamond. The colour scale indicates the degree of pairwise LD (r^2) between the ORM1 duplication and other SNPs. Green translucent box marks the duplication. (**B**) The box plot shows altered *ORM1* expression depending on the ORM1 duplication genotypes. (**C**) eQTL mapping result for *ORM1*. Legend is same as panel (**A**). (**D**) The box plot shows altered *POPDC3* expression depending on the POPDC3 duplication genotypes.

Accordingly, these duplications are likely the underlying variants driving differences in the gene expression (Figure 3.6). Extrapolating the literature, we could hypothesise that the *ORM1* duplication allele, which leads to high *ORM1* expression, will decrease the feed intake in postpartum cows (McGuckin et al. 2020; Brown et al. 2021). In mice, administration of exogenous ORM supressed food intake, via binding leptin receptors, which induce activation of signal transducer and activator of transcription 3 (STAT3) signalling (Sun et al. 2016). Recent dairy cattle study showed that high ORM expression suppressed postpartum feed intake, yet without triggering STAT3 signalling, leaving the underlying appetite suppression mechanism elusive (McGuckin et al. 2020). It is worth noting that this variant is highly frequent (AF=0.49), despite its association with the reduced postpartum feed intake, which is considered detrimental for cows. One possible explanation could be that this variant is under balancing selection. In an attempt to identify target trait(s) under selection, the animal QTL database was screened (Hu et al. 2016), however, there was no QTL reported in the region of interest.

3.4. Discussion

In this study, we used 266 deeply sequenced Dutch dairy cattle genomes to discover SVs. SV discovery and our understanding of SVs have been hindered by low detection sensitivity and inaccurate genotyping issues often arising in low sequencing depth samples (Huddleston and Eichler 2016). Also, SV detection tools relying solely on a single detection principle were shown to generate many false positive calls, compared to ensemble callers (Kosugi et al. 2019; Cameron et al. 2019), resulting in a low-quality call set. To address these issues, we discovered SVs in Damona pedigrees, consisting of deeply sequenced healthy family cohort, consisting of many trios (127 trios, mean sequencing depth=26X; Harland et al. 2017), using a detection method exploiting multiple SV signals. Detection of SVs can benefit from high coverage sequencing data in two ways. First and foremost, it will inevitably improve detection sensitivity. There are more SR and RP evidence supporting SVs for a given locus, leading to an increased number of discovered variants. Secondly, the RD can be measured accurately, hence can be exploited to filter out spurious false positives (e.g. heterozygous deletion without RD change can be filtered out). Of these, we exploited RD to distinguish the clean CNV calls into stringent calls with high genotyping accuracy. Thus, if future studies aim to investigate SVs using allele frequency-based analytical tools (e.g. Fst), the stringent call set may be a good starting point as it contains accurately genotyped CNVs. It is worth noting that duplications had overall higher Mendelian errors than deletions. If a deletion is present in one's genome, it can be in either of two states: a heterozygous deletion (diploid CN1) or homozygous deletion (diploid CN0). Contrary to this, duplications can be biallelic or multi-allelic, and CNs may not ascend sequentially. If the CNs do not ascend sequentially (CNs 2, 5, and 8, instead of CNs 2, 3, and 4), heterozygous animals (e.g. CN5) are genotyped as homozygous for duplication, due to overwhelmingly large number of discordant reads – leading to Mendelian error. Thus, exploiting duplications may require more effort to characterise the true underlying CN states, requiring high sequence coverage to identify read depth differences.

High number of false positives and negatives in SV discovery makes it crucial to perform post-discovery evaluation (Huddleston and Eichler 2016). Commonly used orthogonal validations methods include long-read WGS and PCR amplicons (Bertolotti et al. 2020; Zhao et al. 2021; Collins et al. 2020; Abel et al. 2020). However, often these validations can be costly and time-consuming, and above all, the availability of DNA material can be a bottleneck. Instead, to some extent, we bypassed these issues by incorporating CNV targeting probes into the 50K SNP array that is routinely used in livestock breeding programmes. With this approach, we obtained accurate genotypes of CNVs and SNPs simultaneously, and found that 80% of the targeted CNVs are segregating in the validation cohort. It is likely that this 80% validation is a lower-bound, given the nature of the discovery cohort (266 family animals). We deem rare CNVs private to some families may be confirmed with a larger validation cohort. Particularly, accurate genotyping of some CNVs remain challenging even with ~30X coverage WGS data (lenient call set). As such, unless WGS data allows accurate genotyping (e.g. stringent call set), custom array based direct SV genotyping will be indispensable in identifying trait-associated CNVs. Furthermore, accumulation of SNP and CNV genotypes from ~thousands of animals

will open an opportunity to evaluation the genetic contribution of CNVs, relative to 50K SNPs.

Our and others' work demonstrated a large repertoire of functional SVs, many of which are of interest for livestock breeding (Bickhart and Liu 2014; Clop et al. 2012). In livestock breeding, the genetic merit of animals is estimated based on the genomic prediction that exploits 50K SNPs. Thus, whether the 50K SNPs fully capture the variation from CNVs is a prime question. The CNV-SNP LD shown in our WGS data set revealed that most CNVs in the stringent call set have tagging SNPs (97% deletions and 92% duplications; Figure 3.4 and Supplementary Figure 5), higher than recent reports in human SV studies comprising >10,000 genomes of diverse ethnic background (Collins et al. 2020; Abel et al. 2020). Unlike human studies, we studied a family cohort from a single breed, likely leading to the upper bound of the LD. However, the CNV-SNP LD in the SNP array data set was lower, because (i) the genotyped animals were unrelated and (ii) the MAF of SNPs and CNVs did not match well – CNVs were skewed towards rare variants, whereas SNPs were uniform across the range of MAFs, unlike the sequence level SNPs that include variants from a full spectrum of allele frequencies (Figure 3.3 B). Thus, large scale genomic analyses aiming at investigating CNVs may consider the two options: (1) exploit imputed sequence level SNPs that tags CNVs, or (2) directly genotype CNVs. The first approach may have some limitations, given that many CNVs are rare and so do the tagging SNPs, posing difficulty in imputation. Also, this approach will work only for the CNVs that are accurately genotyped in the sequenced reference population. The ORM1 duplication genotyping was inaccurate in our WGS data (classified as lenient call), whereas the direct genotyping approach showed accurate genotyping results. Of note, the duplication targeting probe was not in LD with the 50K SNPs in vicinity. Thus, associating traits with the direct genotype of the ORM1 duplication may help unravel why this duplication segregate in high frequency, despite the negative association with postpartum feed intake. Lastly, haplotype-based approach was shown to capture CNVs well (Derks et al. 2018; Kadri et al. 2014; Zhang et al. 2012), hence can be an alternative choice, if both of the suggested options (tagging SNPs and direct genotyping) are not feasible.

Due to large event sizes, a single SV may have larger effect, compared to smaller variants (Collins et al. 2020; Abel et al. 2020). Under such circumstances, SVs with deleterious effects, if affecting haploinsufficient gene(s), are expected to be purged rapidly. Hence, it is assumed that most SVs would have a benign effect, unless they confer an adaptive advantage (Campbell and Eichler 2013). Our SV catalogue contains 426 genes affected by SVs, where each animal has on average >100 affected genes.

Mapping deleterious variants can be done exploiting (i) a phenotype driven approach (e.g. GWAS), which requires high allele frequency, and (ii) a genotype driven approach (e.g. homozygosity depletion mapping), which requires a very large genotyped population and high allele frequency. Since the study population is a healthy family cohort of modest size (n=266), discovering novel it was unlikely to detect statistically robust recessive lethal variants. This does not preclude that recessive lethal SVs are segregating in the current population. We confirmed presence of two known recessive lethal deletion is segregating in the current population (Charlier et al. 2012; Schütz et al. 2016), which served as positive controls; however, as expected, we

did not see any homozygous carrier, and the recessive allele was segregating at low frequency (MAF for FANCI deletion=0.06, MAF for TFB1M deletion=0.005). Additionally, we detected a singleton pLoF 50-Kb deletion affecting *CENPC*, shown to result in an early embryonic loss in knock-out mice (Kalitsis et al. 1998). As such, mapping deleterious variants using statistical association may not be suitable for the current data set, however exploiting a wealth of annotation data in human and mouse can help interpretation of the gene-disrupting SVs.

We mapped two promising SV-eQTL for CG events overlapping with *ORM1* and *POPDC3,* respectively (Figure 3.5 and Supplementary Figure 7). *ORM1* encodes acute phase protein and is involved in energy metabolism: mice lacking *ORM1* expression were shown to have increased body weight and fat mass (Sun et al. 2016), whereas upregulation in postpartum cows was correlated with reduced feed intake (Brown et al. 2021; McGuckin et al. 2020). Unfortunately, no feed intake QTL was reported near this duplication in the Animal QTL database and GWAS (Li et al. 2019; Hu et al. 2019). There are several explanations for this conundrum. One possibility is that the duplication itself or tagging SNPs are inaccurately genotyped, leaving no association signals. Another possibility may have to do with the transient expression of *ORM1. ORM1* is strongly upregulated from parturition to postpartum day 14, hence suppressing feed intake during this short period. However, in breeding programmes, feed intake traits are defined as an overall mean during lactation (Veerkamp et al. 2014). Thus, the suppressed feed intake during the first ~2 weeks may be diluted at the end. A future study might define a novel feed intake trait limited to feed intake during ~2 weeks postpartum. Additionally, it is remarkable that the variant has a high frequency despite its negative consequence, hinting that it might be under balancing selection. As with the feed intake QTL, we have not found any QTL associated with other traits, which may be logical if association studies did not have proper variants capturing this duplication. Interestingly, *ORM1* duplication was reported in human populations as well. Diploid CNs of *ORM1* is high in the European population (CN >10) compared to the African population (CNs 2-3; Handsaker et al. 2015), suggesting that upregulation of *ORM1* might confer a generic adaptive advantage across species.

3.5. Conclusion

This study reports a high-quality SV catalogue containing 13,925 SVs detected in deeply sequenced genomes obtained from a Dutch HF family cohort. Using the direct genotyping approach, we genotyped a subset of CNVs in an independent cohort and confirmed that 80% of the targeted CNVs are segregating in the population. In search of high impact SVs, we prioritised two duplications overlapping with *ORM1* and *POPDC3*, associated with feed intake and hoof health traits, respectively. The eQTL mapping results corroborate that these duplications are likely the causal variant driving the gene expression, underpinning the functional importance of SVs. Given the functional impact of SVs, incorporating them in large scale genetic analyses would be crucial. Yet, our LD analyses showed that most CNVs are not captured by 50K SNPs, stressing the importance of incorporating CNVs into routine analyses either by directly genotyping or exploiting CNV tagging SNPs. The current high-quality SV catalogue will serve as an invaluable resource for future population genetics studies.

3.6. Materials & Methods

3.6.1. Whole genome sequencing and variant discovery

3.6.1.1. Whole genome sequence data

The genomes of 266 Dutch HF animals were sequenced. These 266 animals were closely related animals, where 240 were forming 127 parents-offspring trios. The biological materials were either from sperm (males) or whole blood (females and males). Whole genome Illumina Nextera PCR free libraries were constructed (550bp insert size) following the protocols provided by the manufacturer. Illumina HiSeq 2000 instrument was used for sequencing, with a paired-end protocol (2x100bp) by the GIGA Genomics platform (University of Liège). The data was aligned using BWA mem (version 0.7.9a-r786) (Li 2013) to the bovine reference genome ARS-UCD1.2 (Rosen et al. 2020), and converted into bam files using SAMtools 1.9 (Li and Durbin 2009). Subsequently, the bam files were sorted and PCR duplicates were marked with Sambamba (version 0.4.6) (Tarasov et al. 2015). All samples had a minimum mean sequencing coverage of 15X, and the mean coverage of the bam files was 26X.

3.6.1.2. Structural variation discovery

We discovered SVs using Smoove pipeline (https://github.com/brentp/smoove). This pipeline collects split and discordant reads using Samblaster (Faust and Hall 2014) and then discovers SVs per sample. The SV discovery was done using Lumpy, which detects deletions, duplications, inversions, and breakends (non-canonical forms of SVs) (Layer et al. 2014). The per sample SV discovery showed that the number of SVs discovered per sample was in relation to the sequencing coverage (Supplementary Figure 1). We did not find any outlier samples in terms of the total number of SV per sample and the number of singleton SVs per sample. Hence, the entire cohort of 266 animals was kept for further analysis. After the sample level SV discovery, the SVs were merged, creating a population-wide non-redundant SV call set. Subsequently, the entire cohort was genotyped for the non-redundant SV sites using SVTyper (https://github.com/hall-lab/svtyper). Additionally, the fold-coverage change of RD in SV was annotated by Duphold (Pedersen and Quinlan 2019). Duphold annotated two RD values: (i) DHFFC representing sequencing depth fold-change for the variant compared to 1-kb flanking regions, and (ii) DHBFC representing sequencing depth fold-change for the variant compared to genomic regions with similar GC-content. We used DHFFC for filtering deletions and DHBFC for filtering duplications, as recommended by the developer.

3.6.1.3. Site-level filtering

SVs smaller than 50-bp were filtered out following the definition in (Sudmant et al. 2015). True CNVs are expected to show altered RD than diploidregion without CNVs. On the contrary, low-quality false CNVs are often caused by noise (e.g. repeats), hence do not involve RD change. Thus, we applied preliminary filters to remove spurious sites in which the mean RD values for different GTs do not differ. With this filter, 13,731 CNVs were retained out of the 22,267 CNVs .

Afterwards, we classified CNVs into stringent and lenient call sets. The stringent set retained 3,827 deletions and 184 duplications, where RD values per GT correspond unambiguously to the cognate GT, whereas The lenient set retained 8,373 deletions and 1,347 duplications where RD values do not always match with the cognate GT (Figure 3.1). By manually inspecting QC plots of each duplication, we identified sites indicating the presence of >2 alleles. These sites were re-classified as mCNVs (Supplementary Figure 2). Finally, we examined low-quality deletions filtered out by the preliminary filters - some of these false deletions overlapped with the intronic region, indicating processed pseudogene insertions (Supplementary Figure 4) . We detected the source gene of the retrogene insertion using 95% reciprocal overlap between deletions and introns. Also, the insertion sites were identified by retrieving the breakends that are located within +/- 1Kb from the gene body.

3.6.2. Evaluation of the SV call set

3.6.2.1. Mendelian inheritance errors

Using the 127 trios, we coined a quality assessment metric based on Mendelian inheritance. We counted the number of trios showing Mendelian inheritance error and expressed it as a fraction. For example, for a CNV site, if 10 trios showed Mendelian error, we assigned 0.08 (10/127=0.08). Hence, the scale ranged from 0 to 1, where 0 stands for no trios showing inheritance error, whereas 1 indicates all of the 127 trios showing inheritance error. We calculated this metric for all the CNV sites, both lenient and stringent calls.

3.6.2.2. Direct genotyping of CNVs using a 50K SNP array

To avoid highly costly validations, we opted for directly genotyping of a subset of CNVs. We designed probes directly targeting the breakpoint sequences of 372 CNVs (342 deletions and 30 duplications) appeared in non-repetitive regions. These probes were added in the custom part of the EuroGenomics SNP genotyping array (Boichard et al. 2018). Genotyping was done for 815 Dutch HF animals using their ear punch or blood samples. Of note, these samples did not overlap with the WGS data set samples. All of the genotyped samples passed the quality criteria (call rate per sample>0.99). Of the 284 CNVs that passed variant level filter (call rate per variant>0.99), 229 were segregating in the population, confirming the presence of the variant in the population.

3.6.2.3. Comparison with an array-based CNV catalogue

The overall WGS-based CNV call set (including both lenient and stringent calls) was compared to an array-based CNV call set. The array-based CNV call set was obtained based on Illumina BovineHD Genotyping BeadChip (770K) in 315 HF animals, which are independent of the current WGS data set (no overlapping samples; Lee et al. 2020). Of the 315 animals, 34 were overlapping with the WGS samples. The event size determined from an array-based CNV detection is strongly dependent on the local probe density. Thus, applying reciprocal overlap criteria for comparing array- and WGS- call set may underestimate the true overlapping calls. Accordingly, we intersected two call sets in IGV and manually inspected the underlying WGS data for overlapping calls. Where array- and WGS- based sites are overlapping and the un-

derlying WGS data supports true presence of CNVs, we confirmed them as overlapping calls.

3.6.3. Population genetics analyses

3.6.3.1. Linkage disequilibrium in WGS data sets

We investigated the CNV-SNP LD using WGS data sets. SNPs were discovered from the same WGS data set explained above. Variant calling was done using GATK workflow (v4.1.7) and subsequently recalibrated using the following algorithms: BaseRecalibrator, HaplotypeCaller, GenomicsDBImport, GenotypeGVCF, GatherVcfs, Variant Recalibrator (DePristo et al. 2011; McKenna et al. 2010; Auwera et al. 2013). We applied Variant Quality Score Recalibration (VQSR) at a truth sensitivity filter level of 97.5 to remove spurious variants. For calculating CNV-SNP pairwise LD, SNPs located inside the CNVs were removed, and SNPs located within 100-kb distance from the CNV breakpoints were kept. Pairwise CNV-SNP LD (r^2) was obtained from PLINK software (v1.9) (Purcell et al. 2007).

3.6.3.2. Linkage disequilibrium in 50K SNP array data sets

The genotype data obtained from 50K SNP array, augmented with probes targeting the CNV breakpoints, was used to obtain 50K level CNV-SNP LD (explained above). Genotyping was done for 815 Dutch HF animals, and all samples passed the quality criteria (call rate per sample>0.99). Of 53,917 SNPs and 284 CNVs that passed variant level filter (call rate per variant>0.99), 50,342 SNPs and 229 CNVs were segregating in the population. As the number of segregating duplications was low (n=19), we only performed the LD analyses on 210 deletions. We compared SNP-SNP and CNV-SNP LD depending on the inter-marker distance. The number of CNVs was lower than SNPs, and hence SNP-SNP pairs outnumbered CNV-SNP pairs. To compare the same number of pairs, we subset an equal number of SNP-SNP pairs 1,000 times and compared the mean LD with the CNV-SNP pairs. Pairwise SNP-SNP and CNV-SNP LD (r^2) was obtained from PLINK software (v1.9) (Purcell et al. 2007). The analyses were ran for common (MAF≥0.05) and rare (MAF<0.05) variants separately.

3.6.4. Functional impact of SVs

3.6.4.1. Coding sequence disrupting SVs

We classified coding sequence (CDS) disrupting SVs into predicted loss-of-function, copy gain, and intergenic exon duplication following (Collins et al. 2020). The CDS disrupting SVs were identified using Variant Effect Predictor (Ensembl release 98) (McLaren et al. 2016). Retrogene insertions were screened in the case when the breakends indicated insertion sites.

3.6.4.2. Regulatory elements disrupting SVs

Bovine liver ChIP-seq data (H3K27ac and H3K4me3) was obtained from ArrayExpress (E-MTAB-2633; Villar et al. 2015). This ChIP-seq data was aligned to the bovine reference genome ARS-UCD1.2 using Bowtie2 (Langmead and Salzberg 2012), and peaks were called using MACS2 (Zhang et al. 2008). The SVs overlapping with enhancer or promoter signals were identified using BedTools software (Quinlan and Hall 2010). Based on the strength of the regulatory elements signal, the allele frequency of the SVs (MAF > 0.05), and literature sug-

gesting their functional roles in phenotypes (Butty et al. 2021; Brown et al. 2021; McGuckin et al. 2020), we selected two SVs for subsequent SV-eQTL mapping.

3.6.5. SV-eQTL mapping

3.6.5.1. Genotype data and imputation

Liver biopsy samples were collected from ~14 day postpartum HF cows (n=178). The procedures had local ethical approval and complied with the relevant national and EU legislation under the European Union Regulations 2012 (S.I. No. 543 of 2012). These samples were genotyped using Illumina BovineHD Genotyping BeadChip (770K). The genotype file was subset to have a 10-Mb region harbouring the SV of interest, and rare variants were filtered out (MAF<0.02). We included both SNPs and the SVs of interest in the imputation panel to see whether SV(s) is the underlying variant driving the gene expression. The SNPs were discovered, as explained above. Two different SV genotypes were used: (i) the original SV genotypes obtained from SVtyper (https://github.com/hall-lab/svtyper) based on RP and SR evidence and (ii) manually determined RD-based genotypes. The initial imputation panel included SNPs in a 10-Mb region harbouring the SV of interest and two different genotypes of SVs (contained 266 WGS animals). This panel was phased, and variants with low phasing accuracy and allele frequency were filtered out (DR2<0.95 and MAF<0.02). Subsequently, the BovineHD genotypes were imputed to the sequence level variants and variants with low imputation accuracy and low minor allele frequency were filtered out (allele R^2<0.9 and MAF<0.025). Phasing and imputation were done using Beagle 4 (Browning et al. 2018).

3.6.5.2. RNA-seq data and eQTL mapping

We mapped eQTL for a subset of liver regulatory elements overlapping SVs. The RNA-seq data was obtained from the GplusE consortium (http://www.gpluse.eu; EBI ArrayExpress: E-MTAB-9348 and 9871) (Wathes et al. 2021). RNA-seq libraries were constructed using Illumina TruSeq Stranded Total RNA Library Prep Ribo-Zero Gold kit (Illumina, San Diego, CA) and sequenced on Illumina NextSeq 500 sequencer with 75-nucleotide single-end reads to reach average 32 million reads per sample. The reads were aligned to the bovine reference genome UMD3.1, and its corresponding gene coordinates from UCSC as a reference using HISAT2 (Kim et al. 2019). Transcript assembly was conducted with StringTie (Pertea et al. 2016), using a reference-guided option for transcript assembly. Reads were counted at gene level using StringTie. Subsequent quality control on the RNA-seq data set removed three samples with suboptimal quality (QC steps are explained in detail elsewhere; Lee et al. 2021), normalised gene expression was associated with the imputed WGS variants for 175 samples, using a linear model in R package "MatrixEQTL" (Shabalin 2012).

3.7. Acknowledgements

The Dutch HF whole genome sequence population data set was funded by the DAMONA ERC advanced grant to MG. CC is senior research associate from the Fonds de la Recherche Scientifique–FNRS (F.R.S.-FNRS).

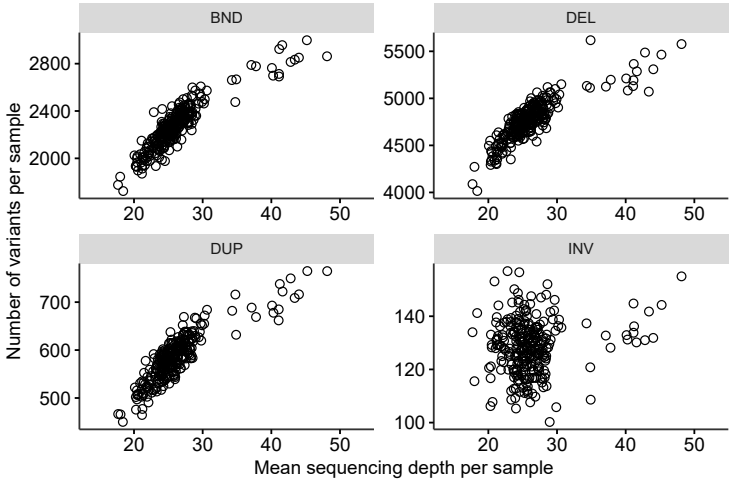
3.8. Research Funding

YLL, ACB, MB, MAMG, and RFV are financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. This work was supported by grants from the European Research Council (Damona to MG; award number: ERC AdG-GA323030), and the EU Framework 7 program (GplusE to MG and HT; award number: 613689). GCMM is post-doctoral fellow of the H2020 EU project BovReg (Grant agreement number: 815668).

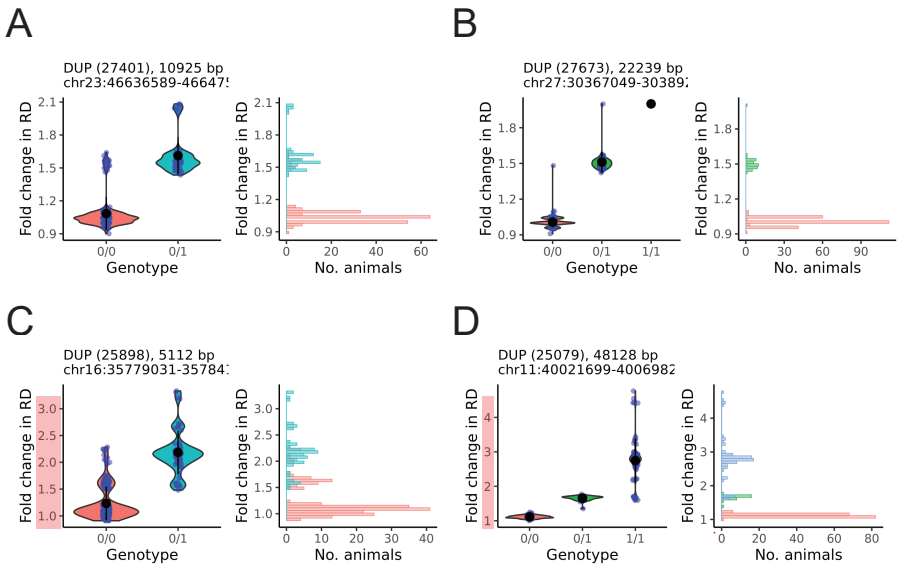
3.9. Authors' contributions

RFV, MAMG, and MG contributed to conception of the study. YL performed the analyses and drafted the manuscript under supervision of CC, ACB, and MB. WC and LK generated WGS data and GCMM mapped the data. HT generated the eQTL mapping data. COE and TD performed variant calling. All authors approved the manuscript.

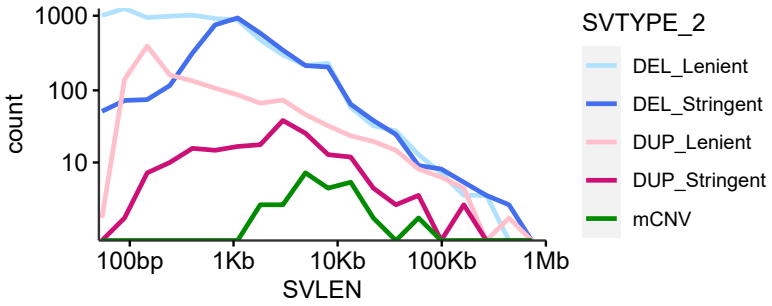
3.10. Supplementary figures



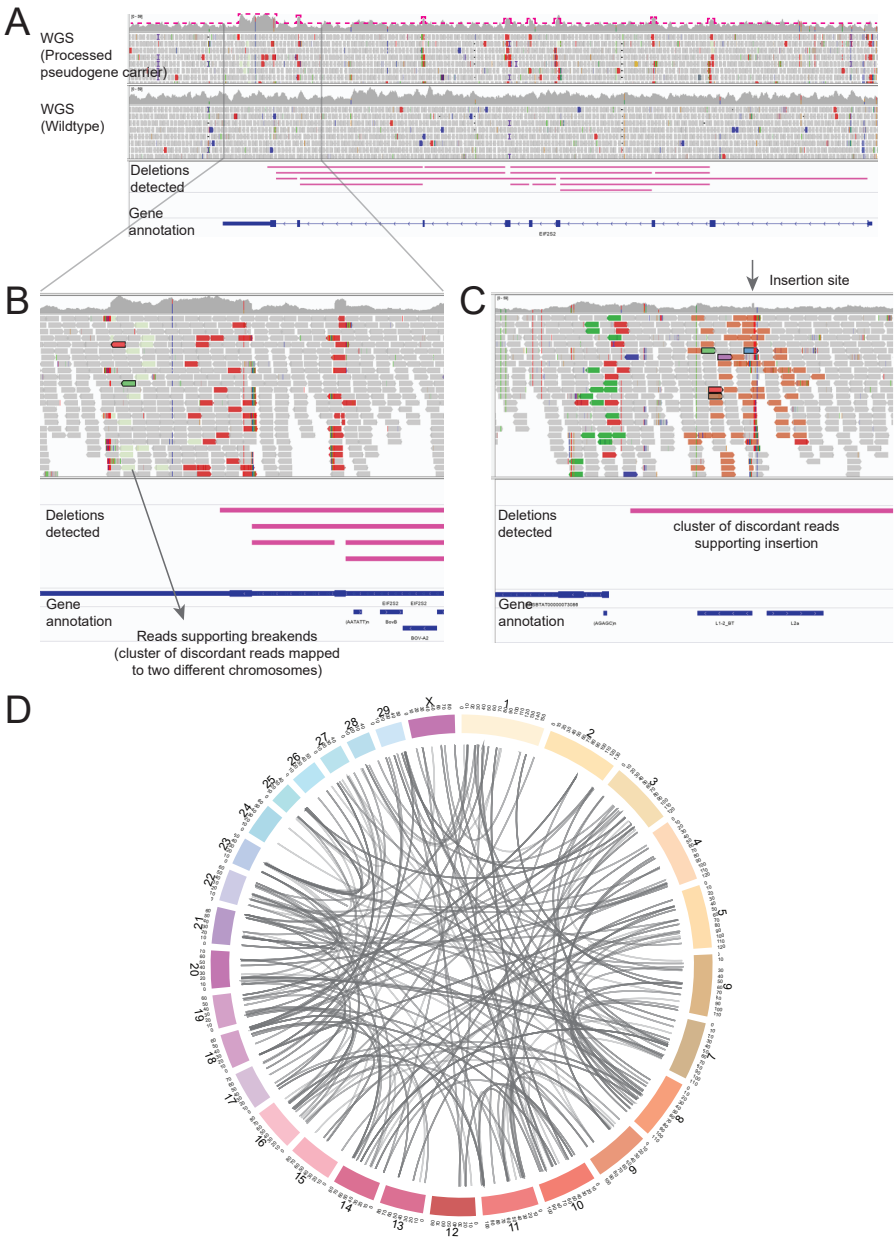
3.10.1. Supplementary Figure 1 Number of discovered SVs depending on mean sequencing depth. Our SV discovery results showed that the number of SVs discovered per sample increased linearly in relation to the mean sequencing depth. Each panel stands for four types of SVs: BND (breakends), DEL (deletions), DUP (duplications), and INV (inversions). Each dot represents a sample. For BND, DEL, and DUP, the increase of discovered variants according to the sequencing depth was evident. We did not observe spurious batch effects. There was no spurious sub-clusters (e.g. batch effects).



3.10.2. Supplementary Figure 2 Distinctive RD distribution discerning mCNVs from biallelic duplications. We re-classified 2 2 duplications as mCNVs based on their RD distribution. (**A**, **B**) Panel A and B show examples of biallelic duplication sites where RD values form the trimodal distribution. (**C**, **D**) In the case of multi-allelic duplications, there are more than three allelic combinations in diploid states and hence show distinctive multi-modal RD distribution. Also, the range of RD values is broader than biallelic duplications (x-axes marked with translucent red).

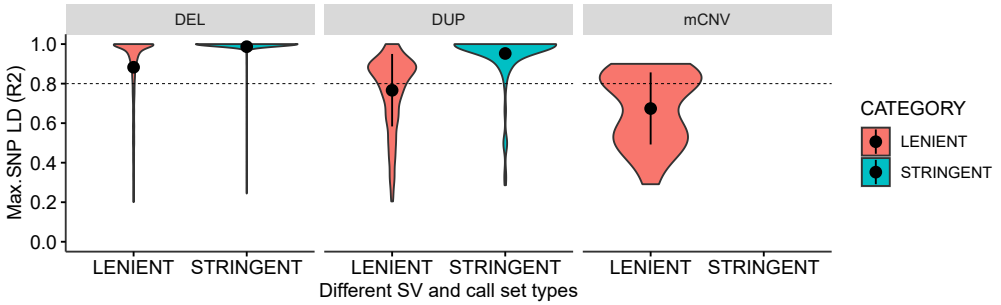


3.10.3. Supplementary Figure 3 Size distribution of overall CNVs. Size distribution of overall CNVs are shown (deletions, duplications, and mCNVs). The deletions are marked with light blue (lenient calls) and dark blue (stringent calls). The duplications are marked with light pinks (lenient) and magenta (stringent). mCNVs are marked with green and they are discovered in >1-kb size.

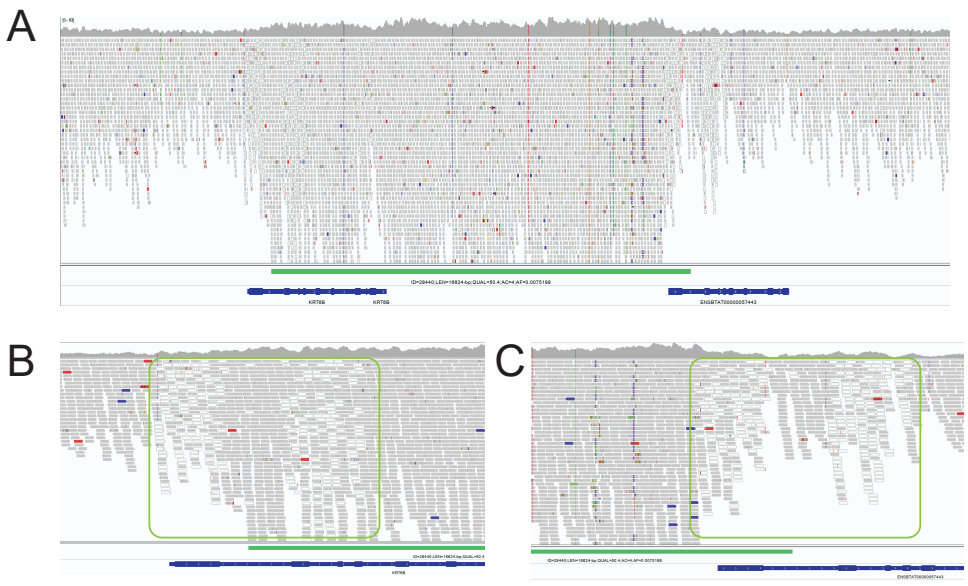


3.10.4. Supplementary Figure 4 Detection of processed pseudogenes. A processed pseudogene event involved insertion of translated mature mRNA. Signatures of processed pseudogenes include the increased read depth at exonic regions, multiple deletions appear in intronic regions, and cluster(s) of discordant reads spanning over the source gene and the insertion site. Furthermore, there may be discordant reads spanning over the source gene and an insertion site. (**A**) The WGS data of a processed pseudogene carrier shows (i) elevated coverage for exons marked with dotted pink line, (ii) many false deletions at intronic regions shown as pink color bars, (iii) discordant reads at the end of the gene body. (**B**) A cluster

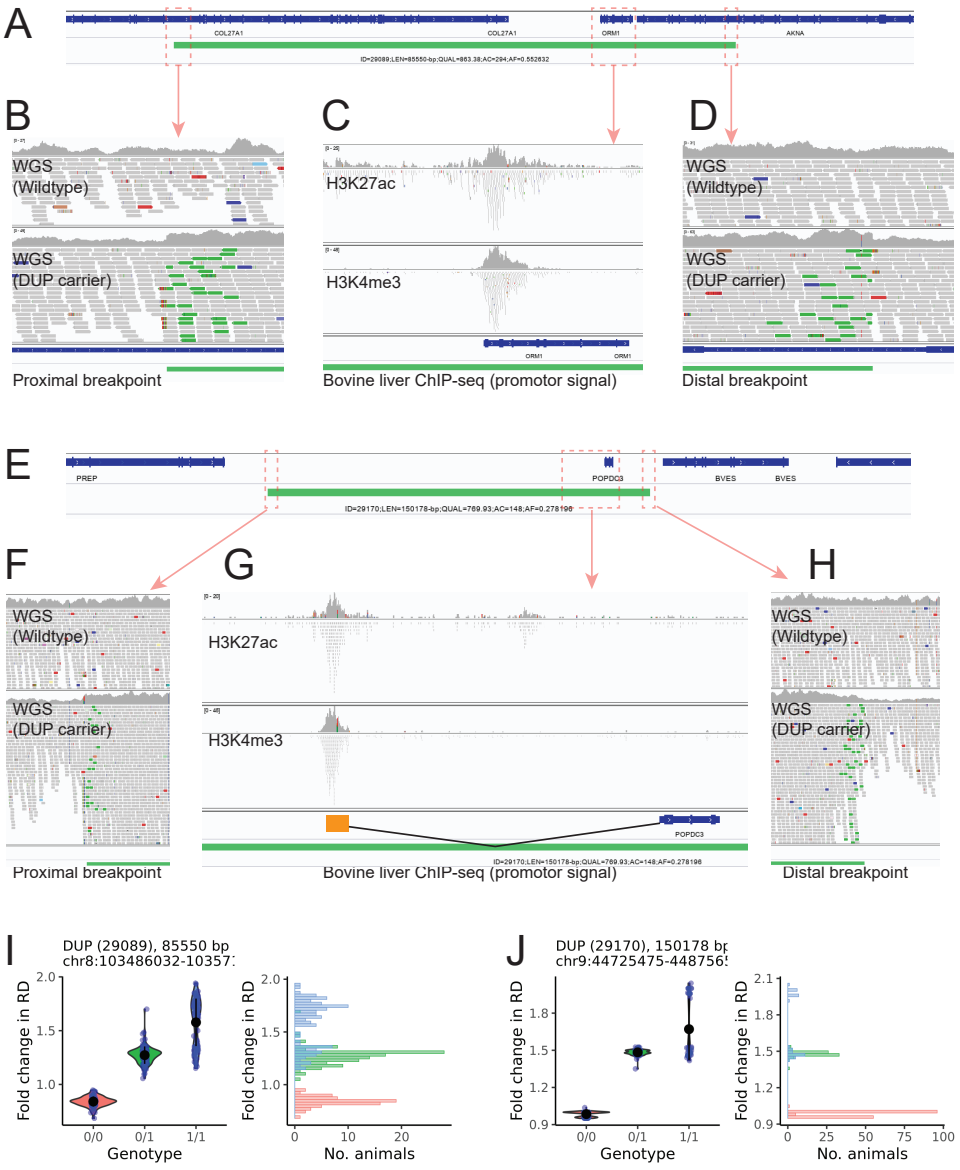
of discordant reads at the end of the gene body (shown in light green) indicates discordant reads mapped to a different chromosome. (**C**) Inspection of the insertion site (bridged by the light green reads shown in panel **B**) suggests true insertion event marked by poly-A signature and target site duplication. (**D**) Circo plot showing the connection between source genes and the corresponding insertion sites.



3.10.5. Supplementary Figure 5 Maximum LD (r^2) between CNV-SNP pairs in WGS data set. We took the maximum LD (r^2) a CNV has between SNPs located within 100-Kb distance. For both deletions and duplications, stringent calls have a higher maximum r^2 than those in the lenient call set. mCNV showed overall lower LD than biallelic CNVs.



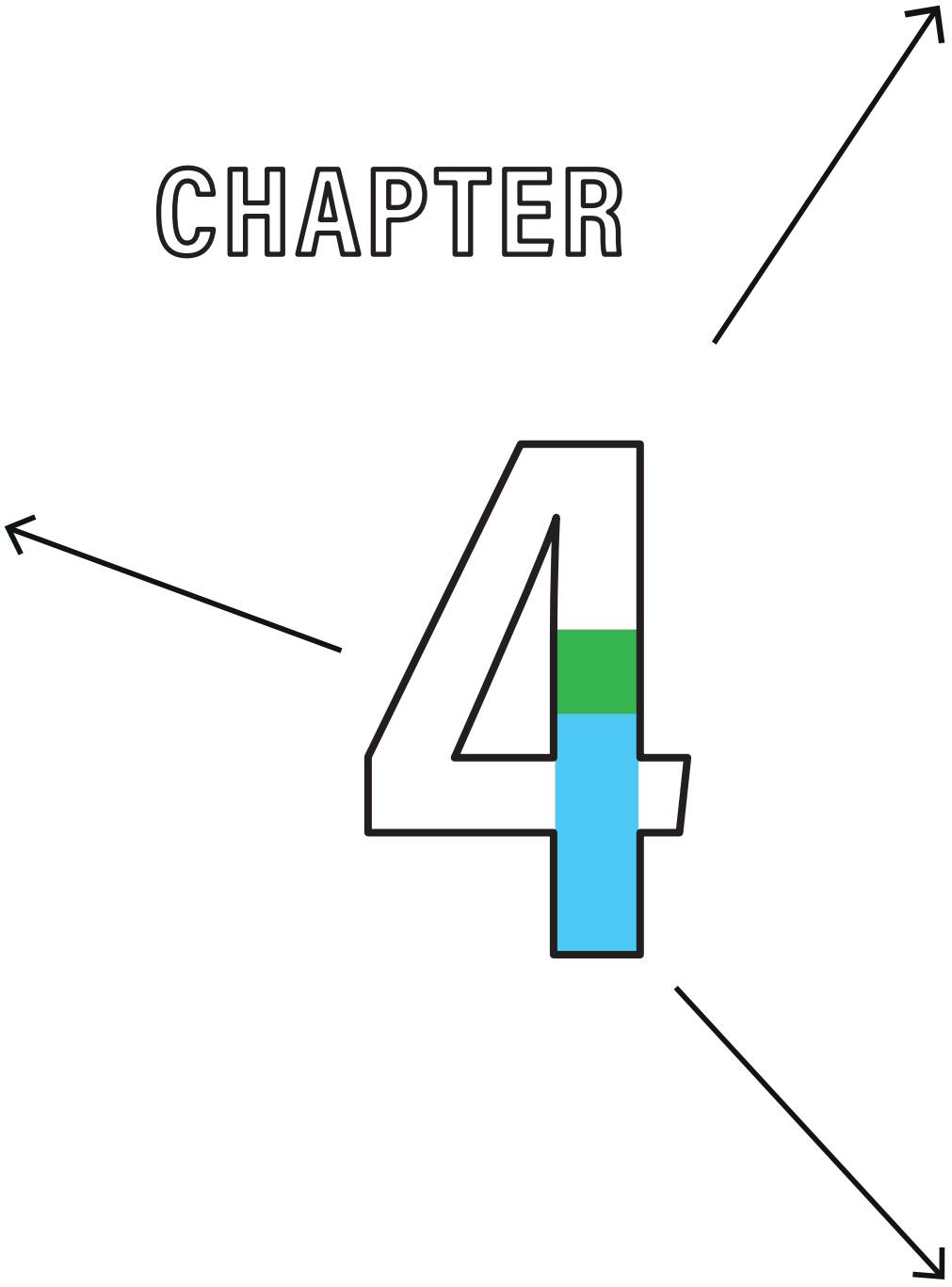
3.10.6. Supplementary Figure 6 Underlying WGS data for a 16-kb mCNV inspected in IGV. (**A**) The underlying WGS data of the 16-kb mCNV was inspected. The green bar indicates the mCNV, and the lower bar shows two keratin genes disrupted by the mCNV. (**B**) The proximal breakpoint of the mCNV. The grey colored bars indicate reads with high mapping quality. The white bars in the green box indicate reads with low mapping quality. (**C**) The distal breakpoint of the mCNV. The legends are identical to panel (**B**).



3.10.7. Supplementary Figure 7 Underlying WGS data for two copy gain duplication events. (**A**) A schematic overview of the ORM1 duplication (marked with green) and the overlapping genes (gene structure shown in dark blue). (**B**) The proximal breakpoint of the ORM1 duplication zoomed-in in IGV. Grey color reads stand for well-aligned ones, whereas green ones stand for discordant ones indicating tandem duplications. (**C**) The bovine liver ChIP-seq data (H3K27ac and H3k4me3) and the gene annotation were inspected. The ChIP-seq signals shown at the start of the ORM1 supports the presence of a liver promoter. (**D**) The distal breakpoint of the ORM1 duplication is zoomed-in IGV. The

legends are identical to panel B. (**E**) A schematic overview on the POPDC3 duplication (marked with green) and the overlapping gene (gene structure shown in dark blue). (**F**) The proximal breakpoint of the POPDC3 duplication is zoomed-in IGV. The legends are identical to panel B. (**G**) The bovine liver ChIP-seq data (H3K27ac and H3k4me3) and the gene annotation were inspected. The ChIP-seq signals indicating a liver promoter appeared upstream of the *POPDC3*. We confirmed an unannotated exon of the *POPDC3* gene coinciding with the ChIP-seq signal (unpublished data). (**H**) The distal breakpoint of the POPDC3 duplication is zoomed-in IGV. The legends are identical to the panel B. (**I**,**J**) QC plots for the ORM1 duplication and POPDC3 duplication, respectively. In both cases, the RD distribution of GT 0/1 and GT 1/1 are overlapping, indicating inaccurate genotyping results.

CHAPTER



A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle

Young-Lim Lee[1*], Haruko Takeda[2], Gabriel Costa Monteiro Moreira[2], Latifa Karim[3], Erik Mullaart[4], Wouter Coppieters[2,3], The GplusE consortium[5], Ruth Appeltant[2], Roel F. Veerkamp[1], Martien A. M. Groenen[1], Michel Georges[2], Mirte Bosse[1], Tom Druet[2], Aniek C. Bouwman[1], Carole Charlier[2]

[1] Wageningen University & Research, Animal Breeding and Genomics, Wageningen, the Netherlands [2] Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Liège, Belgium [3] GIGA Genomics Platform, GIGA Institute, University of Liège, Liège, Belgium [4] CRV B.V., Arnhem, the Netherlands [5] http://www.gpluse.eu/

Published in PLoS Genetics , 2021, 17(7): e1009331, doi:http://dx.doi.org/10.1371/journal.pgen.1009331l

4.1. Abstract

Clinical mastitis (CM) is an inflammatory disease occurring in the mammary glands of lactating cows. CM is under genetic control, and a prominent CM resistance QTL located on chromosome 6 was reported in various dairy cattle breeds. Nevertheless, the biological mechanism underpinning this QTL has been lacking. Herein, we mapped, fine-mapped, and discovered the putative causal variant underlying this CM resistance QTL in the Dutch dairy cattle population. We identified a ~12 kb multi-allelic copy number variant (CNV), that is in perfect linkage disequilibrium with a lead SNP, as a promising candidate variant. By implementing a fine-mapping and through expression QTL mapping, we showed that the group-specific component gene (*GC*), a gene encoding a vitamin D binding protein, is an excellent candidate causal gene for the QTL. The multiplicated alleles are associated with increased *GC* expression and low CM resistance. Ample evidence from functional genomics data supports the presence of an enhancer within this CNV, which would exert *cis*-regulatory effect on *GC*. We observed that strong positive selection swept the region near the CNV, and haplotypes associated with the multiplicated allele were strongly selected for. Moreover, the multiplicated allele showed pleiotropic effects for increased milk yield and reduced fertility, hinting that a shared underlying biology for these effects may revolve around the vitamin D pathway. These findings together suggest a putative causal variant of a CM resistance QTL, where a *cis*-regulatory element located within a CNV can alter gene expression and affect multiple economically important traits.

4.2. Author summary

Clinical mastitis (CM) is an inflammatory disease that negatively influences dairy production and compromises animal welfare. Although one major genetic locus for CM resistance was mapped on bovine chromosome 6, a mechanistic description of this association has been lacking. Herein, we report a 12-kb multiallelic copy number variant (CNV), encompassing a strong enhancer for group-specific component gene (*GC*), as a likely causal variant for this locus. This CNV is associated with high *GC* expression and low CM resistance. We speculate that upregulation of *GC* leads to a large amount of vitamin D binding protein, which in turn, reduces biologically available vitamin D, leading to low CM resistance. Despite the negative effect on CM resistance, the CNV contributes to increased milk production, hinting at balancing selection. Our results highlight how multiplication of a regulatory element can shape economically important traits in dairy cattle, both in favourable and unfavourable directions.

4.3. Introduction

Clinical mastitis (CM) is an inflammation in the mammary glands. This condition is often seen in dairy cattle and the repercussions of CM include production loss, use of antibiotics, and compromised animal welfare (Halasa et al. 2007). CM resistance has a genetic component, with estimated heritabilities ranging between 0.01 and 0.10 (Zwald et al. 2004; Bloemhof et al. 2009; Negussie et al. 2010; Jamrozik et al. 2013; Pritchard et al. 2013). Genome-wide asso-

ciations studies (GWAS) identified several quantitative trait loci (QTL) associated with CM resistance or somatic cell score (SCS), an indicator trait of CM (Zwald et al. 2004; Bloemhof et al. 2009; Negussie et al. 2010; Jamrozik et al. 2013; Pritchard et al. 2013; Sahana et al. 2014). For instance, the six most significant CM resistance QTLs together capture 8.9% of the genetic variance in Danish Holstein Friesian (HF) cattle, underlining the polygenic nature of CM resistance (Sahana et al. 2014). Of these six QTL, the most significant QTL has been mapped near 88 Mb on the Bos taurus autosome (BTA) 6 in various dairy cattle populations (Abdel-Shafy et al. 2014; Sahana et al. 2013; Sodeland et al. 2011; Freebern et al. 2020), including Dutch HF (Veerkamp et al. 2016).

Several fine-mapping studies, using imputed whole genome sequence (WGS) variants, reported non-coding candidate causal SNPs at the group-specific component (*GC*) gene (Sahana et al. 2014; Olsen et al. 2016; Cai et al. 2018; Freebern et al. 2020; Tribout et al. 2020). One of these studies investigated, albeit unsuccessfully, whether one of the non-coding candidate SNP obtained from the GWAS was associated with *GC* expression, leaving the functional mechanisms underlying this association elusive (Olsen et al. 2016). Interestingly, the proposed candidate SNPs showed antagonistic allele effects for milk yield (MY) (the high CM resistance allele was linked to low MY, and vice versa (Sahana et al. 2014; Olsen et al. 2016; Koivula et al. 2005), and one study concluded that a single pleiotropic variant regulates both of the traits (Cai et al. 2020). Furthermore, this locus harbours QTL for many traits including body conformation, fertility, and longevity (Freebern et al. 2020; Jiang et al. 2019; Abo-Ismail et al. 2017; Nayeri et al. 2017; Pausch et al. 2016; Tribout et al. 2020; Xiang et al. 2020), implying pleiotropy, which remains to be investigated.

Until now, *GC*, a gene that encodes the vitamin D binding protein (DBP), has been considered the most promising candidate gene for the CM resistance QTL on BTA 6 (Cai et al. 2018; Olsen et al. 2016). A growing body of literature underpins the importance of DBP, which acts as a macrophage activating factor, modulates immune responses (Gomme and Bertolini 2004), and is central to the vitamin D pathway (Horst et al. 2005). For instance, polymorphisms in *GC* have been shown to cause vitamin D deficiency and inflammatory diseases in humans (Jolliffe et al. 2016). Moreover, a therapeutic use of vitamin D in lactating, CM-infected cows reduced inflammation, implying a link between vitamin D and inflammation (Poindexter et al. 2020; Merriman et al. 2018; Lippolis et al. 2011). Yet, the GWAS lead SNP was not associated with *GC* expression or alternative transcription, leaving the functional mechanism elusive (Olsen et al. 2016). Some researchers hypothesized that a copy number variant (CNV) might be the causal variant underlying this QTL (Sahana et al. 2014). Indeed, a CNV in high linkage disequilibrium (LD) with the GWAS lead SNP was found in the 3' alternative exon of *GC*, however, the functional role of this CNV was not well characterized (Olsen et al. 2016).

In this study, we aimed at fine-mapping the prominent CM resistance QTL in the Dutch HF population, and identifying a candidate causal gene and a variant that may explain the functional mechanism(s) of the QTL. Our findings show that (1) the CM resistance QTL on BTA 6 is confirmed in our Dutch HF population; (2) a 12-kb multi-allelic CNV, encompassing the 3' alternative exon of *GC,* harbours a putative enhancer, which exerts *cis*-regulation on the candidate causal gene, *GC*; (3) a haplotype associated with the CNV allele is strongly selected

for, and the CNV has pleiotropic effects on MY, body conformation and fertility traits. These findings together highlight a functional CNV that contains a *cis*-regulatory element, affecting gene expression and subsequently altering the economically important traits.

4.4. Results

4.4.1. A major CM resistance QTL on BTA 6 segregates in the Dutch HF cattle population

A strong CM resistance QTL has been identified on BTA6 in several cattle populations, including Dutch HF (Freebern et al. 2020; Olsen et al. 2016; Sahana et al. 2014; Cai et al. 2018; Abdel-Shafy et al. 2014; Sahana et al. 2013; Sodeland et al. 2011; Veerkamp et al. 2016). To fine-map this QTL in the Dutch HF population, we first performed an association analysis on BTA6 using 4,142 progeny tested bulls. These animals were genotyped using a custom 16K array, and imputed sequentially, firstly to the Illumina Bovine 50K array, and then to higher density (770K). All analyses were performed according to the Bovine genome assembly UMD3.1 (Zimin et al. 2009) (See Supplementary Table 1 for SNP positions in ARS-UCD1.2 genome). De-regressed estimated breeding values of CM resistance were used as phenotypes in a single SNP association analysis (materials and methods).

As expected, we replicated the strong association signal near BTA 6:88.6 Mb found in Norwegian Red (Olsen et al. 2016), Danish HF (Sahana et al. 2014; Cai et al. 2018), and Dutch HF populations (Veerkamp et al. 2016) ($-\log_{10}P=7.35$; Figure 4.1 A, Supplementary Table 2). This association signal was found downstream of *GC*, a reverse-oriented gene located at BTA 6:88.68-88.74 Mb. We then focused on a 10-Mb window encompassing the QTL (BTA 6:84-93 Mb), using 45,782 imputed WGS level variants present in the window, aiming at (1) confirming the *GC* as a positional candidate gene and (2) identifying a candidate causal variant. This 10-Mb window contained the previously reported CNV, which was in high LD with the GWAS lead SNP (Olsen et al. 2016), and included SNPs located within the CNV in the WGS imputation panel, as they could possibly tag the CNV. The association signal peaked in a 200-Kb region (BTA 6:88.5-88.7 Mb), spanning over *GC* and the region downstream of *GC*. The lead SNP, rs110813063, located at BTA 6:88,683,517 ($-\log_{10}P=9.59$) was ~4 kb downstream of *GC* (Figure 4.1 B). The T allele of the lead SNP (allele frequency (AF) = 0.58), was associated with low CM resistance, whereas the C allele was associated with high CM resistance (AF=0.42).

Finally, a conditional analysis was performed by including the lead SNP as a covariate in the association model. SNPs in the association peak were no longer significant, with the exception of a minor signal at the left side of the association peak (Figure 4.1 C). Our results confirmed the presence of the CM resistance QTL in the Dutch HF population, however, as with previous studies (Cai et al. 2018; Olsen et al. 2016; Sahana et al. 2014; Tribout et al. 2020), the non-coding lead SNP (rs110813063) did not have any evidence of a functional role. There were no coding variants amongst the variant in high LD with the lead SNP ($r^2>0.9$). Provided that the significant association signals do not necessarily indicate causality (Gallagher and Chen-plotkin 2018), we further characterized the lead SNP using WGS data.

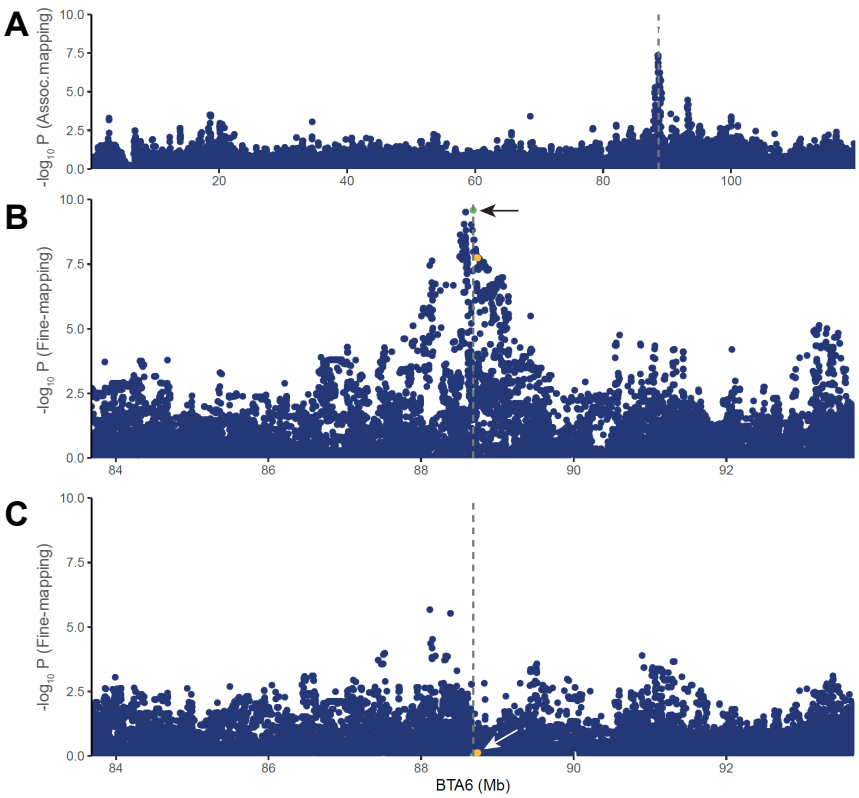


Figure 4.1. Association mapping and fine-mapping of the major clinical mastitis resistance QTL on BTA6. (**A**) Association mapping performed with imputed BovineHD variants on BTA6. The association signal near BTA 6:88.6 Mb, shown in other dairy cattle populations was replicated in the current Dutch HF population. (**B**) Fine-mapping performed with imputed WGS variants in BTA 6:84-93 Mb region. A strong association signal was shown in a 200-Kb region (BTA 6:88.5-88.7 Mb), spanning over *GC* gene. Our lead SNP (rs110813063, marked with an arrow and vertical dotted line) has not been reported as a candidate SNP in other CM fine-mapping studies. CM candidate SNPs from other fine-mapping studies are marked as yellow. (**C**) Conditional analyses including GC CNV as a covariate nullify the association signal (lead SNP is marked with an arrow).

4.4.2. A multi-allelic CNV is in high LD with the lead associated SNP for CM resistance QTL on BTA 6

Our lead associated SNP (rs110813063, hereafter shorten as lead SNP) was located within a ~12 kb CNV encompassing the 14th exon of *GC* (Figure 4.2 A). This CNV, present in both dairy and beef cattle populations (Kommadath et al. 2019), was reported to be in high LD with the candidate SNP for CM resistance QTL in a Norwegian Red population (Olsen et al. 2016). Thus, we hypothesized that the CNV might be the causal variant underlying this QTL. Using our CNV calling pipeline, we characterized the CNV using WGS data of 266 HF animals, exploiting split-read, pair-end mapping, and read depth evidence, confirmed the presence

of the ~12 kb CNV at BTA 6:88,681,767-88,693,545 (hereafter referred to as GC CNV; Supplementary Fig. 1). The 14th exon of *GC*, encompassed by the CNV, is a 3' alternative exon, only accounting for minority of the total *GC* expression (Olsen et al. 2016). The distribution of normalized read-depth of the CNV region suggested a multi-allelic locus, where putative copy number (CN) alleles include CNs 1, 4, 5, and 6 (Figure 4.2 B-C).

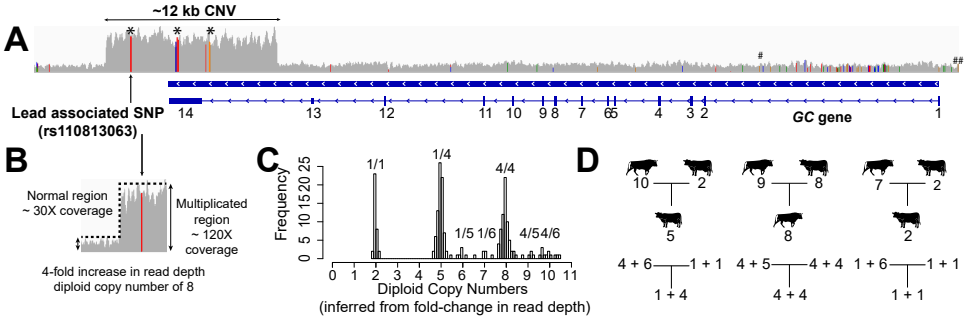


Figure 4.2. Discovery of multiallelic GC CNV using deeply sequenced genomes and familial structure. (**A**) Schematic overview showing the lead associated SNP and ~12 kb CNV overlapping with *GC*. *GC* is a reverse oriented gene, consisting of 14 exons, of which two last exons are non-coding. Five CNV tagging SNPs were present within the GC CNV and marked with black asterisks (the middle asterisk covers three tagging SNPs). Among them, the first SNP, which was also the lead SNP, was in perfect LD with the GC CNV (r^2=1), whereas the rest were in high LD (r^2>0.98). The hash marks at the upstream and the intronic region of *GC* indicate CM resistance candidate SNPs reported by others (Cai et al. 2018; Olsen et al. 2016). (**B**) Sequencing depth difference between the CNV region and normal region was used to infer copy numbers. (**C**) A Histogram of read depth values shows that majority of animals fall into diploid copy number of 2, 5 and 8, and some minor peaks occur at diploid copy number of 6, 7, 9 and 10. Based on this diploid CNs, we inferred haploid CNs of 1, 4, 5, and 6. We showed possible allelic combination(s) above each diploid CN. The diploid CN10 could be comprised of either CN5/CN5 or CN4/CN6; however, our results showed that it was always CN4/CN6. (**D**) Familial information and background haplotypes were used to phase the copy number and thus revealed how the CNV segregates in trios. The upper family tree shown with animal signs stands for diploid copy numbers, and the lower tree shows haploid copy numbers (the phase results of the diploid CNs).

We determined corresponding CN genotypes (e.g., individuals with CN 2 and 6 would be respectively carriers of alleles CN1/CN1 and CN1/CN5) for all but those individuals with 10 copies, that could be either CN4/CN6 or CN5/CN5. In all sequenced duos or trios, these inferred genotypes were compatible with Mendelian segregation rules (Figure 4.2 D), and no genotype incompatibility was observed. Genotypes from relatives of individuals with 10 copies allowed us also to deduce that all carriers of 10 copies were CN4/CN6. Carriers of alleles CN5 or CN6 were restricted to a limited number of families and the segregation of these two alleles perfectly matched haplotype transmission from parents to offspring (Supplementary Fig. 2). An analysis of homozygosity-by-descent (HBD) in all sequenced individuals revealed that alleles CN 4, 5 and 6 share a common haplotype, identical-by-descent for at least 200 kb (≥600 SNPs; Supplementary Table 3). The HBD segments were rare and extremely short at the CNV position in CN 1 individuals, indicating that this allele was associated with different haplotypes.

In summary, we identified four alleles at the GC CNV locus: CN 1 corresponds to a single copy and considered wildtype (Wt) given the high haplotypic diversity, whereas alleles CNs 4-6 correspond to multiple copies (Mul), with four to six copies (Figure 4.3 A). In our population, CN 1 and CN 4 were the most frequent alleles (0.39 and 0.54, respectively), while CN 5 and CN 6 were rare (0.03 and 0.05, respectively; Figure 4.3 A). Furthermore, the GC CNV, coded as a biallelic variant, where CN 1 (Wt) was the reference allele and CNs 4-6 (Mul) were grouped together as the alternative allele, was in perfect linkage (r^2=1) with the lead SNP (rs110813063). The C allele, associated with high CM resistance, tagged CN1, whereas the T allele, associated with low CM resistance, tagged CNs 4-6. Thus, this lead SNP was used as a surrogate marker for the GC CNV in subsequent analyses.

The GC CNV contained five tagging SNPs, including the lead SNP (r^2≥0.98; Figure 4.3 B). These five tagging SNPs showed an allelic imbalance pattern in WGS data of Wt/Mul individuals (Figure 4.3 C) due to a disproportionally high number of reads supporting alternative alleles on the duplication haplotypes. There was no SNP uniquely tagging CNs 4-6 separately, due to their similar and recent haplotypic background.

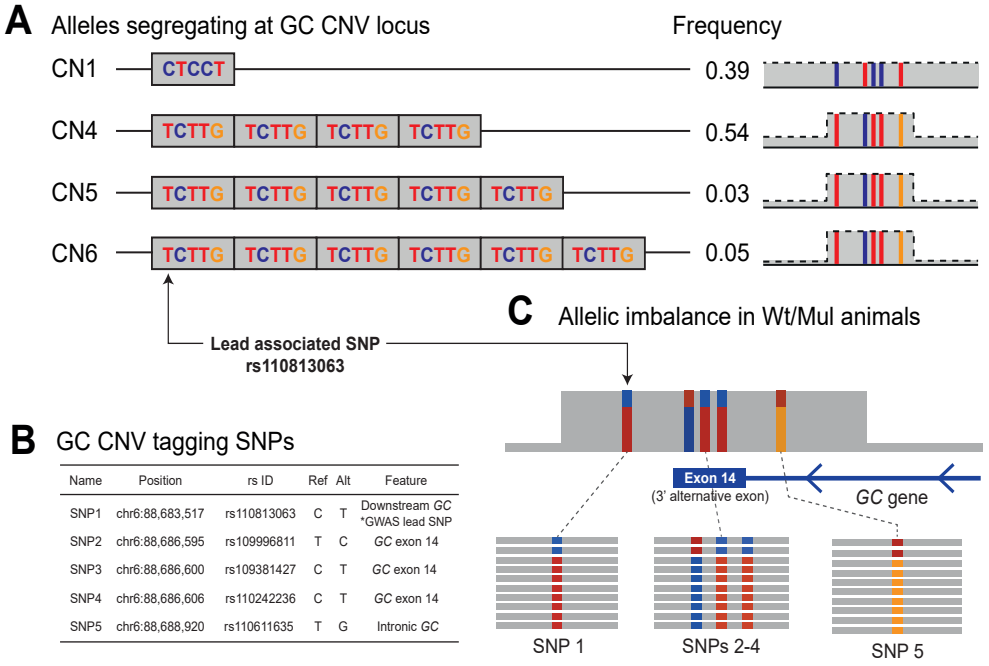


Figure 4.3. Characterization of the GC CNV tagging SNPs and allelic imbalance pattern. (**A**) A schematic overview of four structural haplotypes and the five tagging SNPs inside the GC CNV, shown together with allele frequencies. (**B**) Five GC CNV tagging SNPs, shown with their positions, rs ID, alleles, location within GC. (**C**) Allelic imbalance pattern shown in Wt/Mul animals. Animals will get more supporting reads for alternative alleles for the five CNV tagging SNPs, thus the tagging SNPs will be called as heterozygous but with strong allelic imbalance

4.4.3. Recent positive selection strongly favoured a haplotype harbouring the GC CNV

In livestock breeding, artificial selection for economically important traits (i.e. high CM resistance) potentially drives desired alleles to fixation and removes alleles with negative effects (i.e. low CM resistance) from the population. Thus, it is intriguing that the allele associated with low CM resistance (CNs 4-6) is highly frequent in Dutch HF cattle (combined AF=0.58). We postulated two alternative hypotheses: (1) GC CNV is pleiotropic, conferring positive effects on different traits under selection, contrary to a negative effect on CM resistance, or (2) GC CNV is in high LD with a causal variant of a strongly selected trait, and therefore, genetic hitch-hiking increased the frequency of the low CM resistance allele. In both cases, the GC CNV would be associated with a selected haplotype.

Thus, we scanned BTA 6 for selection signatures based on integrated Haplotype Score (iHS; Gautier et al. 2017; Voight et al. 2006), using WGS haplotypes from the 266 HF animals. Of the two strong signals of selection identified near the 79 and 89 Mb regions ($-\log_{10}P>5$; Supplementary Fig. 3), the latter was only ~200 kb away from the GC CNV, and thus was further inspected (Figure 4.4 A). The extended haplotype homozygosity (EHH), centred at the iHS lead SNP (BTA 6:88,861,709, AF=0.28) revealed a strongly favoured haplotype, which extended outwards further than the non-selected haplotypes (Figure 4.4 B). As expected, CNs 4-6 were located in the strongly selected haplotype, whereas CN 1 was in the non-selected haplotypes. In addition, this finding was in line with our HBD results, where homozygous CNV carriers (CNs 4-6) shared long HBD haplotypes, whereas homozygous non-CNV carriers (CN 1) did not.

These findings supported our hypothesis that a strong positive selection acted upon the region containing the GC CNV. Given the antagonistic effects of the CM resistance QTL on MY (Olsen et al. 2016) and relevance to dairy cattle breeding (Miglior et al. 2017), we deemed MY as a potential target of the selection signature we identified. To confirm whether CM resistance and MY are modulated by the same variant (pleiotropy) or two different variants (LD), the 10-Mb window (BTA 6:84-93 Mb) harbouring the GC CNV was fine-mapped for MY. A strong association signal appeared in 89.08 Mb region, ~400 kb away from the GC CNV ($-\log_{10}P=7.5$; Figure 4.4 A). The MY lead SNP was found at 89,077,838-bp and the G allele was associated with high MY, whereas the T allele was associated with low MY (AF=0.45). Regardless of the ~400-Kb distance, the MY lead SNP and the GC CNV were in high LD ($r^2 = 0.88$). Of note, the iHS lead SNP was located between the association signals for CM resistance and MY (Figure 4.4 A). We re-evaluated the EHH results and found that the strongly selected haplotype harbours CNs 4-6 alleles of the GC CNV and the G allele of the MY lead SNP, implying that the strong selection resulted in low CM resistance and high MY (Figure 4.4 B). We sought to disentangle these two QTL further to elucidate whether they are in pleiotropy or LD. However, due to high LD in the region and limitation in the data set (only 13 out of ~4,000 bulls carrying favourable recombinant haplotypes), this was not possible with the current data.

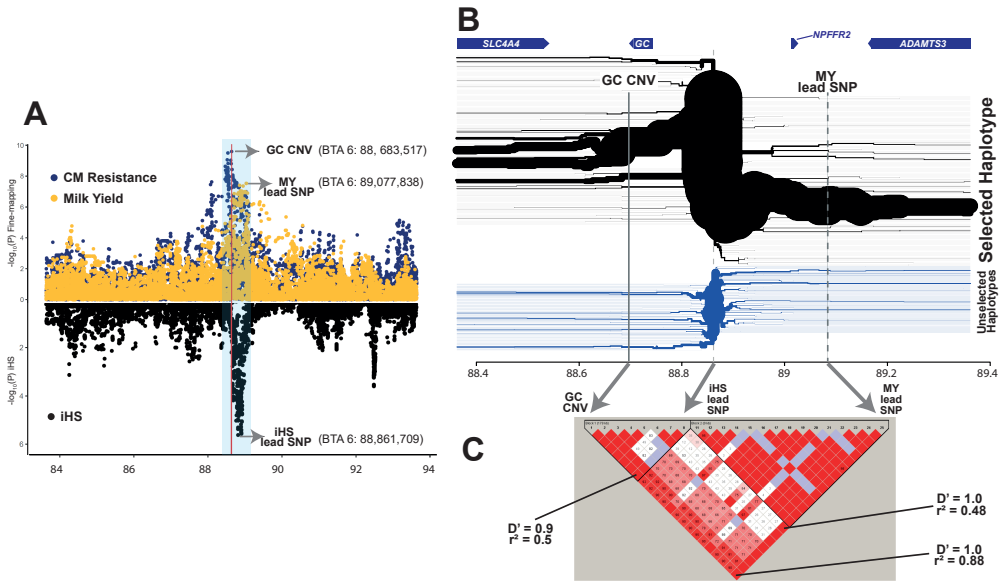


Figure 4.4. Selection signature scan and trait association (clinical mastitis resistance and milk yield) plot. (**A**) A 10-Mb region with a strong selective signature signal was zoomed in (BTA 6: 84-93 Mb). Association mapping results from imputed WGS variants on CM resistance (dark blue) and MY (yellow) are shown in the upper panel; iHS results are shown in the lower panel (black). The CM resistance association peak occurs at the left side of the iHS peak, whereas MY association peak appears on the right side of the iHS peak. The red vertical line marks GC CNV. A 1-Mb region covering GC CNV, iHS lead SNP, and MY lead SNP are marked with translucent blue. (**B**) The extended haplotype homozygosity of the 1-Mb region marked in panel (**A**) is shown, together with four genes annotated in this region (top of the figure). The major haplotype shown in the upper part (black) branches outwards, implying recent positive selection acted upon this haplotype. The non-selected haplotypes, shown in the lower side (blue) rapidly break down from the iHS lead SNP. (**C**) Pairwise D' and r^2 values between GC CNV, iHS lead SNP, and MY lead SNP in the ~4,000 daughter proven bulls. A screenshot made from Haploview software (Barrett et al. 2005)

Additionally, we fine-mapped other dairy cattle traits, in an attempt to identify other potential selection target trait(s). Our results showed that the GC CNV was the lead variant for body condition score (BCS) and calving interval (CI; -$\log_{10}$P=21.4 and 6.6, respectively). Remarkably, GC CNV had a stronger association signal for BCS than it did for MY (Supplementary Fig. 4). The CNs 4-6 allele, which was associated with low CM resistance, was correlated with low BCS (meaning low body fat content) and longer CI (meaning low fertility). The SNP association p-values obtained from either CM resistance and BCS or CM resistance and CI, clearly colocalized, underscoring that these QTL are driven by the same variant, which is likely to exert pleiotropic effects on each of these traits (Supplementary Fig. 3 and Supplementary Table 4).

4.4.4. GC is the most functionally relevant gene underlying the CM resistance QTL on BTA 6

Our fine-mapping results hinted at transcriptional regulation as an underlying mechanism(s) of the CM resistance QTL, as our lead SNP was found in non-coding region. Thus, we mapped *cis*-expression QTL (further referred to as eQTL), to (1) identify shared variant(s) that are driving both local association and eQTL signals, and (2) to corroborate the causality of the candidate gene *GC* in the major CM resistance QTL. Prior to eQTL mapping, we firstly determined the most biologically relevant tissue(s) for our investigation. In the human transcriptome database (Ardlie et al. 2015; Brawand et al. 2011; Lizio et al. 2019; Fagerberg et al. 2014; Papatheodorou et al. 2018; The ENCODE Project Consortium 2012), *GC* is predominantly expressed in the liver, whereas breast tissue showed no expression (Supplementary Table 5). Also, previous dairy cattle transcriptome studies showed that *GC* is expressed in the liver, kidney, and cortex, but not in the mammary gland (Fang et al. 2017; Olsen et al. 2016; Freebern et al. 2020).

Therefore, we performed eQTL mapping within the *cis*-regulatory range (+/- 1Mb region from the GC CNV), using liver RNA-seq data and Bovine HD genotype of lactating HF cows (n=175). Since GC CNV is not in the BovineHD array, the GC CNV genotypes were obtained by (1) imputing BovineHD genotypes to WGS level variants and (2) genotyping the GC CNV directly. Direct genotyping was done by targeting six polymorphic sites (CN 1 as reference allele and CNs 4-6 as alternative allele) in the GC CNV (Supplementary Table 6). The best probe (BTA6:88,683,517) among the six showed 100% compatibility with the imputed GC CNV genotypes, underlining that our imputation approach was robust. Hence, we used the imputed genotypes (BTA6:87-89 Mb) to map eQTL further.

The RNA-seq data was mapped using a reference guided method, where both reference annotated transcripts and novel transcripts can be discovered. Our RNA-seq data detected two different *GC* transcript isoforms: a canonical and an alternative transcript (Figure 4.5 A). The former consisted of 13 exons, not encompassing the GC CNV, and accounted for a majority of overall *GC* expression (˜98%). The latter shared the first 12 exons with the canonical transcript, however, it used an alternative 3' exon, the 14th exon, which is located inside the GC CNV. Expression of the alternative transcript was relatively low (˜2% of the total GC expression). The reference gene set used for transcript assembly included the canonical form, but not the alternative transcript. Thus, *GC* gene-level eQTL mapping was done on the canonical form, and we additionally mapped transcript-level eQTL for the alternative *GC* transcript.

Of the 13 genes annotated within the *cis*-regulatory range of CNV (+/- 1 Mb), *GC* was abundantly expressed (≥5,000 transcripts per million (TPM)) and other genes were either lowly expressed (0.1<TPM<50), or not expressed (Figure 4.5 B). Our gene-level eQTL mapping results discovered highly significant *cis*-eQTL for *GC*. The GC *cis*-eQTL was found in the BTA 6:88.68-88.88 Mb region ($-\log_{10}P>23$; Figure 4.5 D). The GC CNV was one of the top variants within the eQTL peak ($-\log_{10}P=24.4$) and was in high LD with the GC eQTL lead SNP ($-\log_{10}P=25.4$; $r^2=0.88$). P-values for GC expression and CM resistance were highly correlated

(ρ=0.68; Figure 4.5 E), where CNs 4-6 were associated with increased *GC* expression and low CM resistance (Figure 4.5 F).

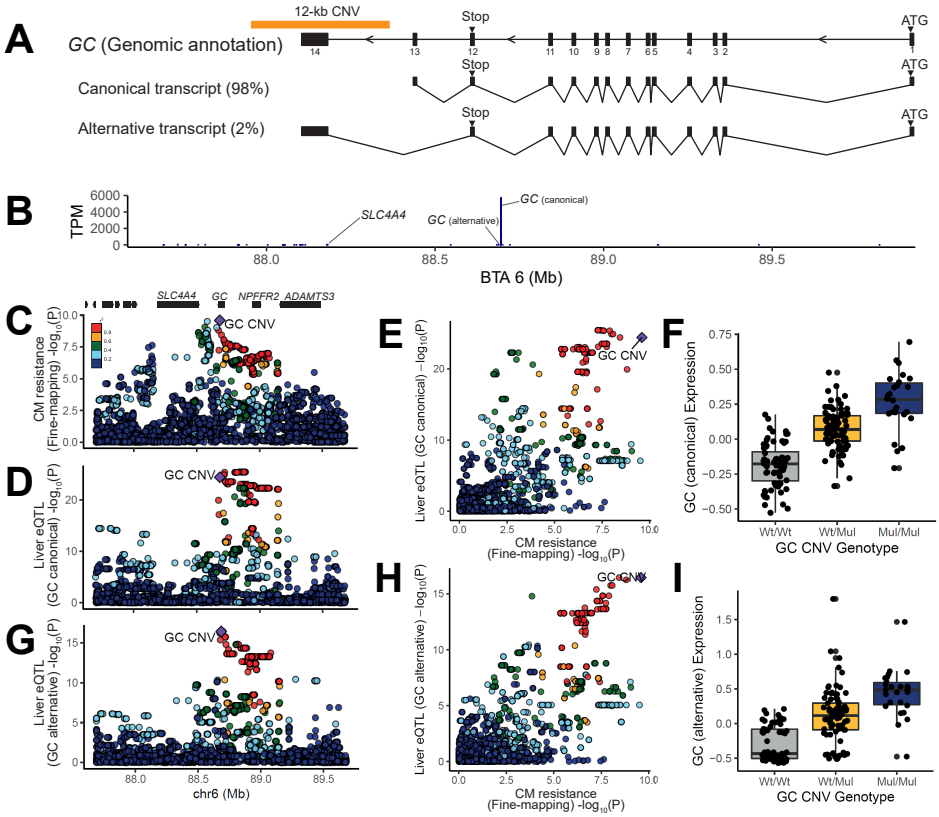


Figure 4.5. eQTL mapping and colocalization of fine-mapping and eQTL mapping results for GC and the non-coding RNA. (**A**) A Schematic overview of the *GC* gene structure and position of the GC CNV. Our data detected two *GC* transcripts, where the canonical form account the majority of the expression (98%) and an alternative form only counting for minor expression (2%). (**B**) eQTL was mapped for the genes located in a 2-Mb bin (BTA6:87.68-89.68). Of the 13 genes annotated in this bin, *GC* showed predominantly high expression (>5,000 TPM), whereas the rest were lowly expressed or not expressed at all. The eQTL were mapped for *GC* and *SLC4A4*. (**C**) CM resistance fine-mapping results were shown for the 2-Mb bin, where eQTL was mapped. The color scale indicates the degree of pair-wise LD (r^2) between the GC CNV and other SNPs. Annotation of genes in this region is drawn as black bars. Six genes on the left part are *AMBN, JCHAIN, RUFY3, GRSF1, MOB1B*, and *DCK*. (**D**) eQTL mapping results for *GC* (canonical transcript). (**E**) P-values obtained from CM resistance fine-mapping and GC eQTL mapping (canonical transcript) were correlated. The GC CNV is located in the right upper corner (ρ=0.68), showing that it is significant for both fine-mapping and eQTL mapping. (**F**) The box plot shows altered *GC* (canonical transcript) expression depending on GC CNV genotypes. (**G**) eQTL mapping result for GC (alternative transcript). (**H**) P-values obtained from CM resistance GWAS and GC eQTL mapping (alternative transcript) were correlated. The GC CNV is located in the right upper corner (ρ=0.74), showing that it is significant for both fine-mapping and eQTL mapping. (**I**) The box plot shows altered *GC* (alternative transcript) expression depending on GC CNV genotypes. Panels C-E, G, H were made with LocusCompare programme (Liu et al. 2019)

Additionally, *cis*-eQTL for Solute Carrier Family 4 Member 4 gene (*SLC4A4*), involved in bicarbonate secretion and associated with renal tubular acidosis, was found in BTA 6:88.67-89.07 Mb ($-\log_{10}P>7$, Supplementary Fig. 5). The lead SNP for *SLC4A4* eQTL (BTA 6:88,672,979, $-\log_{10}P=7.75$) was found ~9 kb away from GC CNV, which also showed high significance ($-\log_{10}P=7.1$). The GC CNV and the *SLC4A4* eQTL lead SNP were in high LD ($r^2=0.99$). P-values obtained from *SLC4A4* eQTL mapping and CM resistance were highly correlated ($\rho=0.82$), and CNs 4-6 were associated with increased *SLC4A4* expression (Supplementary Fig. 5).

Additionally, we mapped a transcript-level eQTL for the alternative *GC* transcript. Intriguingly, the eQTL signal was driven by GC CNV tagging SNPs, followed by the second most significant variant, GC CNV ($-\log_{10}P=16.9$ and 16.4, respectively; Figure 4.5 G). P-values for alternative *GC* transcript expression and CM resistance were correlated even stronger than the canonical transcript ($\rho=0.74$; Figure 4.5 H), where CNs 4-6 corresponded to an increased expression of the alternative *GC* transcript (Figure 4.5 I).

Our results indicate *GC* as a promising candidate gene, given the strong eQTL signal. On the contrary, high LD between GC CNV and the *SLC4A4* eQTL lead SNP ($r^2=0.99$) implied that *SLC4A4* could be a candidate gene. To prioritize between the two candidate genes, *GC* and *SLC4A4*, we used summary data-based Mendelian randomization (SMR) analysis (Zhu et al. 2016), which estimates associations between phenotype and gene expression, aiming at identifying a functionally relevant gene, underlying GWAS hits. Our result showed *GC* to be the only gene whose expression was significantly associated with CM resistance association ($-\log_{10}P=6$), whereas *SLC4A4* was below the statistical threshold ($-\log_{10}P=4.9$; Supplementary Table 7). A subsequent analysis, heterogeneity in dependent instruments (HEIDI; Zhu et al. 2016), was conducted to test whether the significant association between association hit and eQTL shown for *GC* was induced by a single underlying variant or two variants that are in LD (i.e. association hit variant is in LD with eQTL lead SNP). The result suggested a single underlying variant modulating both CM resistance association and *GC* eQTL ($P_{HEIDI}=0.06$). Thus, we confirmed *GC* as the most promising causal gene underlying CM resistance QTL, whose expression affects the CM resistance phenotype.

4.4.5. A putative enhancer located in the *GC* CNV likely modulates the level of *GC* expression

We subsequently exploited epigenomic data sets to infer the functions of candidate variants in the 200-kb *GC* eQTL region (BTA 6:88.68-88.88 Mb; Figure 4.6 A). Bovine liver epigenomic data, interrogating two histone modifications (ChIP-seq for H3K27ac and H3K4me3 marks) (Villar et al. 2015) and open chromatin regions (ATAC-seq) were investigated to infer functional contexts (i.e. active promoters and enhancers).

The ChIP-seq data predicted one active promoter and three active enhancers, overlapping with ATAC-seq peaks, supporting the active status of the regulatory elements along *GC* (Figure 4.6 A). Of the three putative enhancers identified, *GC* harbours two putative enhancers, one in the 1st intron and the other inside the GC CNV. The one located inside the GC CNV (further

referred to as GC CNV enhancer) could be considered highly active, given the strong H3K27 acetylation mark. According to the comparative genomics catalogue of regulatory elements in liver (Villar et al. 2015), the GC CNV enhancer is cattle-specific, whereas the intronic enhancer is conserved between human, mouse, cow, and dog. These two putative enhancers were both supported by corresponding ATAC-seq signals, with a stronger peak shown at the GC CNV enhancer (Figure 4.6 A).

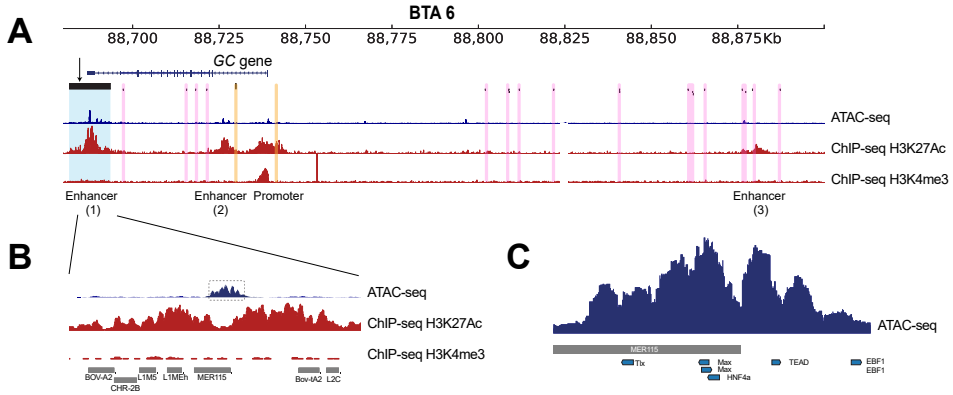


Figure 4.6. Inspection of functional elements near the GC CNV. Functional elements were inspected in *GC* eQTL region using ChIP-seq (H3K27ac and H3K4me3) and ATAC-seq data. The GC CNV genotype of the ATAC-seq sample was Mul/Mul (inferred based on CNV tagging SNP genotypes and read-depth increase in WGS data). The GC CNV genotype of the ChIP-seq sample was unknown (Supplementary Table 9). (**A**) The *GC* eQTL region was zoomed in. In this region, *GC* is the only annotated gene. The GC CNV is marked with the translucent blue, and CM resistance candidate SNPs reported by other studies (Olsen et al. 2016; Cai et al. 2018) are marked with translucent yellow. Other significant eQTL lead SNPs in this region are marked with translucent pink. We overlaid ChIP-seq data to identify putative enhancers and promoters (ChIP-seq tracks; red). Furthermore, liver ATAC-seq data revealed highly accessible chromatin regions, supporting the regulatory elements discovered by ChIP-seq data sets (ATAC-seq tracks; blue). (**B**) We further zoomed in to the ATAC peak within the GC CNV, and discovered that the ATAC peak overlaps with MER 115. The predicted hepatic transcription factor binding sites are marked with translucent grey. (**C**) Transcription factor binding motifs are shown together with the ATAC signal located inside the GC CNV.

Additionally, this strong ATAC signal overlapped with a repetitive element, MER115, from the DNA transposon family, hAT-Tip100 (Figure 4.6 B). Some DNA transposons obtain enhancer function during evolution (Cao et al. 2019), and hAT-Tip100 was shown to function as enhancer in humans (Villar et al. 2015). However, although the human orthologous region harbours MER115, this repeat did not show enhancer activity in humans (The ENCODE Project Consortium 2012), underlining cattle-specific enhancer activity. Finally, we scanned the ATAC peak region inside the GC CNV, searching for transcription factor binding (TFB) motifs using Homer (Heinz et al. 2010). We found strong evidence for five motifs, including transcriptional enhancer factor (*TEAD*) and hepatocyte nuclear factor 4 alpha (*HNF4A*), supporting for the presence of the active enhancer within the GC CNV (Figure 4.6 C).

4.5. Discussion

In this study, we dissected a prominent CM resistance QTL in Dutch HF cattle, by integrating fine-mapping, eQTL mapping, and functional (epi)genomics data. Our findings revealed that the lead variant is the GC CNV, a 12-kb multi-allelic CNV, which harbours a putative *cis*-regulatory element which targets *GC*. This CNV drives both the CM resistance association and the *GC* eQTL signals, underscoring *GC* as the likely causal gene underlying this QTL.

Identifying causal variants from GWAS can be challenging, since, often non-causal SNPs in high LD with the causal variant appear as lead SNPs (Gallagher and Chen-plotkin 2018). Notably, our lead candidate SNP (rs110813063) has not been considered a strong candidate in other fine-mapping studies (Sahana et al. 2014; Cai et al. 2018; Olsen et al. 2016). The function of our lead SNP, located 4-kb downstream of *GC*, was equally elusive, compared to other candidate SNPs that were located in the intronic region of or upstream of *GC* (Sahana et al. 2014; Cai et al. 2018; Olsen et al. 2016; Tribout et al. 2020). Nonetheless, the allelic imbalance pattern in our candidate SNP (Figure 4.3 C), together with a previous report about a CNV in high LD with the GWAS candidate SNP (Olsen et al. 2016), motivated us to hypothesize that the CNV might be the causal variant. As expected, we corroborated that our lead SNP, rs110813063, is a perfect tag SNP of the GC CNV (r^2=1).

There are several explanations why previous fine-mapping studies missed rs110813063. Possibly, this SNP was absent in the GWAS variant set, as the standard SNP quality control (QC) criteria (i.e. removing SNPs with high depth and/or unbalanced allelic ratio) tend to eliminate SNPs inside CNVs. Alternatively, rs110813063 was present, but wrongly genotyped, due to highly disproportional allelic depth (Figure 4.3 C). Hence, one may wonder whether a SNP-based GWAS approach, relying on a stringent QC, is sufficient in identifying causal variants, in a form other than point mutations. The answer might depend on the type of variant: studies showed that most deletions are well captured by tagging SNPs, whereas most duplications and multi-allelic CNVs are poorly tagged (Sudmant et al. 2015; Handsaker et al. 2015). Note that the EHH analysis could clearly discern the GC CNV carrying haplotype (Figure 4.4 B), showing the merit of a haplotype-based approach in delineating trait-associated structural variations, as demonstrated in earlier examples (Zhang et al. 2012; Durkin et al. 2012; Kadri et al. 2014; Mishra et al. 2017). Also, an exploratory check for presence of CNVs might be beneficial for fine-mapping studies.

To connect the discovery from statistical associations to the molecular function, we showed that the association mapping signal was driven by the underlying molecular signal, expression of *GC* (Figure 4.5). Many heritable diseases are manifested in a tissue-specific manner (Hekselman and Yeger-Lotem 2020). Our trait of interest, CM, is manifested in the mammary glands, and hence it was considered a biologically relevant tissue for eQTL mapping. However, both large-scale human transcriptome databases (Supplementary Table 5) and cattle transcriptome studies indicated liver as the main organ of *GC* expression (Freebern et al. 2020; Olsen et al. 2016). This discrepancy shows that prior knowledge of tissue-specific manifestation of a trait

is crucial for elucidating its molecular basis. In cattle, eQTL data was generated from diverse tissues (adrenal gland, blood, liver, mammary gland, milk, and muscle ;Littlejohn et al. 2014, 2016; Kemper et al. 2016; Brand et al. 2016; Lopdell et al. 2017; Leal-Gutiérrez et al. 2020; Van Den Berg et al. 2019), and some studies utilized the data sets in confirming causality of candidate genes (Littlejohn et al. 2014, 2016; Kemper et al. 2016; Lopdell et al. 2017). We expect that the availability of eQTL data sets of diverse tissues and cell types will lead to rapid discovery and confirmation of candidate genes in farm animals in the future.

Bovine epigenomic data sets were utilized to prioritize candidate variants for the CM resistance QTL and eQTL for *GC* expression. The ChIP-seq and ATAC-seq data supported three active enhancers and one active promoter in the region of interest (Figure 4.6 A). Of these regulatory elements, the GC CNV enhancer is considered the most likely causal variant, as multiple copies of enhancers can increase the target gene expression (Ngcungcu et al. 2017). Hence, we propose that altered *GC* expression, mediated via multiplicated enhancers, is likely the key regulatory mechanism underpinning this QTL. Interestingly, candidate causal variants reported by previous studies (Olsen et al. 2016; Cai et al. 2018) were found in the 1st intron of or upstream of *GC*, where an active enhancer and an active promoter were found (Figure 4.6 A). In humans, tissue-specific enhancers outnumber protein coding genes (The ENCODE Project Consortium 2012; Long et al. 2016). Hence, a gene can be regulated by more than one enhancer, and a secondary enhancer is referred to as a shadow enhancer (Scholes et al. 2019). A recent study showed that redundant enhancers function in an additive manner, thus conferring phenotypic robustness (e.g. activities of multiple enhancers act together, thus removal of an enhancer still results in discernible phenotypes; Osterwalder et al. 2018). In light of this finding, we speculate that the two enhancers located in *GC*, might have additive effects on *GC* expression.

Mammalian enhancers evolve rapidly, compared to promoters (Villar et al. 2015). Also, rapidly evolving enhancers were exapted from ancestral DNA sequences and associated with positively selected genes (Villar et al. 2015). The GC CNV enhancer is likely one of these rapidly evolving enhancers, given that (1) it exapted MER115, which does not function as enhancer in other species, (2) it is found in a selective sweep which harbors *GC*, and (3) it is cattle-specific. These findings strongly suggest that utilizing epigenomic data of species other than the one of interest, can be misleading, as it lacks species-specific regulatory elements. This provides a compelling reason to build species-specific epigenome maps, as already embarked upon by the international Functional Annotation of Animal Genomes (FAANG) project (Giuffra and Tuggle 2019; The FAANG Consortium et al. 2015). With this community effort, a wider range of species-specific regulatory elements underlying economically important QTL is expected to be unraveled in the future.

The major CM resistance QTL is known to have antagonistic effects for MY and our results confirmed this (Olsen et al. 2016; Cai et al. 2018). The trade-off between CM resistance and MY suggests that this locus might be under balancing selection (Hedrick 2015). One of the drivers of balancing selection is strong directional selection, which is common in livestock breeding (Georges et al. 2019). Of the two traits of interest, MY has been the primary goal of

dairy cattle breeding (Miglior et al. 2017), and hence it seems plausible to assume that this locus is under balancing selection. Next to this, we had a particular interest in understanding the genetic basis of the antagonistic effects. The genetic modes considered were (1) a single pleiotropic variant affecting two traits and (2) two independent causal variants for each trait, in high LD. In case of pleiotropy, a particular genomic region cannot enhance both traits simultaneously through breeding. On the contrary, in case of LD, selection for recombinant haplotypes, containing favourable alleles for both traits, enables simultaneous improvement on the two traits. Only a small number (0.3%) of the studied animals had recombinant haplotypes (13 out of 4,142 bulls), and hence our data set lacks sufficient power to distinguish LD from pleiotropy. Future studies might consider two approaches for delineating this issue. The first is to exploit a daughter design by obtaining sufficient daughters of the 13 recombinant haplotype carrier Dutch HF bulls (Georges 2007). Another possibility would be to harness data from different breed(s). A dairy cattle breed that has both MY and CM resistance recorded, yet with low LD in the QTL region, would be most useful. Otherwise, a meta-GWAS of multiple cattle populations can aid in distinguishing between LD and pleiotropy.

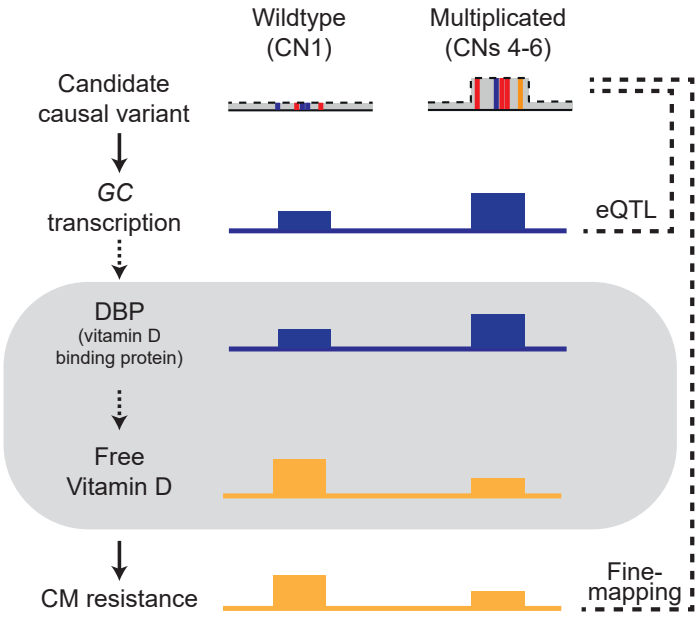


Figure 4.7. Summary of the key findings and hypothesis of physiological aspects linking *GC* expression and CM resistance. A schematic overview summarizing the allele effects of wildtype (CN 1) and multiplicated (CNs 4-6) alleles of the GC CNV, a likely causal variant for the major CM resistance QTL. The two alleles at the GC CNV locus lead to altered *GC* transcription, where the multiplicated alleles correspond to high *GC* expression. On the bottom shows the phenotypic association between the GC CNV and CM resistance, where the multiplicated allele is associated with low CM resistance. Finally, the area marked with grey shade shows our hypotheses that the amount of DBP is positively related with the GC expression. Further, we speculated that the amount of DBP and free vitamin D is inversely correlated, as long as vitamin D is bound by DBP, it is not biologically available. The solid arrows indicate the relations based on our findings. The dotted arrows indicate the relations based on our speculation.

We attempted to integrate the findings from the current study and further speculate on how *GC* expression and CM resistance are linked, assuming that the level of *GC* expression and DBP is correlated; meaning absence of post-transcriptional/translational regulation (Figure 4.7). DBP binds to and transports vitamin D (Daiger et al. 1975), yet it plays additional roles in bone development, fatty acid transport, actin scavenging, and modulating inflammatory responses (see Gomme and Bertolini 2004 and Bouillon et al. 2020 for review). To start with, DBP modulates immunity as a macrophage activating factor (DBP-MAF), when deglycosylated via glycosidases of T- and B- cells (Yamamoto and Kumashiro 1993; Swamy et al. 1997) and therefore enhances the immune response. Accordingly, we could hypothesize that CNV carriers have larger amounts of DBP, and subsequently improved immunity, leading to higher CM resistance. However, this hypothesis contradicts with our findings, as CNV carriers were shown to have lower CM resistance. Alternatively, DBP regulates the amount of freely circulating vitamin D metabolites (Bikle and Schwartz 2019). According to the free hormone hypothesis, only free vitamin D metabolites are able to cross the cell membrane, and are thus biologically available (Chun et al. 2014). In humans, only 0.03 % of 25(OH)D, an indicator of vitamin D, is free, whereas the majority of vitamin D is bound either to DBP (85%) or to albumin (15%) (Bikle and Schwartz 2019). Under these circumstances, CNV carriers, having a large pool of DBP, can be hypothesized to have lower levels of biologically available vitamin D, as postulated in human studies (Sinotte et al. 2009; Lauridsen et al. 2005). Thus, presumably, if CNV carriers have low amounts of free vitamin D, which may be termed as 'vitamin D deficiency', low CM resistance in these animals seems like a logical consequence (Figure 4.7).

Although vitamin D has been considered crucial in bone health (Horst et al. 2005), a recent review showed an inverse correlation between vitamin D concentrations and an extensive range of ill-health outcomes in humans (Autier et al. 2014). Given the pervasive associations between vitamin D and health conditions, there may be other associated traits induced by vitamin D deficiency. Indeed, GC CNV showed strong associations with BCS and CI, which are known to be negatively correlated with MY (e.g. cows with low BCS have longer CI and higher MY (Berry et al. 2003; Pryce et al. 2000), indicating pleiotropy (Supplementary Fig. 4). The CNs 4-6, which were associated with low CM resistance, were associated with longer CI, meaning poor fertility. CI, a measure of duration from one calving to the next, is an indicator of fertility issues such as perturbations in the oestrous cycle, and/or anovulation (Santos et al. 2016). Human studies reported ovarian disfunction induced by vitamin D deficiency (Irani et al. 2014). This finding provides convincing evidence that the GC CNV might be the underlying variant inducing vitamin D deficiency and suboptimal female fertility. Furthermore, pleiotropic effects of the GC CNV indicated that CNs 4-6 were associated with low CM resistance and low BCS. This finding fits with the results found in HF cattle where low BCS was genetically correlated with high disease incidence (Dechow et al. 2004), although the causal relationship between CM resistance and BCS remains unknown. Intriguingly, human studies showed contradictory results: obesity, a condition analogous to high BCS, predisposes patients to vitamin D deficiency, leading to disease susceptibility (Chun et al. 2014). These opposing consequences of body composition traits possibly hint at an 'optimum' body fat amount, where an organism can function well without compromising its health.

4.6. Conclusions

In this study, we dissected the major CM resistance QTL on BTA 6, integrating fine-mapping, CNV calling, eQTL mapping, and functional prioritization of candidate variants. We revealed a multi-allelic CNV harbouring a strong enhancer targeting *GC*, as the likely causative variant. Our findings revealed that the candidate causal gene *GC* is likely regulated by an enhancer located in the GC CNV. We speculate that GC CNV carriers which were shown to have high *GC* expression would have a larger amount of DBP, and by extension, low amount of biologically available vitamin D. This physiological condition probably puts animals into a state comparable to vitamin D deficiency and leads to low CM resistance. Moreover, we report evidence of pleiotropic effects of the GC CNV for other economically important traits as BCS, CI, and MY, which likely revolves around the vitamin D pathway. The current study provides a novel example of multiplicated *cis*-regulatory elements playing pleiotropic roles on various polygenic traits in dairy cattle.

4.7. Materials and methods

4.7.1. Ethics statement

RNA-seq data used for eQTL mapping was obtained from liver biopsy samples collected from ~14 day post-partum HF cows (n=178). The procedures had local ethical approval and complied with the relevant national and EU legislation under the European Union Regulations 2012 (S.I. No. 543 of 2012). The institutes involved in samples collection and the respective local ethical approval information is as following: 1) University College Dublin, approved by University College Dublin Animal Research Ethics Committee (approval number: AE18982/P046), 2) Agri-Food and Biosciences Institute, approved by Animal Welfare Ethical Review Body of Livestock Production Science Branch (approval number: PPL2754), 3) Aarhus University, approved by Danish Veterinary and Food administration Animal Experiments Inspectorate (approval number: 2014-15-0201-00282), 4) Walloon Agricultural Research Centre, approved by ethical commission of Liège University (approval number: 14-1617), and 5) Leibniz Institute for Farm Animal Biology, approved by the State Office for Agriculture, Fishery and Food safety of Mecklenburg-Western Pomerania (approval number: MV 7221.3-1.1-053/13).

4.7.2. Whole genome sequencing and variant discovery

4.7.2.1. Whole genome sequence data

The genomes of 266 Dutch HF animals were sequenced. These 266 animals were closely related animals, where 240 were forming parents-offspring trios. The biological materials were either from sperm (males) or whole blood (females and males). Whole genome Illumina Nextera PCR free libraries were constructed (550bp insert size) following the protocols provided by the manufacturer. Illumina HiSeq 2000 instrument was used for sequencing, with a paired end protocol (2x100bp) by the GIGA Genomics platform (University of Liège). The data was aligned using BWA mem (version 0.7.9a-r786) (Li 2013) to the bovine reference genome UMD3.1 and converted into bam files using SAMtools 1.9 (Li and Durbin 2009). Subsequently, the bam files

were sorted and PCR duplicates were marked with Sambamba (version 0.4.6) (Tarasov et al. 2015). All samples had minimum mean sequencing depth of 15X and the mean coverage of the bam files was 26X.

4.7.2.2. SNP calling and imputation panel construction

Variant calling was done using GATK Haplotype caller in N+1 mode. We applied Variant Quality Score Recalibration (VQSR) at truth sensitivity filter level of 97.5 to remove spurious variants. Using the trusted SNP and indel data sets which are explained elsewhere (Kadri et al. 2016), we observed that the VQSR step filtered out GC CNV tagging SNPs, possibly due to the overwhelmingly high depth in this region. Yet, GC CNV tagging SNPs were considered crucial in our research, as we considered that they could tag the GC CNV. Therefore, the genotypes of the GC CNV tagging SNPs obtained from the raw variant calling format (VCF) file were inserted in the VQSR filtered VCF file. While inspecting the CNV tagging SNPs, we discovered genotyping errors (three errors among 5 tagging SNPs of 266 individuals), where Ref/Alt was wrongly genotyped as Alt/Alt, due to severe allelic imbalance (where the number of alternative allele supporting reads is predominantly high). Using parent-offspring relationships available in our data set, we confirmed that these were true errors, and hence they were manually corrected. Finally, the VCF file of sequence level variants in BTA 6:84-94Mb, containing manually corrected GC CNV tagging SNPs was obtained. This VCF file, consisting of 45,820 variants of 266 animals was used as an imputation panel for fine-mapping analyses. For analyses related to haplotype segregation among the 266 sequenced animals, haplotype sharing among carriers and identification of selection signatures, we further applied stringent variant filters on the VCF file as explained in Kadri et al (2018) to conserve only highly confident variants and to remove spurious genotyping and map errors.

4.7.3. QTL mapping and fine-mapping analysis

4.7.3.1. Phenotype and genotype data

We obtained BTA6 genotype and phenotypic data of 4,142 progeny tested HF bulls from Dutch HF cattle breeding programme (CRV B.V., Arnhem, the Netherlands). The phenotype data was consisting of 60 traits, which are routinely collected in the breeding programme, including clinical mastitis resistance (Supplementary Table 8). CM resistance was recorded as a binary trait, depending on the disease status registration done by farmers and on somatic cell count level of routine milk recording samples (CRV 2020). The estimated breeding values (EBVs; obtained based on BLUP as published in the national evaluation by cooperation CRV (Arnhem, the Netherlands; CRV 2020) were de-regressed to correct for the contribution of family members and de-regressed EBVs were used as phenotypes in an association analysis. The effective daughter contributions of the 4,142 bulls for CM resistance ranged between 25 and 971.3, with an average of 204.3 and a standard deviation of 217. The 4,142 bulls were genotyped with a low density genotyping array (16K). Afterwards, the 16K genotype data was imputed in two steps, firstly to 50K density, based on two private versions of a Bovine 50K genotyping array, and the panel was consisting of 1,964 HF animals. It was further imputed to a higher density, using a panel of 1,347 HF animals genotyped with Illumina BovineHD Bead-

Chip (770K). Subsequently, this genotype data was imputed to sequence level for the 10-Mb QTL region on BTA6 using the HF WGS imputation panel described above. The imputation was done with Beagle 4 (Browning et al. 2018), and variants with low minor allele frequency (MAF<0.025) and low imputation accuracy (allele R^2<0.9) were filtered out.

4.7.3.2. Association model and conditional analysis

To confirm the presence of CM resistance QTL on BTA 6 in the Dutch HF population, we performed association mapping on BTA6 using imputed high density genotypes (28,669 SNPs on BTA6). The association mapping was performed SNP-by-SNP with a linear mixed model in GCTA (Yang et al. 2011). The following model was fitted :

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{b} + \mathbf{g} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of phenotypes (de-regressed EBVs), $\mathbf{1}$ is a vector of ones, μ is the overall mean, $\mathbf{X}$ is a vector of SNP genotypes coded as biallelic variant (0, 1, or 2), $\mathbf{b}$ is the additive effect of the SNP which is being tested for association, $\mathbf{g}$ is the polygenic effect captured by a genomic relationship matrix (GRM; random effect), and $\mathbf{e}$ is the residual. The GRM was built with GCTA, based on 50K genotypes to avoid losing statistical power by including causal markers (Yang et al. 2014). The model assumed equal residual variances. A SNP was regarded significantly associated with CM, when -$\log_{10}$P value was above 6.45 (chromosome-wide Bonferroni multiple-testing correction for 28,669 tests), at a nominal significance of p = 0.01. Subsequently, we repeated the association analysis in the BTA 6:84-94 Mb region, using the imputed sequence level variants (n=45,820) to fine-map the QTL. The association model used was same as described above. Finally, to test if the lead SNP explained the QTL signal completely we ran a conditional association analysis for the imputed sequence variants in BTA 6:84-94 Mb region with the lead SNP as a covariate in the model.

4.7.4. CNV discovery and characterization of GC CNV

The CNV was called from the WGS data of 266 animals, using the Smoove pipeline (https://github.com/brentp/smoove). This pipeline collects split and discordant reads using Samblaster (Faust and Hall 2014), and then calls CNVs using Lumpy (Layer et al. 2014). Afterwards, the CNV sites were genotyped using SVTyper (https://github.com/hall-lab/svtyper). Additionally, we used Duphold (Pedersen and Quinlan 2019) to calculate read depth of the CNs at the GC CNV locus. Duphold exploits read depth between copy number variable regions and normal regions and calculates the ratio between these two. The integer diploid CNs obtained based on the Duphold coverage ratio values were assigned to the 266 animals. The CN distribution showed peaks at diploid CNs of 2, and 5-10, implying four haploid CNs (1, 4, 5, and 6) segregating at the GC CNV locus. Thus, possible haplotypic combinations for the diploid CNs would be: 2 (1/1), 5 (1/4), 6 (1/5), 7 (1/6), 8 (4/4), 9 (4/5), 10 (4/6 or 5/5). There was only one haplotypic combination possible for all the diploid CNs, except 10, were either (4/6) or (5/5) could form diploid CN 10. We confirmed the true presence of these CN alleles by taking advantage of our family structure. First, we verified that observed genotyped followed the Mendelian segregation rules. Next, we also checked that CN alleles transmission within the pedigree was in agreement with haplotype transmission. Haplotypes were reconstructed using

familial information and linkage information using LINKPHASE3 programme (Druet and Georges 2015). The program estimates also, at each marker position and for each parent-offspring pair, the probability that a progeny inherited the paternal or the maternal haplotype for its parents.

To study the relationship between different CN alleles, we estimated homozygous-by-descent (HBD) probabilities at the CNV locus, and measured length of identified HBD segments. To that end, we ran with RZooRoH (Bertrand et al. 2019) a multiple HBD-class model described in (Druet and Gautier 2017), with four HBD classes and one non-HBD class with rates equal to 10, 100, 1000, 10,000 and 10,000, respectively. As this approach compares haplotypes within individuals, it does not require haplotype reconstruction and is not affected by eventual phasing errors.

4.7.5. Selection signature analyses

The BTA 6 was scanned for haplotype based selection signatures observed in the sequence variants from the 266 animals. We used the integrated haplotype homozygosity score (iHS; Voight et al. 2006) using 'rehh' R package (Gautier et al. 2017) for within-population analysis of recent selection signatures. In order to unravel the selection target trait(s) we identified a number of traits within the routinely collected catalogue of 60 traits showing QTL signals in the 84-94Mb region on BTA 6 based on 16K GWAS results (EuroGenomics custom SNP chip; Boichard et al. 2018). Hence, we performed GWAS for these traits (BCS, CI, and MY) in BTA 6:84-94Mb region based on imputed WGS variants to fine-map those QTL. The input files and the association model for the GWASs were the same as described above, except that phenotypes used de-regressed EBVs of BCS, CI, and MY, respectively. The GWAS results of these traits were plotted against the GWAS result of CM resistance to characterize the colocalization of QTL signals, using the R package "LocusComparer" (Liu et al. 2019).

4.7.6. eQTL mapping and Summary data–based Mendelian randomization analysis

4.7.6.1. RNA-seq data and eQTL mapping

We used RNA-seq data produced by the GplusE consortium (http://www.gpluse.eu/; EBI ArrayExpress: E-MTAB-9348 and 9871; Wathes et al. 2021). RNA-seq libraries were constructed using Illumina TruSeq Stranded Total RNA Library Prep Ribo-Zero Gold kit (Illumina, San Diego, CA) and sequenced on Illumina NextSeq 500 sequencer with 75-nucleotide single-end reads to reach average 32 million reads per sample. The reads were aligned to the bovine reference genome UMD3.1 and its corresponding gene coordinates from UCSC as a reference using HISAT2 (Kim et al. 2019). Transcript assembly was conducted with StringTie (Pertea et al. 2016), using reference-guided option for transcript assembly, which enables discovery of novel transcripts that are not present in the reference gene set. Reads were counted at gene- or transcript-level using StringTie. After data normalization using DESeq2 (Love et al. 2014), we performed principal component (PC) analyses and removed outliers (PC>3.5 standard deviations from the mean for the top four PCs, n=2). Subsequently, the gene expression levels were corrected with Probabilistic Estimation of Expression Residuals (PEER; Stegle et al. 2012).

All animals were genotyped using Illumina BovineHD Genotyping BeadChip (770K). The genotype data was imputed to WGS level, using the imputation panel explained above, using Beagle 4 (Browning et al. 2018) and variants with low minor allele frequency (MAF<0.025) and low imputation accuracy (allele R2<0.9) were filtered out. Finally, the PEER corrected normalized gene expression was associated with the imputed WGS variants for 175 samples, using a linear model in R package "MatrixEQTL" (Shabalin 2012).

4.7.7. Prioritizing the causal gene

We prioritized the most functionally relevant gene for the CM resistance QTL on BTA 6, using SMR (version 1.03; Zhu et al. 2016), to estimate association between phenotype and gene expression. Input data required were summary statistics from CM resistance fine-mapping and eQTL mapping results explained above. Additionally, the program requires plink format genotype data to analyse LD in the region of interest, for which the imputed WGS variants (BTA 6:84-94 Mb) of 4,142 bulls were used. A subsequent analysis, heterogeneity in dependent instruments (HEIDI), was conducted to test whether the significant association between lead hit and eQTL shown for GC was induced by a single underlying variant or two variants that are in LD (i.e. lead associated variant is in LD with eQTL lead SNP). The statistical thresholds for SMR (-log10P>5) and HEIDI (P>0.05) were benchmarked from the original paper (Zhu et al. 2016).

4.7.8. Functional genomics assay data

The human transcriptome data bases were examined to find out in which tissue GC is highly expressed via Ensembl website (release 101;Hunt et al. 2018). We downloaded liver ChIP-seq data (H3K27ac and H3K4me3) generated from four bulls from ArrayExpress (E-MTAB-2633; Villar et al. 2015). This ChIP-seq data was aligned to the bovine reference genome UMD3.1 using Bowtie2 (Langmead and Salzberg 2012) and peaks were called using MACS2 (Zhang et al. 2008). A catalogue of mammalian regulatory element conservation (https://www.ebi.ac.uk/research/flicek/publications/FOG15) was used to infer the conservation of the regulatory elements predicted from the ChIP-seq data sets. Next, ATAC-seq data was explored to see if chromatin accessible regions coincided with the histone marks obtained from the ChIP-seq data. We obtained a two weeks old male HF calf's liver ATAC-seq data from the GplusE consortium (ArrayExpress accession number: E-MTAB-9872). Data was analysed by following the ENCODE Kundaje lab ATAC-seq pipeline (https://www.encodeproject.org/pipelines/ENCPL792NWO/). Sequences were trimmed using Trimmomatic (Bolger et al. 2014) and aligned on the bovine reference genome UMD3.1 using Bowtie2 (Langmead and Salzberg 2012). After filtering out low quality, multiple mapped, mitochondrial, and duplicated reads using SAMtools (Li et al. 2009) and the Picard Toolkit (http://broadinstitute.github.io/picard/), fragments with map length 146 bp were kept as nucleosome-free fraction. Genomic loci targeted by TDE1 were defined as 38-bp regions centered either 4 (plus strand reads) or 5-bp (negative strand reads) downstream of the read's 5'-end. ATAC-seq peaks were called using MACS2 (Zhang et al. 2008) (narrowPeak with options --format BED, --nomodel, --keep-dup all, --qvalue 0.05, --shift -19, --extsize 38). We inspected the presence of an enhancer in the

human orthologous region of the GC CNV, using ENCODE data (The ENCODE Project Consortium 2012) in the UCSC genome browser (Kent et al. 2002). Transcription factor (TF) binding motifs in ATAC-seq peak regions were discovered using Homer (Heinz et al. 2010). The TF motifs that are expressed in bovine or human liver are kept and shown in the figure (de Souza et al. 2018; Liu et al. 2008).

4.7.9. EuroGenomics custom array genotyping

We attempted to directly genotype the GC CNV in a biallelic mode (CN 1 as reference allele and CNs 4-6 as alternative allele). Probes targeting six polymorphic sites were designed in Illunima DesignStudio Custom Assay Design Tool were added to custom part of the EuroGenomics array (Boichard et al. 2018; Supplementary Table 6). DNA was extracted from the same biological material used for RNA-seq used for eQTL mapping (described above), and genotyped for the EuroGenomics array by the GIGA Genomics platform (University of Liège). The SNP genotypes were assessed using Illumina GenomeStudio software. Average call rate per probe was calculated to assess the quality of the probes. Finally, genotypes obtained from the highest quality (BTA6:88,683,517) was compared to the imputed GC CNV genotypes.

4.8. Acknowledgements

The Dutch HF whole genome sequence population data set was funded by the DAMONA ERC advanced grant to MG. CC and TD are senior research associates from the Fonds de la Recherche Scientifique–FNRS (F.R.S.-FNRS). The authors are grateful to Elias Kaiser for discussion on the Vitamin D pathway, Ole Madsen for discussion on functional genomics data, Martijn Derks for providing advice on CNV calling pipeline, and the GplusE consortium for providing data for eQTL mapping and ATAC-seq. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

4.9. Author contribution

Conceptualization: Michel Georges, Carole Charlier, Aniek Bouwman

Data Curation: Wouter Coppieters, Latifa Karim, Gabriel Costa Monteiro Moreira, Haruko Takeda, Erik Mullaart, Ruth Appeltant

Formal analysis: Young-Lim Lee, Tom Druet, Haruko Takeda

Funding Acquisition: Roel Veerkamp, Michel Georges

Investigation: Young-Lim Lee, Tom Druet, Haruko Takeda

Methodology: Aniek Bouwman, Tom Druet, Haruko Takeda

Resources: Erik Mullaart

Supervision: Carole Charlier, Mirte Bosse, Tom Druet, Aniek Bouwman, Michel Georges, Martien Groenen, Roel Veerkamp

Visualization: Young-Lim Lee

Writing – original draft preparation: Young-Lim Lee
Writing – review & editing: Young-Lim Lee, Carole Charlier, Mirte Bosse, Tom Druet, Aniek Bouwman, Michel Georges, Martien Groenen, Roel Veerkamp, Wouter Coppieters, Latifa Karim, Gabriel Costa Monteiro Moreira, Haruko Takeda, Erik Mullaart, Ruth Appeltant

4.10. Data reporting

Genome sequence data of the CM resistance QTL region (BTA 6:84-93 Mb) of the 266 Dutch Holstein Friesian animals are deposited in the European Nucleotide Archive under accession PRJEB45439/ERP129554. The RNA-seq data is deposited under EBI ArrayExpress accession E-MTAB-9348 and 9871. The genotype data used for eQTL mapping is available in the supporting information. The ATAC-seq data is deposited under EBI ArrayExpress accession E-MTAB-9872. Genotype and phenotype data of 4,142 bulls used for QTL mapping on BTA 6 are available upon request directed to CRV B.V. (chris.schrooten@crv4all.com) and require a material transfer agreement.

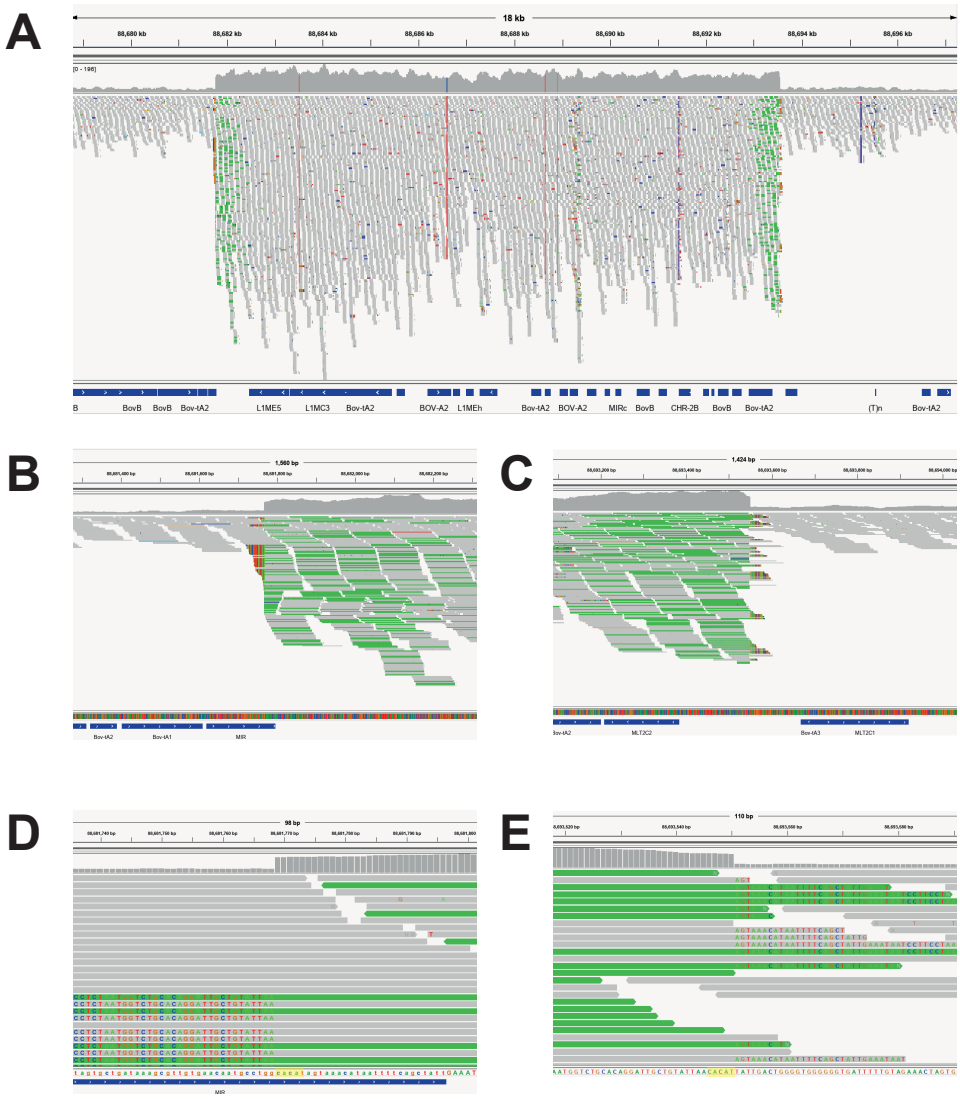
4.11. Research Funding

YLL, ACB, MB, MAMG, and RFV are financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. This work was supported by grants from the European Research Council (Damona to MG; award number: ERC AdG-GA323030), and the EU Framework 7 program (GplusE to MG and HT; award number: 613689). GCMM is post-doctoral fellow of the H2020 EU project BovReg (Grant agreement number: 815668). The funders had no role in study design or decision to publish.

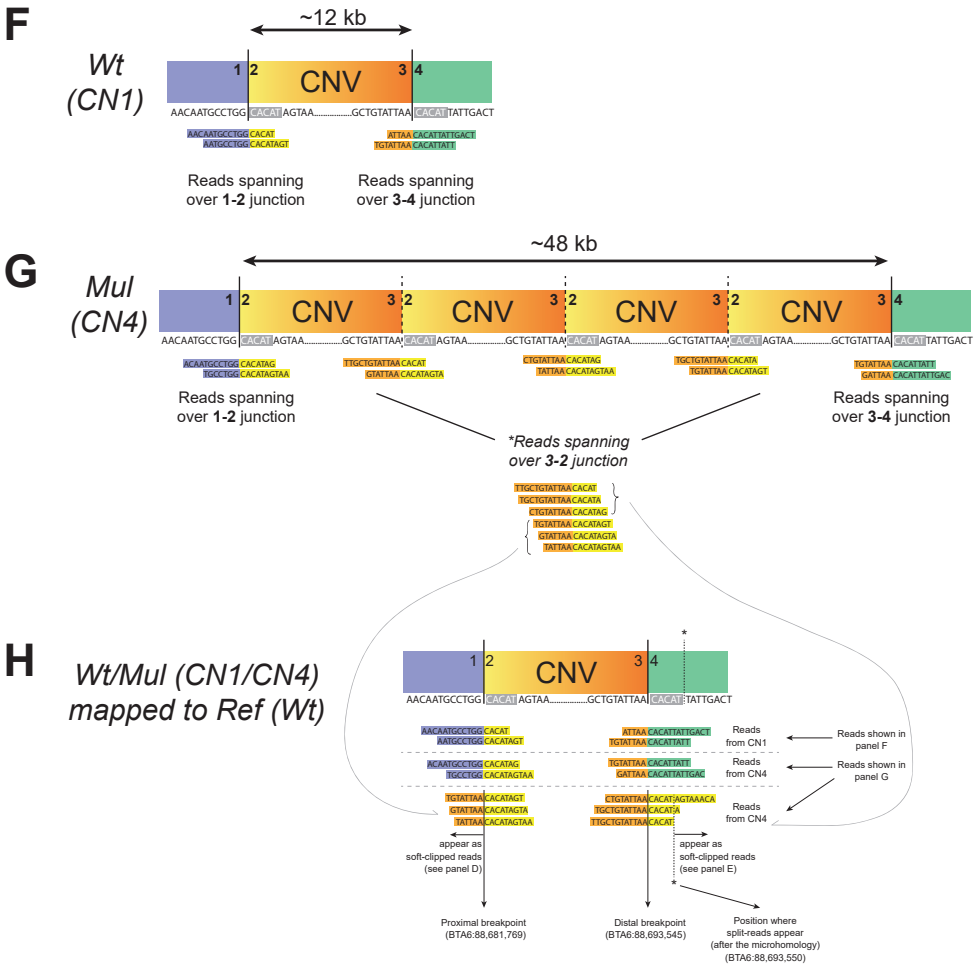
4.12. Competing interests

EM is an employee of CRV B.V., one of the partners of the Breed4Food consortium. All other authors declare that they have no conflict of interest.

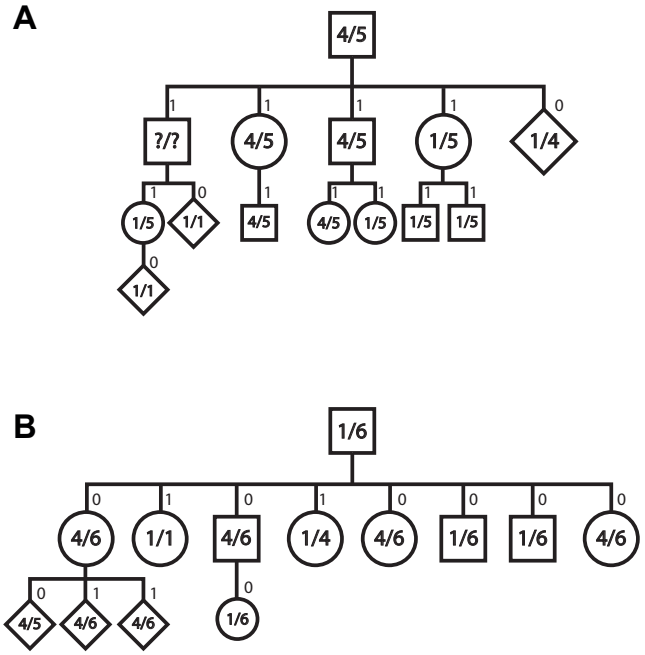
4.13. Supplementary figures



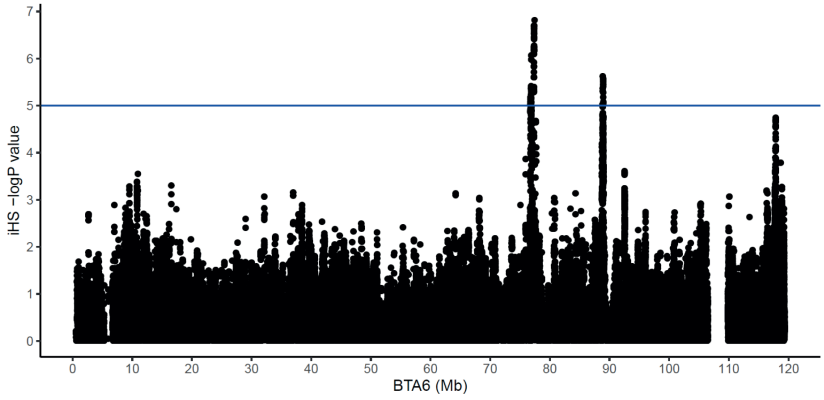
4.13.1. Supplementary Figure 1 IGV screen shots for the GC CNV and the breakpoints. (**A**) The GC CNV (BTA 6:88,681,767-88,693,553) is shown in the IGV screen shot. The grey reads are normally mapped reads, whereas green ones are discordantly mapped reads, providing evidence for a tandem duplication. The sequencing coverage of the CNV region is higher than non-CNV region. (**B**) The left breakpoint is flanked by MIR repeat. (**C**) The right breakpoint does not overlap with repeats. (**D**, **E**) The proximal and distal breakpoints are zoomed in and the soft-clipped reads (positions where nucleotide sequences are written) information revealed the 5-bp microhomology "CACAT" at the breakpoints (marked as yellow). (**F**-**H**) A schematic overview demonstrating a tandem configuration of the GC CNV.



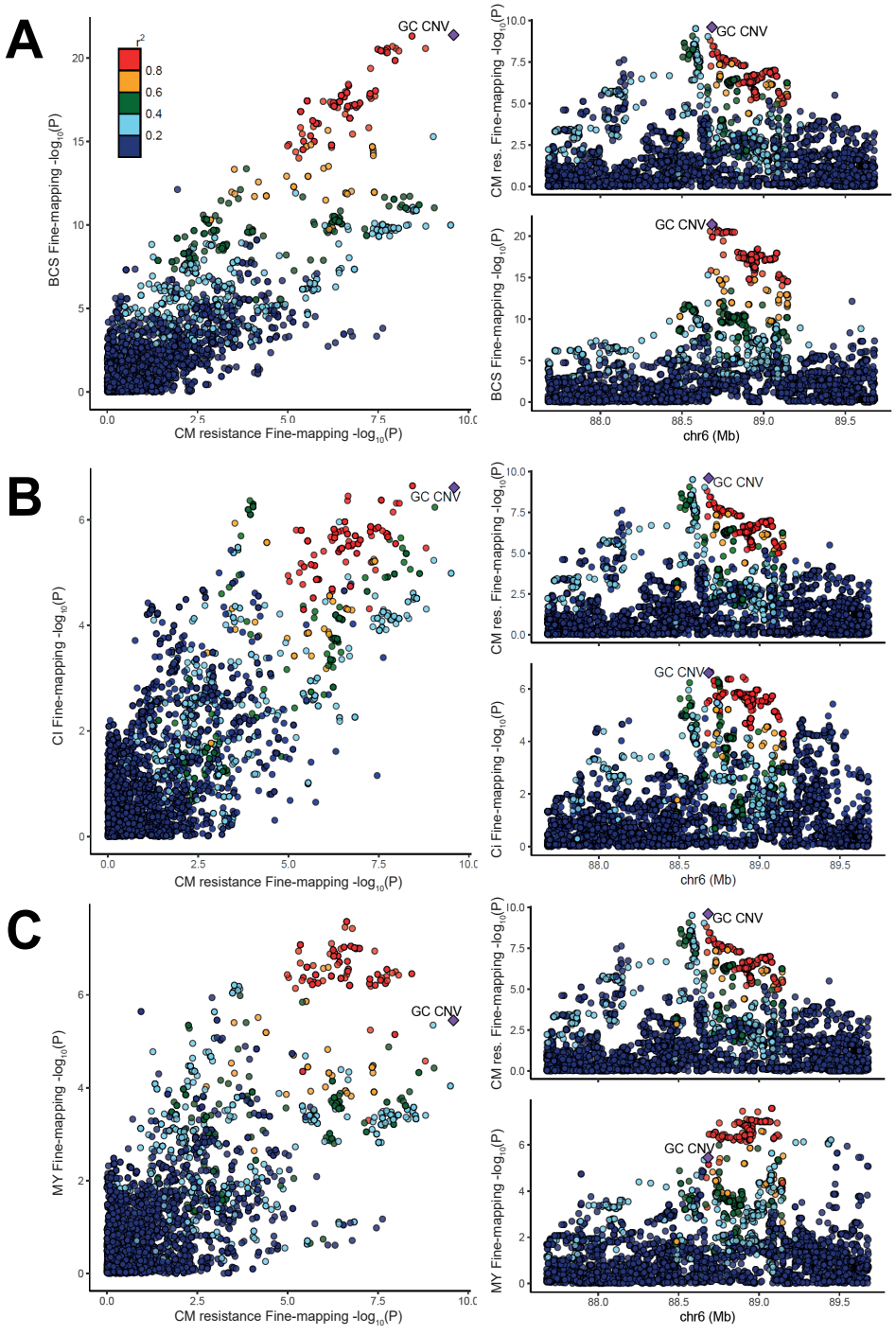
(**F**) The start and end of the GC CNV (shown in yellow-to-orange gradient colour) is flanked by non-CNV background, forming two junctions shown as 1-2 and 3-4 (marked with solid vertical lines). Sequencing reads from Wt (CN1) haploid genome can uniquely aligned to the 1-2 junction (blue-yellow reads) and the 3-4 junction (orange-green reads). The microhomology sequence is marked with grey shade. (**G**) A haploid CN4 genome forms unique junctions spanning over 3-2 formed by tandemly aligned 12-kb segments (marked with dotted vertical lines). Thus, reads from haploid CN4 would have reads spanning over 3-2 junction (orange-yellow reads), in addition to reads aligned to junctions 1-2 and 3-4. (**H**) Alignment of a heterozygous genome (CN1/CN4) to the reference genome (Wt) is shown. Both CN1 and CN4 have reads spanning over 1-2 and 3-4 junctions that are present in the reference genomes; these reads can be uniquely mapped. Reads spanning over 3-2 junctions cannot be uniquely mapped to the reference genome; these reads will be discordantly mapped either at 1-2 or 3-4 junctions (grey lines). The 3-2 junctions reads can be aligned to the junction 1-2, however orange reads will appear as soft-clipped. Likewise, these reads can be aligned to the 3-4 junction. However, the 4 junction started with the "CACAT" sequence. Hence, the sequences after the microhomology will appear as soft-clipped (marked with dotted line and asterisk).



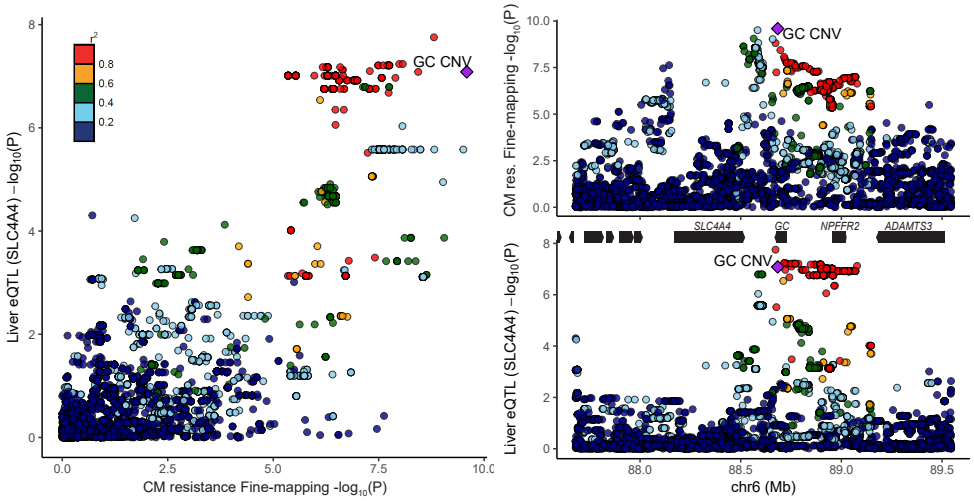
4.13.2. Supplementary Figure 2 Family tree of animals having CN 5 and CN6 allele on GC CNV locus. In the panel of 266 animals, we observed CN 5 and CN6 alleles are mostly segregating among a small number of related animals. To show that CN5 and CN6 alleles are truly segregating, transmission probabilities at each marker position were calculated with LINKPHASE3. Transmission probabilities were estimated for paternal haplotypes (1 and 0 indicated transmission of the paternal and maternal allele, respectively). (**A**) Largest family of CN5 carriers. Each circle indicates one individual and the number inside the circle indicates the GC CNV copy number genotype. The numbers above each individual stand for the transmission probability. Questions marks mean individuals with no copy number information. Male animals are shown as squares, female animals are shown as circles, and animals with unknown gender are marked with diamonds. (**B**) Largest family of CN6 carriers. The legends are identical to panel (**A**).



4.13.3. Supplementary Figure 3 Chromosome-wide scan of selection signatures using integrated Haplotype Score (iHS) on BTA6. Chromosome-wide scan of integrated Haplotype Score revealed two iHS peaks with high significance (-log10P>5), near BTA6:78 Mb and BTA6:89 Mb.


A
B
C
GC CNV
r²
0.8
0.6
0.4
0.2
BCS Fine-mapping -log10(P)
CM resistance Fine-mapping -log10(P)
CM res. Fine-mapping -log10(P)
chr6 (Mb)
CI Fine-mapping -log10(P)
MY Fine-mapping -log10(P)

4.13.4. Supplementary Figure 4 Three traits that showed strong association signals in CM resistance QTL region on BTA 6. Colocalization of fine-mapping p-values of a pair of traits are showing that body condition score (BCS), calving interval (CI), and milk yield (MY) are having association signal near or at the CM resistance QTL region on BTA 6. (**A**) Colocalization between body condition score and CM resistance is shown in the left panel, together with the separate Manhattan plots for body condition score and CM resistance on the right. (**B**) Colocalization between calving interval and CM resistance. (**C**) Colocalization between milk yield and CM resistance. Panel layout of (**B**) and (**C**) is the same as panel (A). In each panel, the colours of dots indicate degree of LD (r^2) with GC CNV (colour scale shown in the left upper corner of panel **A**). The purple diamond marks GC CNV.

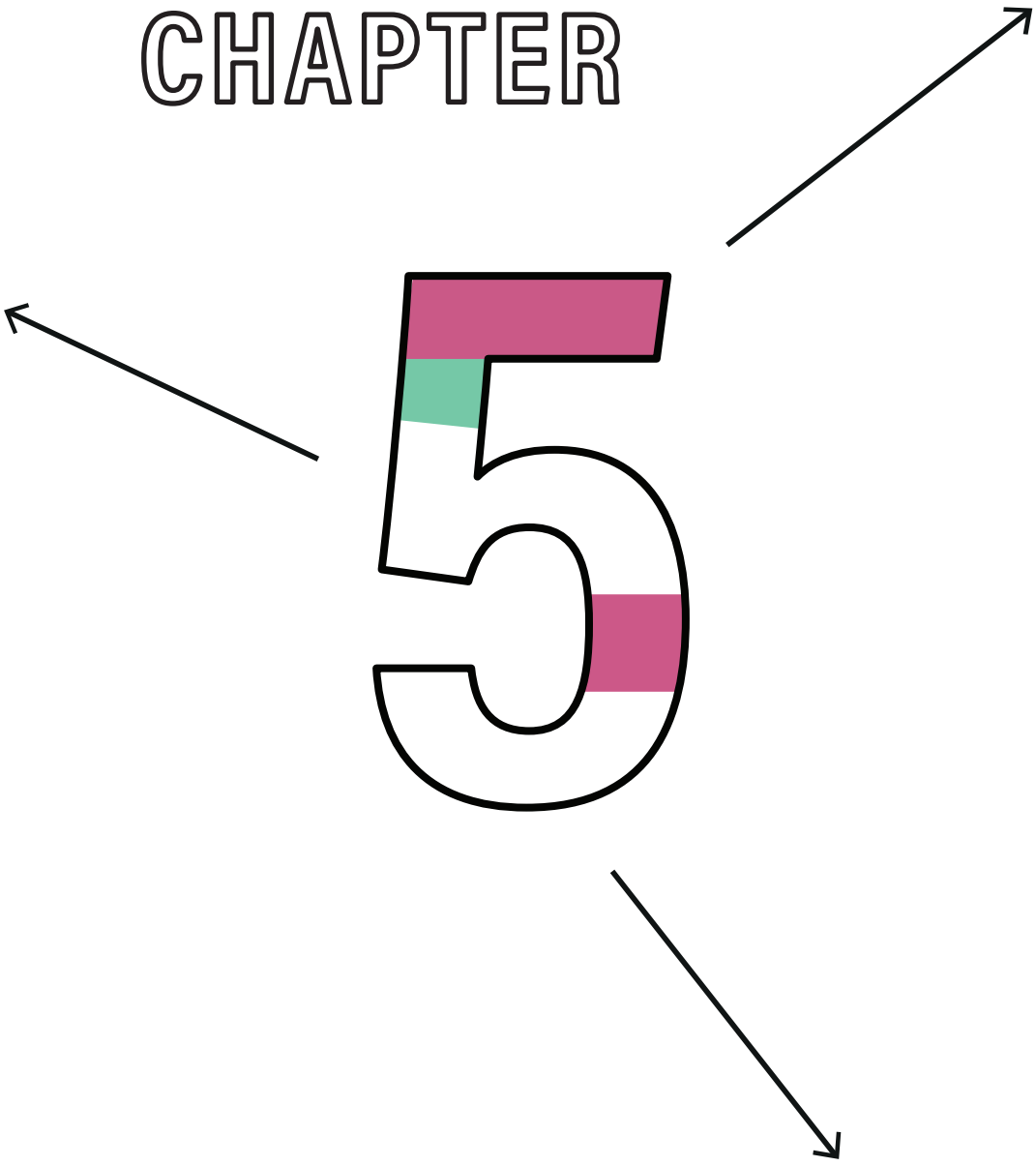


4.13.5. Supplementary Figure 5 eQTL mapping and colocalization of fine-mapping and SLC4A4 eQTL mapping results. A colocalization plot for CM resistance fine-mapping and *SLC4A4* eQTL mapping results is shown on the left side. The right upper panel is the CM resistance fine-mapping results and the right lower panel is the *SLC4A4* eQTL mapping results. In between these two right panels are the genes located in this region. Six genes on the left part are *AMBN, JCHAIN, RUFY3, GRSF1, MOB1B*, and *DCK*. In each panel, the colours of dots indicate degree of LD (r^2) with GC CNV (colour scale shown in the left upper corner of the colocalization figure. The purple diamond marks GC CNV.

4.14. Supplementary tables

Supplementary tables are available at the online version of the article.

CHAPTER



Extreme paternal bias in bovine *de novo* structural mutations in *in vitro* produced embryos including a high proportion of post-fertilization events

Young-Lim Lee[1], Aniek C. Bouwman[1], Mirte Bosse[1], Gabriel Costa Monteiro Moreira[2], Chad Harland[2], Latifa Karim[3], Roel F. Veerkamp[1], Martien A. M. Groenen[1], Erik Mullaart[4], Wouter Coppieters[3], Michel Georges[2], Carole Charlier[2*]

[1]Wageningen University & Research, Animal Breeding and Genomics, Wageningen, the Netherlands [2] Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Liège, Belgium [3] GIGA Genomics Platform, GIGA Institute, University of Liège, Liège, Belgium [4] CRV B.V., Arnhem, the Netherlands

In preparation

5

5.1. Abstract

Germline *de novo* mutations (DNMs) hold paramount importance for genetic and evolutionary studies. In recent years, whole genome sequencing (WGS) data of hundreds of trios unravelled a strong parental age effect, mainly for *de novo* single nucleotide variants (dnSNVs). Compared to dnSNVs, the *de novo* structural variation (dnSV) rate was shown to be far less frequent, based on array data. Nevertheless, the availability of WGS data opens a new avenue to detect different types of dnSVs at a finer scale, providing opportunities to estimate an accurate dnSV rate and explore the underlying mutational driver(s). We aimed at detecting dnSVs using 127 multi-generational pedigrees obtained from a healthy dairy cattle cohort. The unique population structure, where each pedigree consists of a trio and a large number of half-siblings and grand-offspring, enabled the detection of dnSVs with high confidence. The 127 probands were produced from three different assisted reproduction technologies, namely (i) artificial insemination, (ii) flushed embryo, and (iii) *in vitro* fertilisation (IVF). Overall, we discovered 20 dnSVs (15 deletions, four duplications, and one inversion), ranging from 58-bp to 1.2-Mb in size. We distinguished 15 late germline dnSVs and five early mosaic dnSVs, amounting to a germline dnSV rate of 0.12 per generation (equivalent to 1 dnSV per ~8.5 births). Furthermore, we observed a strong paternal bias of 14:1 among the 15 late germline dnSVs. Particularly, of the 15 late germline dnSVs, 11 were detected in IVF probands, highlighting a pronounced IVF effect. Of these, five late germline dnSVs were shown to harbour additional small scale DNMs of early developmental origin, pointing towards discrepant timing of mutations. We hypothesize that the strong IVF effect may arise from extensive DNA lesions arising in sperm cells and/or reduced DNA repair capacity of *in vitro* matured oocytes. Our work is the first of its kind to report an extreme paternal bias and a strong IVF effect in cattle.

5.2. Introduction

De novo mutations (DNMs) arising in the germline generate genetic diversity and provide substrates for selection. Thus, understanding of DNMs is crucial in genetic and evolutionary studies. Whole-genome sequencing (WGS) opened up the possibility to directly estimate the DNM rate (Conrad et al. 2011). DNMs are previously unreported genetic variants, where an animal (hereafter referred to as proband) is heterozygous, but neither of its parents is. Until now, most DNM studies focused on *de novo* single nucleotide variants (dnSNVs), discovering 45-65 dnSNVs per generation in human cohorts (Kong et al. 2012; Michaelson et al. 2012; Goldmann et al. 2016) and ~38 dnSNVs per generation in bovine data (Harland et al. 2017). Structural variations (SVs) include diverse classes of genetic variants, including deletions, duplications, and inversions, and are defined as being larger than 50-bp in size (Sudmant et al. 2015). An initial screen utilizing array data showed 1 dnSV per 77 births in humans (Itsara et al. 2010), confirming their rarity compared to dnSNVs. However, recent WGS based studies reported one dnSV per 7-13 births (Collins et al. 2020; Werling et al. 2018; Brandler et al. 2018), with 2.2:1 or stronger paternal bias (Belyeu et al. 2021; Kloosterman et al. 2015). Despite their importance, dnSVs studies have been limited to human cohorts, or a small Rhesus macaque cohort (14 trios; Thomas et al. 2020)

Below, the term 'DNMs' is used to describe overall *de novo* mutations; DNM of particular variant types (e.g. dnSNVs and dnSVs) are further distinguished to avoid confusion. So far, most dnSNV studies exploited two-generational trio pedigrees, focusing on elucidating mutational drivers. The key mutagenic factors replicated in multiple cohorts are sex and age effects. The male germline account for ~80% of the dnSNVs, and the number of dnSNVs increases as the male germline ages (~1 extra dnSNVs per year; Kong et al. 2012). Next to this, a larger cohort (>1,000 trios) showed a subtle yet significant maternal age effect (~1 extra dnSNVs per four years; Goldmann et al. 2016). Additionally, advanced maternal age was associated with a rise in clustered mutations (cDNMs) that harbour multiple DNMs (e.g. dnSNVs and a dnSV within 100-Kb distance; Goldmann et al. 2018). Some human cohorts revealed a higher number of dnSNVs in children conceived by *in vitro* fertilization (IVF), hinting that assisted reproductive technologies (ART) might play a role in DNM formation (Wong et al. 2016; Wang et al. 2021). However, the use of ART in humans is skewed towards couples with low fertility, which may correlate with other sub-optimal health conditions. Thus, whether ART increases DNMs remains to be investigated in an unbiased cohort.

The timing of DNMs can vary: DNMs can arise during gametogenesis or embryonic development (Figure 5.1). Commonly, germline DNMs are assumed to occur during gametogenesis,

resulting in a single or small number of DNM carrying gamete(s). Transmission of the DNM carrying gamete will result in a constitutional mutation in a proband (Figure 5.1 A). Thus, all cells have the DNM as heterozygous in the proband, and ~50% of gametes have the DNM. Furthermore, the DNM, if transmitted to the offspring of the probands (3rd gen), is in perfect linkage with the chromosome that it originally arose on (Figure 5.1 E). On the contrary, DNMs can occur during early embryogenesis. A DNM arising after the 1st cell division of the zygote will affect a subset of the cells, resulting in mosaicism. (Figure 5.1 B-D). The DNM carrying cell lineages, if developed into the germline, can be transmitted to the next generation. Such mutation carrier (the founder of the DNM) has three different haplotypes locally, where the DNM carrying haplotypes accounts for less than 50% (allelic imbalance). All three haplotypes can be transmitted, and hence, the offspring (3rd gen) may inherit the haplotype on which the DNM occurred, with or without the DNM – leading to imperfect linkage (Figure 5.1 F-G). We distinguish the former DNM as "late germline" as it appears late in one's developmental trajectory (gametogenesis of a mature individual), whereas the latter as "early mosaic", appearing as early as a zygotic phase. Elucidating the timing of DNMs (late vs early) requires extended pedigrees, where a two-generation pedigree is supplemented with siblings (2nd gen) or offspring (3rd gen) of the proband (Sasani et al. 2019; Rahbari et al. 2016).

Our data set, Damona, contains 743 dairy cattle genomes, consisting of 127 multi-generational bovine pedigrees obtained from a healthy cattle cohort. In each pedigree, a trio (sire-dam-proband) is complemented with the proband's half-siblings (HS; mean=7.8, 2nd gen) and grand-offspring (GO; mean=5, 3rd gen). Hence, this data set offers the opportunity to study both late germline and early mosaic DNMs (Figure 5.1 H). Furthermore, 127 Damona probands are produced via three different ART, making Damona an ideal data set for investigating the effects of ART in DNM formation. A pilot study focusing on dnSNV in 5 of the 127 Damona pedigrees revealed a paternal bias of 2.6:1 and 17% of mosaicism among all dnSNVs (Harland et al. 2017). However, the occurrence of germline and mosaic dnSVs and the effect of ART on dnSVs has not yet been studied in the Damona pedigrees.

Here, we investigated dnSVs in the Damona pedigrees, aiming at (i) identifying and distinguishing late germline and early mosaic dnSVs and hence obtaining an accurate bovine germline dnSV rate, (ii) investigating parent-of-origin effects in dnSVs, (iii) investigating effects of the various types of ART on dnSVs, and (iv) providing a fine-scale molecular characterisation of dnSVs.

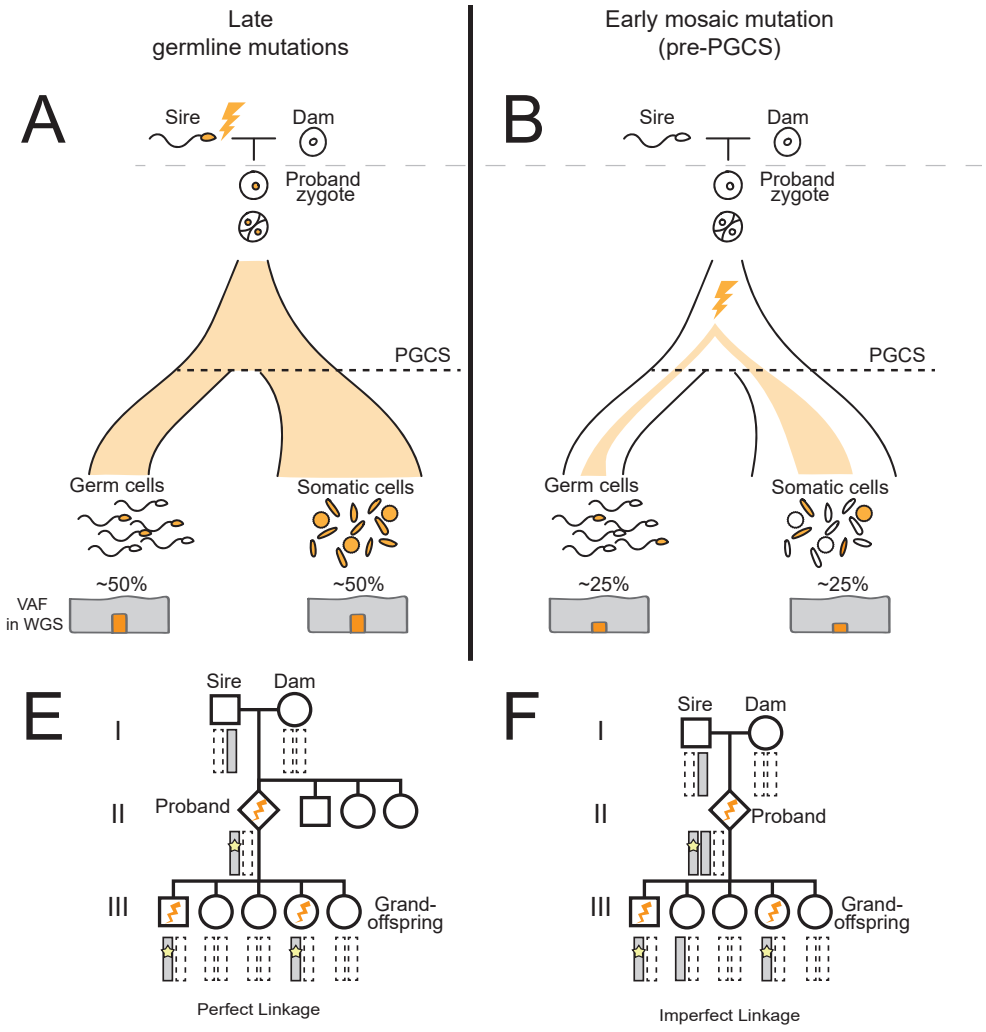
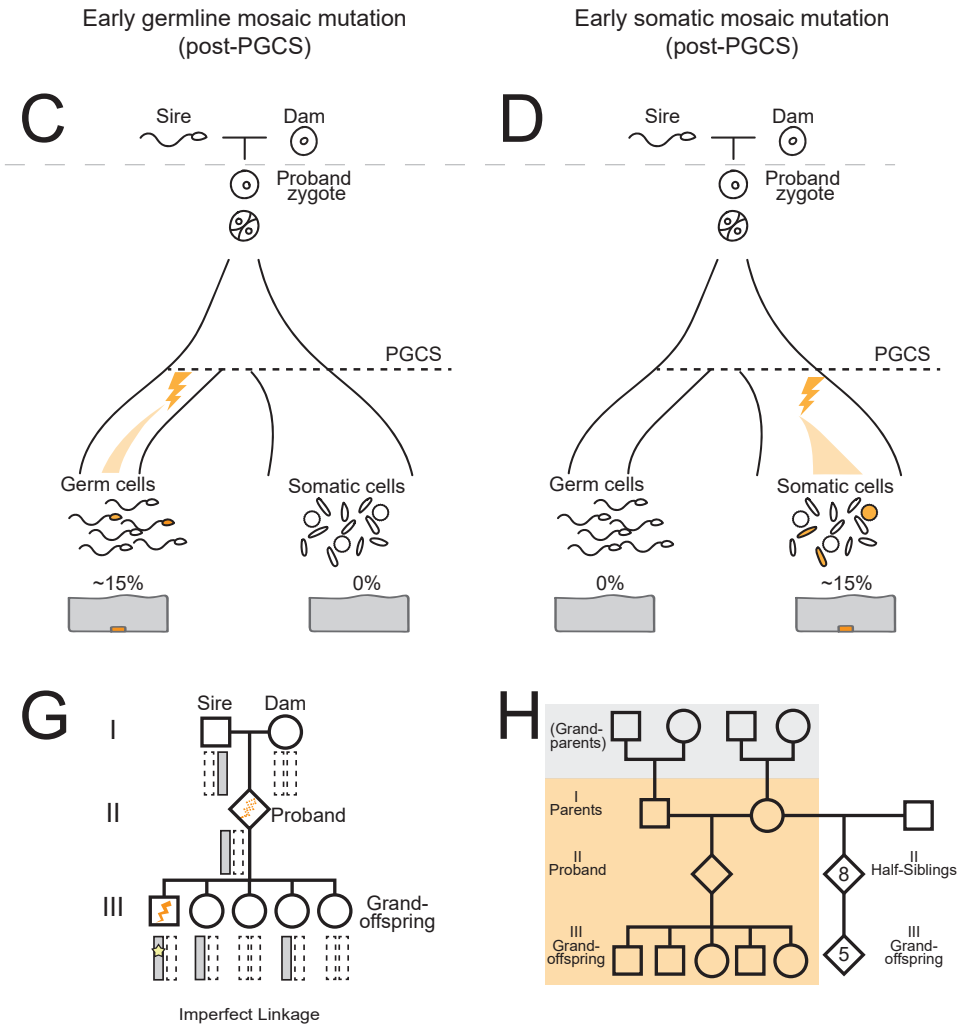


Figure 5.1. The timing and detection of de novo mutations in multi-generational pedigrees. (**A**) A late germline DNM in parents, if transmitted, will appear as a constitutional mutation in a proband. Hence, bulk sequencing of germline and somatic cells will show variant allele frequency (VAF) of ~0.5. (**B**) An early mosaic DNM can affect a subset of cell lineages resulting in mosaicism. The degree of mosaicism differs depending on the time of mutation. A DNM occurring before primordial germ cell specification (PGCS) can affect germline and soma. (**C-D**) After the PGCS, cells in the germline and soma do not exchange further. Thus, a post-PGCS germline DNM can be transmitted to offspring, whereas a post-PGCS somatic DNM cannot. (**E**) Detection of a late germline DNM in a three-generation pedigree. The haplotype of which a DNM occurred is marked with grey, and other parental haplotypes irrelevant



to the DNMs are shown as dotted bars. The DNM detected in the proband (2nd generation; marked with yellow star), if transmitted to the offspring (3rd gen), show perfect linkage with the haplotypic background where it originally occurred. (**F-G**) Detection of an early mosaic DNM. A pre-PGCS DNM can (i) be detectable in the proband, yet with a low fraction (allelic imbalance), (ii) be recurrently transmitted, and (iii) show imperfect linkage in offspring (3rd gen; shown in F). A post-PGCS DNM, if affecting a small fraction of the germline, may not be recurrently transmitted. Instead, imperfect linkage in offspring (3rd gen) can delineate the mosaic nature of the DNM (shown in G). (**H**) A typical Damona pedigree structure. Probands were defined as those with parent and offspring (3rd gen) available. Figure adapted from Jonsson et al. 2021.

5.3. Results

5.3.1. Detection of *de novo* structural variants

Our data set, Damona, consisted of 127 multi-generational bovine pedigrees, sequenced at 26X coverage for the trios and 8X coverage for other animals (e.g. GO and grandparents, when available). The 127 probands were produced via one of three assisted reproductive technologies (ART): (i) artificial insemination (AI, n=34), (ii) flushed embryo (FE, n=47), and (iii) *in vitro* fertilisation (IVF, n=46). The current study defined SVs as deletions, duplications, and inversions larger than 50-bp. SVs were detected using Lumpy (Layer et al. 2014).

5.3.2. Detection of late germline *de novo* structural variants

Our first aim was to detect late germline dnSVs. A late germline dnSV is defined as an SV which is (i) called heterozygous in the proband, (ii) absent in parents and HS, (iii) transmitted to ≥1 GO. By applying these filters and manually inspecting the underlying sequencing data in the Integrative Genome Viewer (IGV; Thorvaldsdóttir et al. 2013), we identified 15 late germline dnSVs, ranging from 58-bp to 1.2-Mb in size, consisting of 11 deletions, three duplications, and one large inversion (Table 1). Transmission of all dnSVs was confirmed, except a 1.2-Mb inversion appeared in a proband with one GO. The formation mechanisms were inferred based on the extent of sequence homology at the breakpoints. Fourteen out of 15 dnSVs showed limited to no sequence homology at the breakpoints (<15-bp), hinting at either microhomology-mediated end joining (MMEJ) or non-homologous end joining (NHEJ). NM-14 occurred in a region enriched for variable number tandem repeats (VNTR) and likely arose from mechanisms associated with VNTR formation such as slipped strand mispairing (Eslami Rasekh et al. 2021)

5.3.3. Bovine germline dnSVs are strongly biased towards male germlines

The 15 late germline dnSVs, identified in 127 pedigrees, indicated a germline mutation rate of 0.12 dnSV per generation (95% CI 0.08-0.21; 1 dnSV per 8.5 births). This rate corresponds to the estimates from healthy human cohorts (Belyeu et al. 2021; Werling et al. 2018; Brandler et al. 2018; Turner et al. 2017; Kloosterman et al. 2015; Figure 5.2). The parent-of-origin of the 15 dnSVs was determined using the three-generation pedigree structure. The dnSVs and the haplotypes of origin were in perfect linkage in GO (dnSV and the haplotype are transmitted together), supporting that the dnSVs arose during gametogenesis. Of the 15 late germline dnSVs, 14 were of a paternal origin, underscoring a striking paternal bias of 14:1 ($P= 9.8x10^{-4}$, two-tailed binomial test). Furthermore, we observed overrepresentation of dnSVs in probands obtained via IVF (11 out of 15 late germline dnSVs; $P=6.6x10^{-3}$, Fisher's exact test).

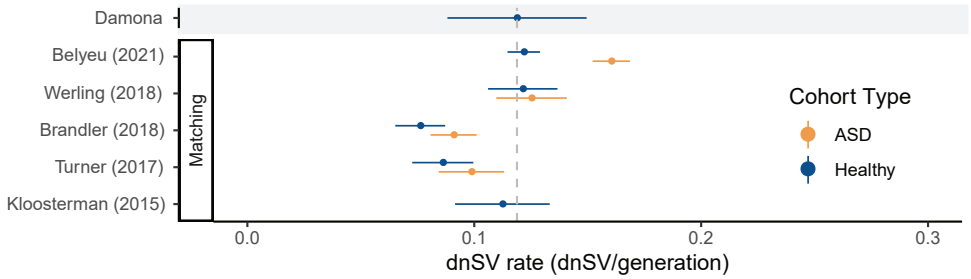


Figure 5.2. Frequency of germline de novo structural variations. The bovine germline dnSV rate was compared to previous reports in humans. The human studies used heterogeneous cohorts (healthy vs autism spectrum disorder (ASD)) and definitions for SVs. Hence, the dnSV rate for healthy and autism spectrum disorder (ASD) cohorts are shown separately (ASD in yellow and healthy in blue, numbers in brackets are the number of trios), limited to SVs that correspond to how it was defined in the current study. The mutation rate was obtained by dividing the number of dnSV by the number of trios. Error bars represent the 95% confidence intervals according to a Poisson distribution. For the dnSV rate including different SV definitions than the current study, see Supplementary Fig. 1 and Supplementary Table 1.

Table 5.1. List of *de novo* structural variations

dnSV ID	Coordinates	Length (bp)	SV Type	Locus
NM-1	chr1:30682619-30682677	58	Deletion	intergenic
NM-2	chr2:40607085-40653044	45,959	Deletion	intergenic
NM-3	chr2:58157986-58161632	3,646	Duplication	intergenic
NM-4	chr5:105850965-105853146	2,181	Duplication	intergenic
NM-5	chr6:42007578-42008209	631	Deletion	*ADGRA3* (intron 14/18)
NM-6	chr6:82111710-82113346	1,636	Deletion	intergenic
NM-7	chr6:83263187-83313423	50,236	Deletion	*CENPC* (5/19 exons), *STAP1* (6/9 exons)
NM-8	chr8:104571290-104580599	9,309	Deletion	intergenic
NM-9	chr10:10660965-10671493	10,528	Duplication	*TENT2* (3/15 exons)
NM-10	chr14:2861104-2861755	651	Deletion	*PTK2* (intron 1/32)
NM-11	chr15:77797101-77798011	910	Deletion	*PTPRG* (intron 1/22)
NM-12	chr16:76526481-76528614	2,133	Deletion	*DENND1B* (intron 16/22)
NM-13	chr17:15502897-15503002	105	Deletion	*INPP4* (intron 5/23)
NM-14	chr17:57212783-58417668	1,204,885	Inversion	*RFC5, KSR2, FBXO21, FBXW8, HRK, RNFT2, C17H12orf49*
NM-15	chr18:22389280-22389678	398	Deletion	*FTO* (intron 8/8)
M-1	chr1:11667465-11668783	1,318	Deletion	intergenic
M-2	chr4:38370379-38375464	5,085	Deletion	*CACNA2D1* (intron 3/38)
M-3	chr11:24355597-24360343	4,746	Deletion	intergenic
M-4	chr11:52854495-52955650	101,155	Deletion	intergenic
M-5	chr11:52955642-52964474	8,832	Duplication	intergenic

Mechanism	Mutational Type	Parent -of-origin	ART	Proband Sex	dnSV Transmission	cDNM
MMEJ	Late germline	Sire	AI	Male	4/5 (80%)	.
MMEJ	Late germline	Sire	IVF	Female	2/5 (40%)	Yes
MMEJ	Late germline	Dam	AI	Female	2/5 (40%)	.
MMEJ	Late germline	Sire	IVF	Female	3/5 (60%)	.
MMEJ	Late germline	Sire	FE	Female	2/5 (40%)	.
MMEJ	Late germline	Sire	IVF	Male	1/4 (25%)	.
MMEJ	Late germline	Sire	IVF	Male	3/5 (60%)	.
MMEJ	Late germline	Sire	IVF	Male	5/6 (83%)	.
NHEJ	Late germline	Sire	IVF	Female	2/5 (40%)	Yes
MMEJ	Late germline	Sire	IVF	Female	1/5 (20%)	Yes
MMEJ	Late germline	Sire	IVF	Male	3/5 (60%)	Yes
NHEJ	Late germline	Sire	IVF	Female	2/5 (40%)	.
MMEJ	Late germline	Sire	IVF	Female	1/1 (100%)	Yes
.	Late germline	Sire	FE	Female	0/1 (0%)	.
MMEJ	Late germline	Sire	IVF	Female	1/1 (100%)	.
NHEJ	Early mosaic	n.a.	AI	Male	2/8 (25%)	.
MMEJ	Early mosaic	Maternal	IVF	Female	3/5 (60%)	.
NHEJ	Early mosaic	Paternal	IVF	Female	2/5 (40%)	Yes
Complex	Early mosaic	Paternal	IVF	Female	1/5 (20%)	.
Complex	Early mosaic	Paternal	IVF	Female	1/5 (20%)	.

5.3.4. Mosaic dnSVs revealed through imperfect linkage and allelic imbalance

DNMs occurring after the first cell division can result in heterogeneous genomes in an individual, leading to mosaicism (Figure 5.3 A). In such case, a mosaicism carrier will (i) show allelic imbalance for the DNM, and (ii) if the germline is affected, the DNM can be recurrently transmitted, (iii) where DNM and the haplotype-of-origin may have imperfect linkage (Figure 5.1 F-G). We screened our pedigrees for dnSVs that manifest these early mosaicism signatures and identified five early mosaic dnSVs, consisting of four deletions and one duplication ranging from 1- to 101-Kb in size (Table 1). Notably, M-4 (deletion) and M-5 (duplication) were found in the same animal on the same haplotype. Furthermore, the breakpoints of M-4 and M-5 were adjacent to each other.

We found one early mosaic dnSV from sperm sequencing data (M-1), and the rest were found in blood sequencing data (Ms 2-5). The sperm mosaicism event, M-1, suggests transmissibility of the dnSV, and indeed the data in the subsequent generation showed both recurrent transmission and imperfect linkage (Figure 5.3 B). On the contrary, soma mosaicism (e.g. Ms 2-5) may affect either soma exclusively or both soma and germline. Therefore, we investigated data in the subsequent generation to confirm the transmissibility mosaic dnSVs discovered from soma sequencing. We confirmed that all soma mosaic dnSVs were transmitted, hence confirming that both germline and somatic tissue are affected (gonosomal mosaicism; Figure 5.3 C-D).

Finally, the mosaic dnSVs were phased using informative SNPs available in the three-generational data to determine the parent-of-origin (Figure 5.3 A). The parent-of-origin of M-1 was not determined, although successfully phased, because the carrier of M-1 was a sire; hence no data from parents (i.e. grandparents from the probands' point of view) was available. We found that M-2 occurred on a maternal haplotype, and the rest (Ms 3-5) appeared on paternal haplotypes. Early developmental DNMs are expected to affect the parental chromosomes with equal probability. Our observation was biased towards the paternal chromosome (3:1). Also, all but one (M-1) early mosaic dnSVs were observed in IVF probands, suggesting an IVF effect in early mosaic dnSVs. However, the paternal bias and the IVF effect were not statistically significant (P=0.3, one-tailed binomial test for paternal bias; P=0.12, Fisher's exact test for IVF effect), due to the insufficient number of observations. Nevertheless, taking the early and late dnSVs together, the IVF effect was significant P= 9.2×10^{-4}, Fisher's exact test).

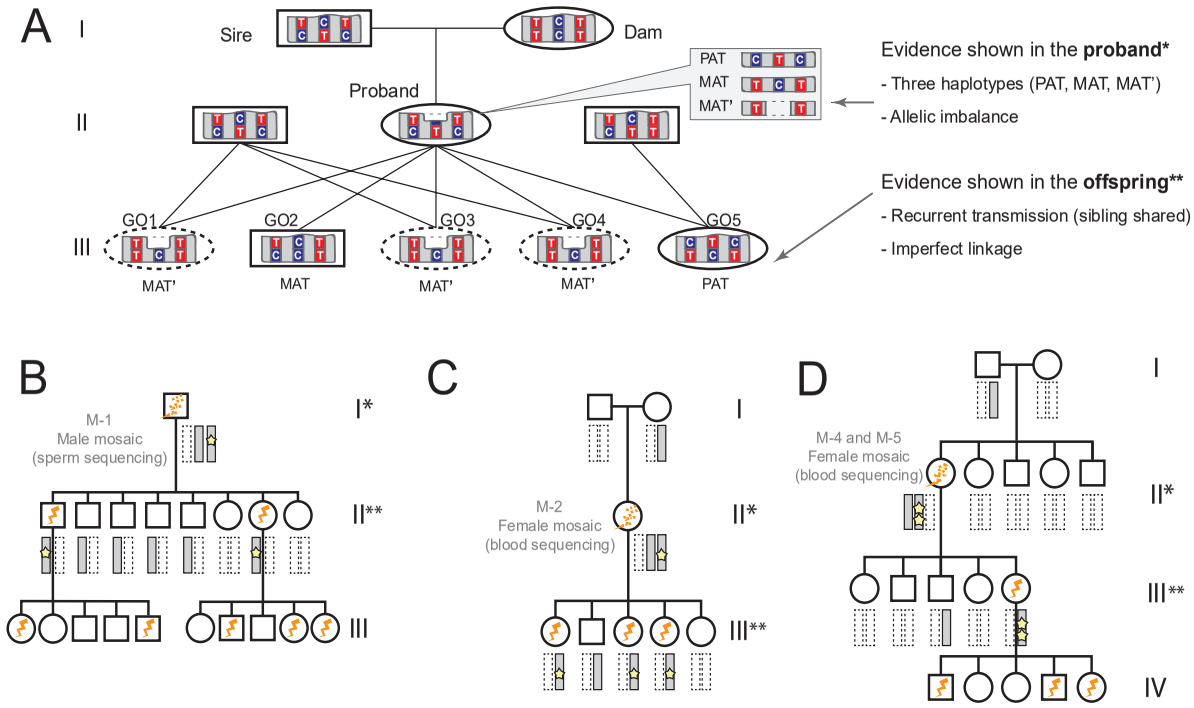


Figure 5.3. Characterisation of early mosaic dnSVs. (**A**) A framework to delineate a mosaic dnSV is demonstrated based on the actual data obtained from M-2. The mosaic dnSV evidence shown in the proband includes: (i) three haplotypes and (ii) allelic imbalance. In the current example, the dnSV occurred on the maternal haplotype, and hence the proband contained both paternal (PAT), maternal (MAT), and the maternal haplotype with the dnSV (MAT'). The mosaic dnSV evidence shown in GO includes: (i) recurrent transmission (multiple offspring carrying MAT' – marked with a dotted circle) and (ii) imperfect linkage. In the current example, all of these criteria were met. (**B-D**) Mosaic dnSV pedigrees and their haplotypes are shown, for M-1, M-2, and M-4/5, respectively. The haplotype legend is as explained in Figure 5.1. The mosaicism carriers are marked with dotted bolts. The transmission of these three haplotypes and imperfect linkage was observed in the subsequent generations (marked with two asterisks).

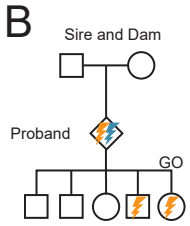
5.3.5. Clustered *de novo* mutations in the male germline

Some DNMs arise together in spatial proximity (e.g. <20-Kb apart). If these DNMs are on the same allele – for instance, two very closely located DNMs arisen on the paternal haplotype – they are assumed to occur from a single mutagenic event. Such DNMs, referred to as clustered DNMs (cDNMs or paired DNMs), account for 2-3% of human dnSNVs (Michaelson et al. 2012). The incidence of maternal cDNMs consisting of dnSNVs were strongly correlated with maternal age (Goldmann et al. 2018). The same study reported 11% (5 out of 45) of the dnSVs to be cDNMs (paired with dnSNVs), all of maternal origin, pointing towards compromised double-strand break (DSB) repair in ageing oocytes. Together, these findings show that the female germline can accrue DNA damage driven DNMs (both dnSVs and dnSNVs), opposing the textbook view that DNMs mostly arise in the male germline due to replication errors. Accordingly, enrichment of cDNMs in a particular condition or genomic locations can provide insights into the novel source(s) contributing to DNM formation.

Thus, flanking regions of the 20 dnSVs were scanned, searching for other DNMs (e.g. dnSNVs and dnINDELs). We found that five late germline dnSVs and one early mosaic dnSVs harbour another DNM in the flanking region (Figure 5.4 A). All of the cDNMs pairs (i) arose on the paternal haplotype, (ii) observed in IVF probands, and (iii) the inter-mutational distances between the mutation pairs were very close (<3-Kb). As elaborated earlier, paired DNMs are assumed to form via a single event. Accordingly, paired DNMs ought to arise during the same developmental phase (either late or early). An early mosaic dnSV (M-3) was paired with a 1-bp deletion, which showed allelic imbalance (indication of early mosaicism). Hence, both M-3 and the paired 1-bp deletion likely arouse during early development of a zygote. To our surprise, while the dnSVs in the remaining five DNM pairs were unambiguously classified as late germline dnSVs, the associated dnSNVs and dnINDELs behaved as early mosaic DNMs (allelic imbalance and imperfect linkage; Figure 5.4 A-B). Thus, while their proximity and occurence on the same allele suggest that they derive from a single event, their distinct status (late germline dnSV paired with early mosaic dnSNV/dnINDEL) suggests that they derive from distinct mutational events. Of the five late germline dnSVs, NM-2 and NM-11 manifest strong evidence of mosaicism (three haplotypes, allelic imbalance, and imperfect linkage; Figure 5.4 C). This would make sense if DNA lesion in sperm cells led to two DNMs of which, the one was repaired upon fertilization hence indistinguishable from the late germline dnSV, whereas the other segregated unrepaired.

A

dnSV ID		dnSNV / dnINDEL	Distance	ART	Proband			Offspring	cDNM evaluation
					Three haplotypes	Allelic (WGS)	Imbalance (Allegro)	Imperfect linkage	
M-3	5-Kb DEL	1-bp DEL	410 bp	IVF	/	0.27	0.25	perfect	Early-Ealry
NM-9	10-Kb DUP	SNV (G>A)	2,913 bp	IVF	/	0.15	0.23	perfect	Late-Early
NM-10	651-bp DEL	8-bp DEL	592 bp	IVF	/	0.30	/	perfect	Late-Early
NM-13	105-bp DeL	11-bp INS	404 bp	IVF	/	0.33	0.4	perfect	Late-Early
NM-2	46-Kb DEL	22-bp DEL	108 bp	IVF	Yes	0.19	/	Imperfect*	Late-Early
NM-11	910-bp DEL	SNV (A>T)	367 bp	IVF	Yes	0.20	/	Imperfect*	Late-Early



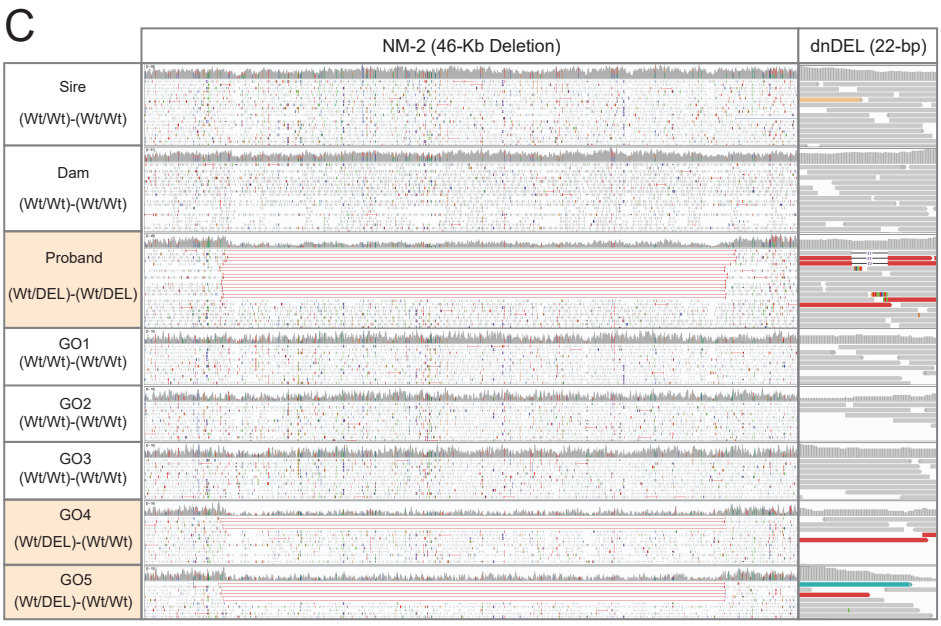
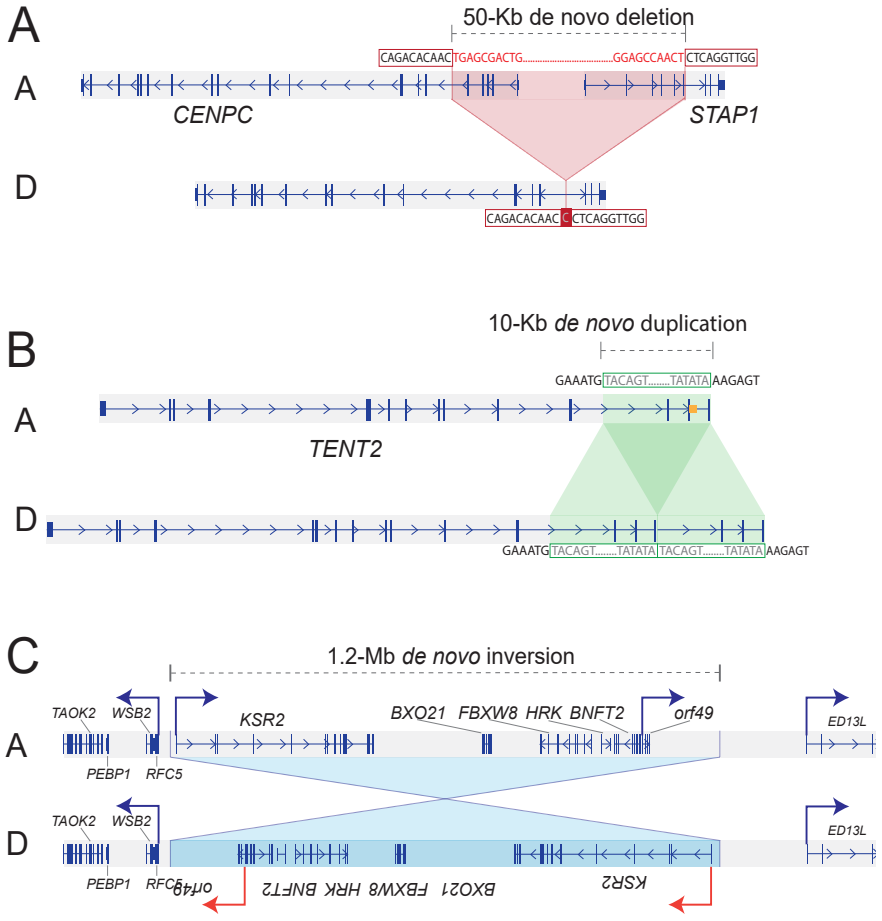


Figure 5.4. Characterisation of de novo cluster mutations. (**A**) List of cDNMs. (**B**) A schematic overview of cDNMs where the two DNMs do not co-segregate, resulting in imperfect linkage. (**C**) IGV screen capture of cDNM consisting of NM-2 and a paired small DNM (22-bp deletion). The left most panel shows families of this pedigree and their genotypes for NM-2 and the paired dnDEL. The panel in the middle shows discordant reads spanning over NM-2 (marked red) and reduced coverage within the deletion. The right panel shows the 22-bp dnDEL, which arisen in the proband, however did not co-segregate with the NM-2.

5.3.6. Fine-scale characterisation of *de novo* SVs

Overall, we identified 21 dnSVs, of which three (NM-8, NM-9, NM-14) directly disrupted coding sequences (CDS), whereas the others were either intronic or intergenic (Table 1). Of the three CDS disrupting dnSVs, NM-8 was considered to be potentially deleterious. This 50-kb deletion ablates transcription starting sites and several exons of two genes, *centromere protein C* gene (*CENPC*), which is related to cell division, and *signal transducing adaptor family member 1* gene (*STAP1*), which encodes a docking protein and is involved in tyrosine-protein kinase Tec activity (Figure 5A). *CENPC* is essential during mitosis, as loss of *CENPC* was related to increased chromosome misalignment (Gopalakrishnan et al. 2009). The importance of *CENPC* was demonstrated in a mice knock-out study: heterozygous mice with one functional copy of *CENPC* were healthy and fertile, whereas the homozygous state resulted in embryonic lethality (Kalitsis et al. 1998). Likewise, matings of NM-8 carriers, if leading to homozygous NM-8 animals, may result in abortion (Figure 5D).



Furthermore, NM-9, a 10-kb duplication, overlapped with the terminal nucleotidyltransferase 2 gene (TENT2), involved in mRNA stabilisation and polyadenylation (Figure 5B). The poly(A) signal appeared after the duplication, suggesting that the transcript might be elongated. An orthologous region in humans contains a regulatory T-cell enhancer, suggesting potential regulatory consequences in the bovine genome if this element is conserved (The ENCODE Project Consortium 2012).

Lastly, the 1.2 Mb inversion (NM-14) spanned over seven coding genes without directly disrupting CDS (Figure 5C). Inversions could have harmful consequences by disrupting CDS or relocating regulatory elements into 'out-of-context' regions (Laugsch et al. 2019). In the current case, we did not observe transmission of NM-16, which could be due to chance (only one GO) or unbalanced gametes formed due to NM-16, which did not produce viable offspring (Wellenreuther and Bernatchez 2018) (Figure 5F).

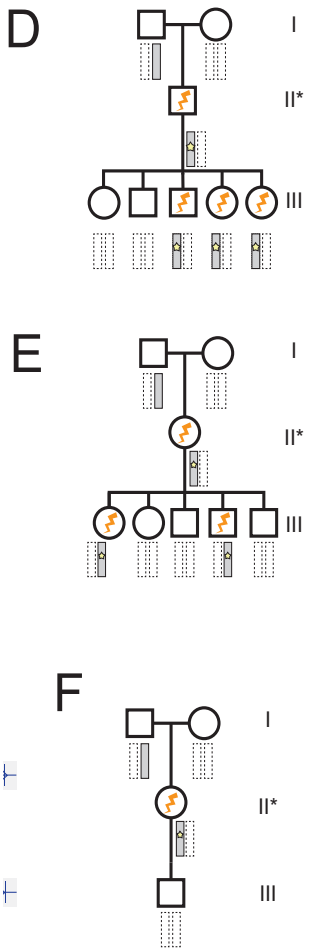


Figure 5.5. A schematic overview and inheritance of CDS disrupting dnSVs. (**A**) NM-8 is a 50-kb deletion overlapping with *CENPC* and *STAP1*. The likely formation mechanism is MMEJ, mediated by 1-bp insertion. The ancestral and derived forms are marked with A and D. (**B**) NM-9 is a 10-kb duplication, encompassing the last three exons of *TENT2*. No sequence homology was observed at the breakpoints, indicating NHEJ mechanism. The enhancer found in the orthologous region in human is marked with yellow. (**C**) The 1.2-Mb inversion encompasses seven coding genes. (**D-F**) Transmission of CDS disrupting dnSVs. Probands are marked with an asterisk. The haplotype legend is as explained in Figure 5.1.

5.4. Discussion

Our study is the first to investigate dnSVs in a healthy cattle cohort, exploiting multi-generational pedigrees with complete ART information. In total, 20 dnSVs were discovered, ranging between 58-bp and 1.2-Mb (15 deletions, four duplications, and one inversion; Table 1). Using three-generation bovine pedigrees, we distinguished the 20 dnSVs into 15 late germline and five early mosaic dnSVs, revealing a germline dnSV rate of 0.12 per generation. Taking the two early mosaic dnSVs arouse in one animal as a single event (M-4 and M-5), our results indicate 21% of detected dnSVs are mosaic (4 out 19). This fraction is rather high compared to 8% mosaic dnSVs reported in two-generation pedigrees in humans (Belyeu et al. 2021). This discrepancy may arise from the different data sets, where the two-generation pedigrees provide limited power to ascertain mosaicism. Thus, re-evaluation of the dnSVs with transmission data may assist in obtaining more accurate germline dnSV rates in humans.

We found an unprecedented paternal bias of 14:1 in bovine germline dnSVs. This bias is far from the paternal bias of 2.6:1 shown in bovine dnSNVs in a subset of Damona pedigrees (Harland et al. 2017). Likewise, human dnSNV studies have reported a paternal bias of approximately 4:1, which was attributed to continuous replication in the male germline, resulting in a paternal age effect (e.g. male germline accrue about one dnSNVs each year; Kong et al. 2012; Goldmann et al. 2016; Jónsson et al. 2017). Contrary to this, dnSVs in humans did not find a significant paternal age effect, despite showing a moderate paternal bias of 2.7:1 (Belyeu et al. 2021). These findings collectively suggest that dnSVs occur more frequently in the paternal genome, although they are unlikely to arise due to DNA replication errors. The post-meiotic phase is a possible route where sperms accumulate more DNA damage than oocytes. Spermiogenesis occurs after meiosis and constitutes the final phase of spermatogenesis. During spermiogenesis, spermatids are matured into sperm cells by removing cytoplasm, and protamine condenses the DNA packaging, leaving chromatin transcriptionally inactive. This sperm maturation phase induces DSBs, yet they are left unrepaired given its haploid nature (no homologous chromosome available for repair) and lack of transcription (Bergero et al. 2021; Grégoire et al. 2018). In comparison, oocytes are transcriptionally active and accumulate mRNA that can repair sperm damages once fertilized (García-Rodríguez et al. 2019). Thus, post-meiotic DSBs preferentially occurring in mature sperm cells seem a plausible cause of the paternal bias in dnSVs.

Yet, to our understanding, post-meiotic DSBs in sperm cells alone do not fully explain the extreme paternal bias shown in our study because: (i) our bias is far more extreme than those of humans and (ii) if the post-meiotic DSBs are the sole source of paternal dnSVs, we ought to observe an even distribution of dnSVs across the three ART methods, which we did not. Instead, we observed a strong IVF effect (11 out of 14 paternal late germline dnSVs were found in IVF probands). This implies a preferential occurrence of dnSVs in IVF condition, particularly on

the paternal genome, possibly due to (i) further accruement of DSBs occurring in sperm cells, (ii) low DSB repair capability of *in vitro* matured oocytes, or (iii) both of them. Two human studies have reported decreased DSB repair efficiency of ageing oocytes as the key source of maternal cDNMs (Goldmann et al. 2018) and early mosaic DNMs (Gao et al. 2019). Whether the strong IVF effect is arising from the sperm cells, or due to compromised repair in oocytes, in an analogous manner as proposed in human aging oocytes, warrants further investigation.

So far, the reported effects of ART on dnSVs have been contradictory. IVF embryos were shown to have more chromosomal aberrations (e.g. aneuploidy), compared to *in vivo* conceived embryos in humans (Voet et al. 2011) and cattle (Tšuiko et al. 2017), suggesting detrimental outcomes, including embryonic losses. On the contrary, a genotyping array-based comparison of children born via natural conception and IVF did not reveal any difference in the number of large *de novo* copy number variants (dnCNV; detection resolution>100-Kb; Zamani Esteki et al. 2019). It is worth noting that the dnSVs detected in our WGS data set are as small as 58-bp, and only M-4 (101-Kb mosaic deletion) would have been detected with the array used in (Zamani Esteki et al. 2019). Thus, our study suggests that the contribution of IVF might be more significant for fine-scale SVs.

The mutation rate obtained in the current study (˜0.12 dnSV per generation) corresponds to estimates from healthy human cohorts. Given the strong IVF effect discovered in the current study, we speculate that limited to the cattle population, IVF has the potential to accelerate the emergence of dnSVs. Undoubtedly, various types of ART have been used widely for cattle breeding, even though, the practice of IVF varies depending on the country (Viana 2020). Therefore, although our results are limited to a single local cattle population, we postulate that population-wide dnSV rates would differ depending on how frequently IVF is practised. Likewise, IVF is widely accepted in human reproduction, and thus the contribution of ART in the formation of fine-scale dnSVs (<100-Kb) warrants further investigation.

Finally, we identified 20 dnSVs, of which 3 directly disrupt CDS. Among these, NM-7, a 50-Kb deletion ablating *CENPC* is considered a potentially recessive lethal mutation. In theory, extensive use of popular sires can increase allele frequency of recessive alleles, which can happen to NM-7. Therefore, it may be worth monitoring the frequency of this deletion in the population to avoid carrier-carrier matings. Furthermore, SVs analysed in the current study are deletions, duplications, and inversions. While such SVs are abundant in bovine genomes, some forms of SVs have not been investigated in the present study, such as *de novo* mobile element insertions and complex inter/intra-chromosomal events (Feusier et al. 2019; Brandler et al. 2016; Richardson et al. 2017). In addition, our catalogue of dnSVs is bound to the inherent technological bias of short reads sequencing (i.e., challenge in mapping repetitive regions; Ho et al. 2019). Thus, we stress that the mutation rate estimated here could be a lower bound of the true SV mutation rate, which will be better estimated with long-read sequencing technology.

5.5. Conclusion

This study utilized unique multi-generational bovine pedigrees to identify and characterize dnSVs. The discovery of 15 late germline and five early mosaic dnSV events underline the importance of multi-generational pedigree in obtaining an accurate germline mutation rate. In addition, our germline dnSVs confirmed the paternal bias shown in dnSNVs; however, it is likely arising from post-meiotic DSBs, instead of replication errors. Our study confirms reports six cDNMS consisting of dnSVs and dnSNVs/dnINDELs. Five of them indicated distinct timing of mutations: dnSVs were detected as late paternal events, whereas accompanied dnSNVs/dnINDELs indicated early mosaicism in probands. This conundrum further supports that IVF may be highly mutagenic and alter the post-fertilization repair capability of oocytes, similar to what was reported in ageing oocytes in humans. Followingly, we expect that the adoption of IVF in cattle breeding have increased the dnSV rate, and this may apply to human cohort, which warrants future studies.

5.6. Materials and methods

5.6.1. Multi-generational bovine pedigrees

The Damona data set was consisted of 743 Dutch Holstein Friesian cattle genomes, forming 127 multi-generational pedigrees (minimum of three generations). For each pedigree, a proband was defined as a second generation animal where dnSVs were scanned. Thus, each pedigree consisted of (i) a trio (sire-dam-proband), (ii) on average 5.6 paternal and 2.2 maternal HS of the proband, and (iii) on average 5 GO (min=1, max=11) (Figure 5.1 H). In addition, some pedigrees had complete or partial grandparents (GP) data available, forming four-generation pedigrees (4 pedigrees had all GP data, five pedigrees had paternal GP data, and 32 pedigrees had maternal GP data). The probands (88 female and 39 male) were produced via one of three assisted reproductive technologies (ART): (i) artificial insemination (AI, n=34), (ii) flushed embryo (FE, n=47), and (iii) *in vitro* fertilisation (IVF, n=46).

5.6.2. Whole-genome sequencing data

Trio animals in the 127 pedigrees were together 236 animals, and their DNA was obtained from blood (144 females and 22 males) or sperm samples (70 males) using standard procedures. Familial relationships were confirmed by genotyping all samples with the 10K Illumina SNP chip. We constructed 550 bp insert size whole-genome Illumina Nextera PCR free libraries following the protocols recommended by the manufacturer. All samples were then sequenced on Illumina HiSeq 2000 instruments, using the 2x100 bp paired-end protocol by the GIGA Genomics platform (University of Liège). The sequence data was mapped using BWA mem 0.7.5a(Li 2013) to bovine reference genome ARS-UCD1.2. Afterwards, SAMtools 1.9 (Li and Durbin 2009) was used to convert sam files into bam files. Subsequently, the bam files were sorted with sambamba 0.6.6 (Tarasov et al. 2015), and the PCR duplicates were removed using Picard Tools 2.7.1 (https://github.com/broadinstitute/picard). The 236 trio animals were

sequenced at a mean coverage of ~26X, and the rest (GO and GP) were sequenced at mean coverage of 8X (Figure 5.1 H).

5.6.3. SV discovery

We discovered SVs using the population calling mode in Smoove (https://github.com/brentp/smoove). Firstly, Lumpy used split- and discordant- reads evidence to detect population-wide SVs in 127 trios (n=236, mean coverage=26X) (Layer et al. 2014). Lumpy was designed to detect deletions, duplications, inversions, and breakends which are junctions that could not be classified into canonical forms of SVs (Abel et al. 2020). Thus, in our study, the scope of dnSVs was limited to deletions, duplications, and inversions. Afterwards, the population-wide SVs were merged using SVtools (Larson et al. 2019), generating a non-redundant SV call set. Subsequently, the full cohort of 743 animals was genotyped (236 animals forming 127 trios and 507 animals either HS, GO, or GP of the probands) using SVtyper (https://github.com/hall-lab/svtyper). Read depth information of CNVs was annotated using duphold (Pedersen and Quinlan 2019).

5.6.4. Late germline *de novo* SVs detection

Late germline dnSVs are assumed to arise in a single gamete of a parent. To detect later germline dnSVs, we scanned the SV call set for the following criteria: (i) both parents are homozygous reference with evidence in sequencing data as follows: max. ALT supporting reads < 4 and ratio of ALT supporting reads < 0.1, (ii) proband is heterozygous with evidence in WGS data as follows: min. ALT supporting reads > 2 and ratio of ALT supporting reads > 0.05, (iii) All of the HS of the probands are homozygous reference, (iv) the dnSV is transmitted to at least 1 GO, and the dnSV and the cognate haplotype where the dnSV arose are in perfect linkage. We made one exception for a detected inversion (was not transmitted to GO), given that an inversion can give rise to unbalanced gametes, likely leading to unviable offspring. All candidate sites that passed these filters were manually inspected using IGV in all members of an extended pedigree.

5.6.5. Determining parent-of-origin

For unbalanced SVs (deletions and duplications), we used informative SNPs located within the dnSVs to determine the parental origin. SNPs located within a *de novo* deletion are expected to be homozygous due to hemizygosity. Thus, the SNP allele(s) present within the deletion are from the parent that transmitted a normal gamete, whereas the absent alleles are from the DNM-affected gamete. Contrary to this, SNPs located within duplications show allelic ratio deviating from 1:1. With a simple tandem duplication, the SNP allelic ratio is expected to be 1:2, where the duplicated allele accounts for twice the amount of reads, revealing the parent of origin. A *de novo* inversion, where SNP allelic depth is not informative, we inspected discordant reads which might carry informative SNPs to assign parent of origin. As no informative SNPs were present within the SVs and/or on the discordant reads, we phased SNPs outside of the dnSV, exploiting genomes of GP and GO.

5.6.6. Computation of mutation rates

The SV mutation rate was computed for late germline dnSVs only (excluding early mosaic dnSVs). The mutation rate per generation was calculated by dividing the number of late germline dnSVs (n=15) by the number of trios (n=127). The standard error was calculated assuming a Poisson distribution; taking a square root over the mutation rate per generation divided by the number of trios. The 95% confidence interval was used in Figure 5.2. To compare the mutation rate, we retrieved the number of dnSV from other studies (Werling et al. 2018; Turner et al. 2017; Kloosterman et al. 2015; Belyeu et al. 2021; Brandler et al. 2018) and calculated the mutation rate and standard error as explained above. As these dnSV studies used various SV calling pipelines and filtering steps, we only kept the SVs matching the SV definition used in our study (i.e. >50-bp, deletions, duplications, and inversions).

5.6.7. Early mosaic *de novo* SVs detection

Early mosaic dnSVs, if transmitted to the next generation, are assumed to affect part of the germline. Our pedigree structure enables detection of mosaicism that occurred in parents (1st gen; based on recurrent transmission to a proband and part of its HS) and probands (2nd gen; based on recurrent transmission to GO). To avoid confusion, either case was referred to as mosaicism carriers. Early mosaic dnSVs can have three signatures: (i) allelic imbalance shown in the mosaic carrier, (ii) recurrent transmission of the dnSVs to multiple offspring, (iii) the dnSV and the haplotype of origin are in imperfect linkage (thus, the haplotype of origin is transmitted to the next generation, with or without the dnSV). Accordingly, we classified dnSVs that met at least one of the three signatures as early mosaic dnSVs. De novo deletions harboring polymorphic SNPs are considered mosaic (allelic imbalance). SNP calling was done using FreeBayes (Garrison and Marth 2012), and low-quality SNPs (QUAL<100) were filtered out. Informative SNPs were phased and assigned to either paternal or maternal alleles.

5.6.8. Detection of cluster DNMs

To identify cDNMs, we detected dnSNVs and dnINDELs near the breakpoints in a 20-Kb window. Variant calling was done using FreeBayes(Garrison and Marth 2012), and for the entire dnSV call set, we manually inspected a 20-Kb window using IGV to confirm any false positive and negative variants. When the occurrence of dnSNVs and dnINDELs was confirmed based on the same criteria used for detecting late germline dnSVs, we checked the transmission to investigate the linkage. When a pair of DNMs belonging to the same mutation cluster does not segregate together (i.e. imperfect linkage), we consider it evidence of mosaicism. Furthermore, we obtained amplicon sequencing of dnSNVs and dnINDELs (unpublished data). Allelic imbalance in the amplicon sequencing data was viewed as evidence of early mosaic mutation.

5.6.9. Fine-scale molecular characterisation of dnSVs

We scanned the overlap between dnSVs and genic features using UCSC Genome Browser and the ARS-UCD1.2 bovine genome assembly. To obtain regulatory elements maps of the bovine liver, we downloaded liver ChIP-seq data (H3K27ac and H3K4me3) generated from four bulls from ArrayExpress (E-MTAB-2633; Villar et al. 2015). This ChIP-seq data was aligned to the bovine reference genome ARS-UCD1.2 using Bowtie2 (Langmead and Salzberg 2012), and

peaks were called using MACS2 (Zhang et al. 2008). To complement the scarcity of the publicly available bovine epigenomic maps, we obtained human orthologous region using UCSC LiftOver function (The ENCODE Project Consortium 2012).

5.7. Acknowledgements

The Dutch HF whole genome sequence population data set was funded by the DAMONA ERC advanced grant to MG. CC is senior research associate from the Fonds de la Recherche Scientifique–FNRS (F.R.S.-FNRS).

5.8. Research Funding

YLL, ACB, MB, MAMG, and RFV are financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. This work was supported by grants from the European Research Council (Damona to MG; award number: ERC AdG-GA323030). GCMM is post-doctoral fellow of the H2020 EU project BovReg (Grant agreement number: 815668).

5.9. Competing interests

EM is an employee of CRV B.V., one of the partners of the Breed4Food consortium. All other authors declare that they have no conflict of interest.

5.10. Authors' contributions

MG and CC contributed to conception of the study. EM generated multi-generational bovine pedigrees registered in Dutch cattle breeding programme. WC and LK generated WGS data. GCMM and CH mapped the data. YLL and CC performed the analyses and YLL drafted the manuscript under supervision of CC, MG, AB, MB, RV, and MAMG

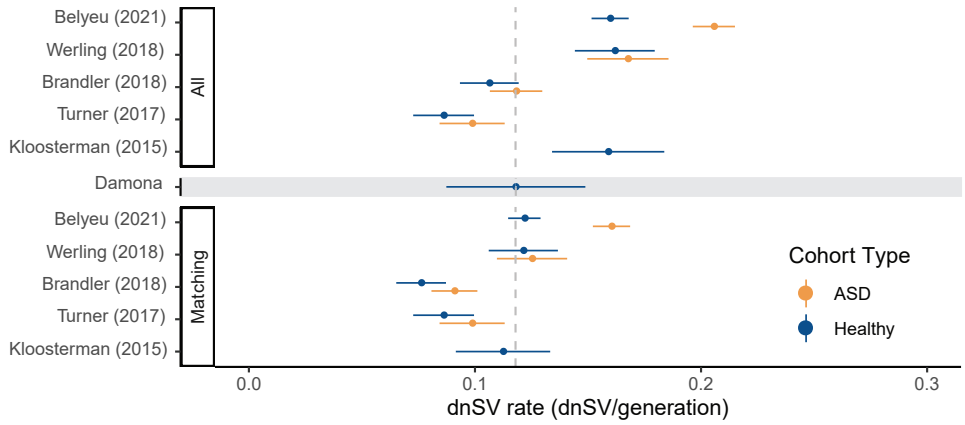
5.11. Supplementary Table

5.11.1. De novo SV rate compared to human cohorts

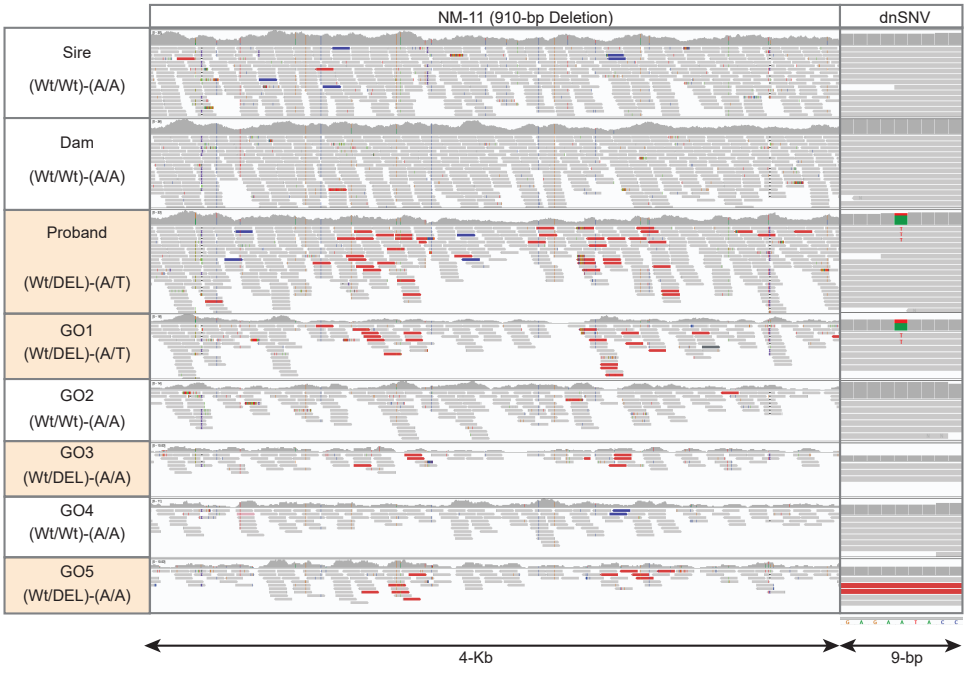
Study	Paper	WGS data spec	Data structure
Damona		WGS (sperm of blood; 26X) Insert length: 550bp Read length: 100bp	multi-generation pedigrees with multiple HS & GO
Belyeu (2021)	De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families DOI: 10.1101/2020.10.06.329011	Seqeuncing data information is not provided in the manuscript	2 generation pedgirees 2,363 ASD trios 2,372 healthy trios Healthy trios include 434 CEPH trios (3 generation pedigrees)
Brandler (2018)	Paternally inherited cis-regulatory structural variants are associated with autism DOI: 10.1126/science.aan2261	WGS (blood, 42.6X) REACH cohort (Read length: 150bp; 50X) SSC1 cohort (Read length: 150bp Insert sizes: 348-420bp; 39-50X)	REACH cohort 112 ASD trios 362 healthy trios SSC1 cohort 517 Autism quads (parents, an autism proband, and a healthy sibling) 2 generation pedigrees
Werling (2018)	An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder DOI: 10.1038/s41588-018-0107-y	WGS (blood; 37.8X) Insert length: 423bp Read length: 150bp	SFARI consortium 519 Autism quads (parents, an autism proband, and a healthy sibling) 2 generation pedigrees
Turner (2017)	Genomic Patterns of De Novo Mutation in Simplex Autism DOI: 10.1016/j.cell.2017.08.047	WGS (blood; 34.8X) Insert length: 423bp	SFARI consortium 476 Autism quads (parents, an autism proband, and a healthy sibling) 2 generation pedigrees
Kloosterman (2015)	Characteristics of de novo structural changes in the human genome DOI: 10.1101/gr.185041.114.19	WGS (blood, 14X) Insert length: 500bp Read length: 90bp	258 healthy trios 2 generation pedigrees from GoNL project

Resolution	SV type	Cohort information	No. DNM	No. trios	Mutation rate/haploid	Mutation rate/gen	1/N birth
50bp	CNV, INV	Healthy cattle	15	127	0.059	0.118	8.5
50bp	CNV, INV	Healthy cattle	5	127	0.020	0.039	25.4
50bp	CNV, MEI, complex	ASD + Healthy	865	4735	0.091	0.183	5.5
50bp	CNV, INV	ASD + Healthy	733	4735	0.077	0.155	6.5
50bp	CNV, MEI, complex	ASD	486	2363	0.103	0.206	4.9
50bp	CNV, INV	ASD	379	2363	0.080	0.160	6.2
50bp	CNV, MEI, complex	Healthy	379	2372	0.080	0.160	6.3
50bp	CNV, INV	Healthy	289	2372	0.061	0.122	8.2
50bp	CNV, MEI, complex	CEPH (Healthy)	62	434	0.071	0.143	7.0
50bp	CNV, INV	CEPH (Healthy)	52	434	0.060	0.120	8.3
43bp	CNV, INV, MEI, complex	ASD + Healthy	175	1510	0.058	0.116	8.6
50bp	CNV, INV	ASD + Healthy	135	1510	0.045	0.089	11.2
43bp	CNV, INV, MEI, complex	ASD	104	880	0.059	0.118	8.5
50bp	CNV, INV	ASD	80	880	0.045	0.091	11.0
43bp	CNV, INV, MEI, complex	Healthy	67	630	0.053	0.106	9.4
50bp	CNV, INV	Healthy	48	630	0.038	0.076	13.1
50bp	CNV, MEI, complex	ASD + Healthy	171	1038	0.082	0.165	6.1
50bp	CNV	ASD + Healthy	127	1038	0.061	0.122	8.2
50bp	CNV, MEI, complex	ASD	87	519	0.084	0.168	6.0
50bp	CNV	ASD	65	519	0.063	0.125	8.0
50bp	CNV, MEI, complex	Healthy	84	519	0.081	0.162	6.2
50bp	CNV	Healthy	63	519	0.061	0.121	8.2
201bp	CNV	ASD + Healthy	88	952	0.046	0.092	10.8
201bp	CNV	ASD	47	476	0.049	0.099	10.1
201bp	CNV	Healthy	41	476	0.043	0.086	11.6
20bp	CNV, MEI, complex	Healthy	41	258	0.079	0.159	6.3
50bp	CNV	Healthy	29	258	0.056	0.112	8.9

5.12. Supplementary figures



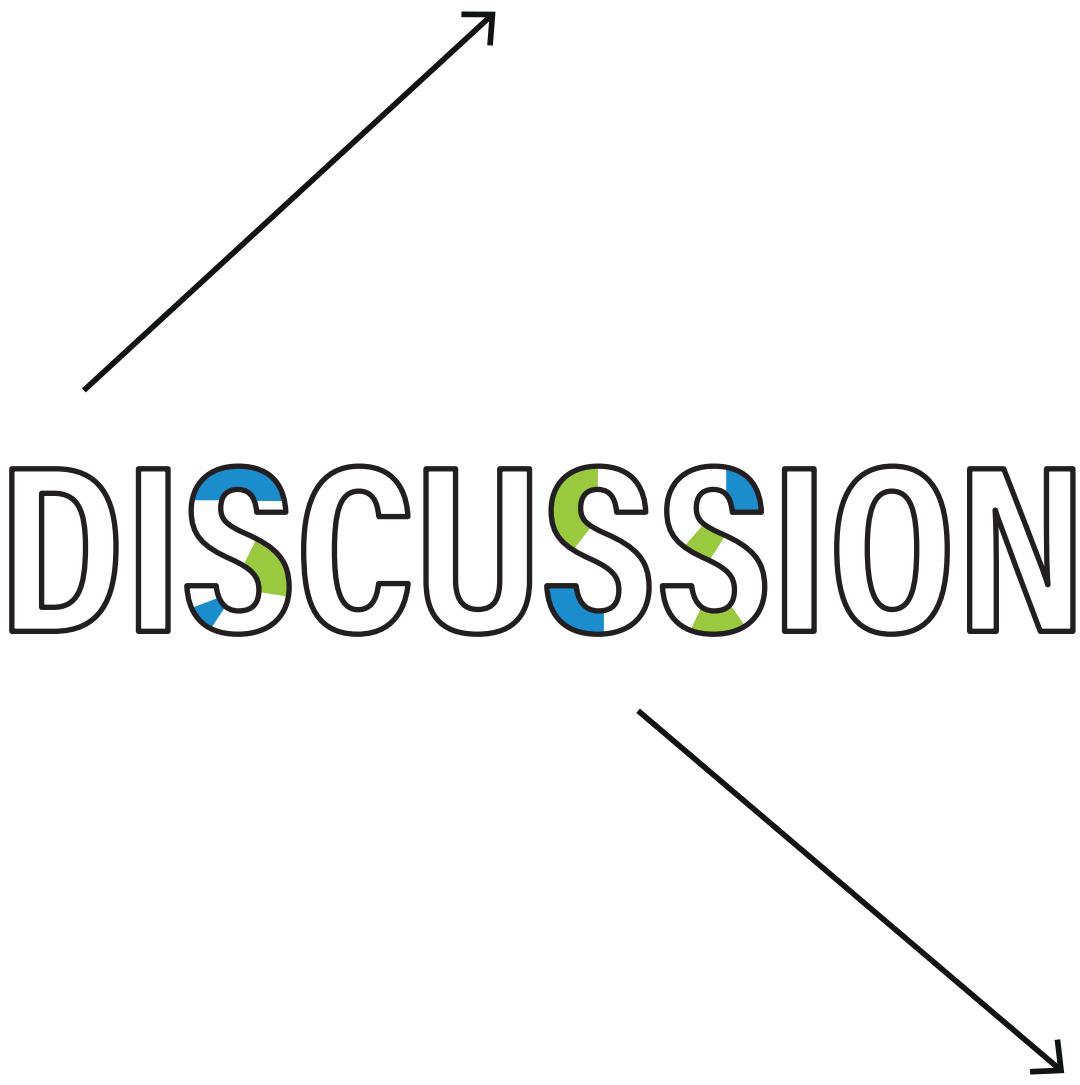
5.12.1. Supplementary Figure 1 dnSV rate. Panel above shows the dnSV rates including all SVs discovered in each study (including mobile elements insersion and complex events); the lower panel is identical to the main figure 5.2.



5.12.2. Supplementary Figure 2 IGV screenshot for a clustered mutation NM-11 and its pair dnSNV. Despite the 910-bp de novo deletion in the proband was transmitted to three GOs (GO 1;3;5), the paired dnSNV was transmitted to only one GO (GO1), manifesting imperfect linkage.

5

GENERAL


DISCUSSION

6

6.1. Introduction

Genome and environment together greatly influence how an organism develops and functions. Efforts to decipher how the genome relates to its phenotype in cattle, a livestock species with high economic significance, thus far have been mostly focused on single nucleotide polymorphisms (SNPs). Hence, exploration of complex and less tractable genetic variants, such as structural variants (SVs), have been scarce. This thesis presents an in-depth investigation of bovine structural variants. I generated two bovine SV catalogues, using a high-density SNP array and WGS data (Chapters 2 and 3). Of the numerous SVs discovered in the catalogues, I focused on highlighting variants at two extreme ends of the frequency spectrum: the GC CNV, segregating at a high frequency (Chapter 4), and an extreme form of rare SVs, *de novo* SVs (dnSVs; Chapter 5).

In this discussion, I compare two functional SVs discovered in this thesis (Chapters 3 and 4) and reflect on the evolutionary forces governing the fate of these SVs. Subsequently, I discuss the extreme form of rare SVs: dnSVs arising in bovine germlines. I discuss how our views on the source of *de novo* mutations (DNMs) have evolved over time and propose a model that fits best with extreme paternal bias and the IVF effect (Chapter 5). Furthermore, I discuss trends and developments in SV research that will advance our knowledge of SVs, and end the discussion with suggestions for discovery and utilization of functional SVs.

6.2. Evolutionary forces governing the fate of duplications and mCNVs

In this thesis, I exploited a wealth of genomic and phenotypic data obtained from dairy cattle populations, to gain insights into high impact SVs. Here, I compare the GC CNV and the ORM1 duplication in terms of the structural alleles segregating at each locus, and discuss the differences between these two SVs. From here onwards, I use the term "functional SVs" to refer to SVs that alter gene functioning, for instance, via eQTL mapping or GWAS association.

6.2.1. The structural alleles at the GC CNV and the ORM1 duplication

The two functional SVs in this context, the GC CNV and ORM1 duplication, are similar in that they are associated or implied to have association with economically important traits in dairy cattle (Lee et al. 2021; Brown et al. 2021; McGuckin et al. 2020). On the other hand, the type of SV (mCNV vs. duplication) and thereby the structural alleles segregating at these loci differ, which may reflect different evolutionary forces operating on the respective locus. Below, I elaborate on these loci; their differences are illustrated in Figure 6.1.

The GC CNV is a multiallelic CNV, where the diploid CNs ranged from 2 and 11. This locus harbours two main structural alleles: CN1 (wild type) and CN4 are the main alleles and two rare alleles: CN5 and CN6. The rare alleles were only present in a small number of related animals within our deep WGS cohort (Figure 6.1). The long stretch of IBD haplotypes shared

across CNs 4-6 animals implied that CNs 5-6 alleles are recurrent mutations that arose from the CN4 allele. The SV catalogue generated by Kommadath et al (2019) confirmed that the GC CNV is segregating in both dairy and beef cattle breeds. However, due to heterogenous sequencing coverage used for constructing the SV catalogues, it remains unknown whether the GC CNV carriers are the carriers of the rare, CNs 5-6 allele carriers.

The ORM1 duplication is a highly frequent biallelic duplication (Wild type allele (Wt)=0.51, duplicated allele (DUP)=0.49), harbouring two structural alleles, CN1 and CN2 (Figure 6.1). Presence of the ORM1 duplication was reported in multiple beef cattle breeds (Kommadath et al. 2019), but with at most diploid CN4 (CN2/CN2), and thus is likely biallelic in these other breeds as well. To further consolidate our findings on the biallelic state of the ORM1 duplication, we genotyped an additional 815 HF animals for the ORM1 duplication, using the direct genotyping approach. The genotyping results showed three canonical clusters, further supporting a biallelic status (results not shown).

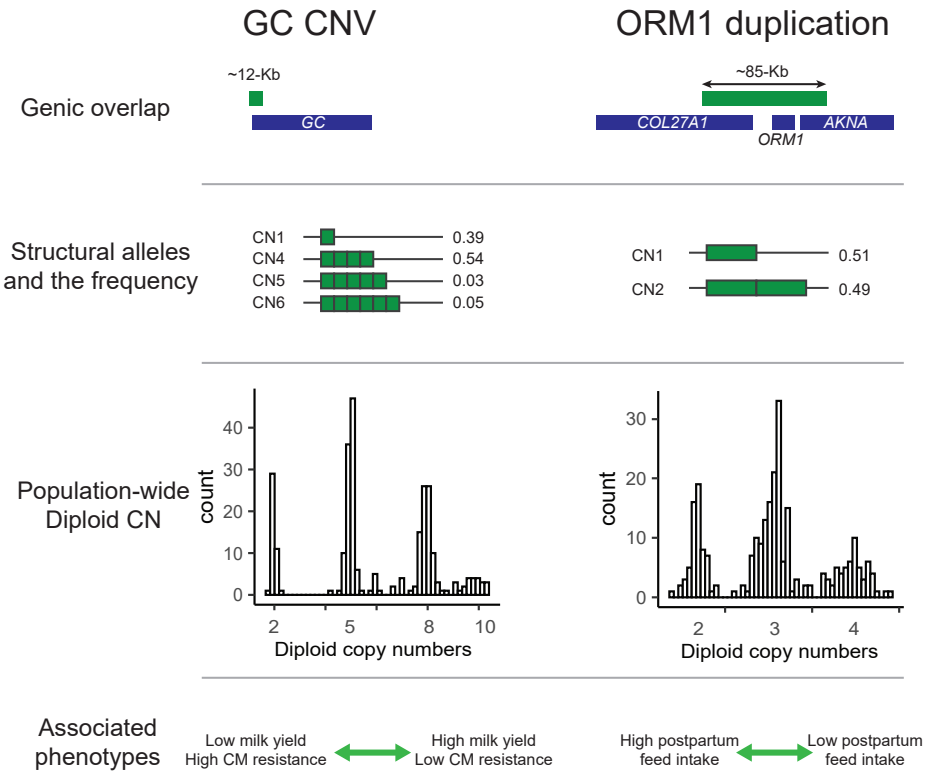


Figure 6.1. A schematic overview on the structural alleles and the population wide distribution of the GC CNV and the ORM1 duplication. The genic overlap, structural alleles, and their frequency are shown. In the bottom figure, the diploid copy numbers of the 266 deep WGS animals .The phenotypic effects associated with the copy numbers are annotated above the CN peaks. The associated phenotypes for the GC CNV was based on the GWAS catalogue obtained from the same study population. The associated phenotypes for the ORM1 duplication is based on literature (Brown et al. 2021;McGuckin et al. 2020)

Overall, presence of these SVs in other populations, including beef breeds, suggests that both SVs likely arose before the formation of modern dairy cattle breeds. When compared to the population-wide biallelic duplications which are predominantly rare (Chapter 3), it is evident that common duplications, like that of ORM1, are exceptions rather than the norm. On the other hand, the 24 mCNV loci discovered in the WGS-based SV catalogue (Chapter 3) were shown to harbour frequent multi-copy alleles, similar to GC CNVs. To sum up, our population-wide SV duplications and mCNVs point towards a seemingly contradictory phenomenon where duplications loci are enriched for rare variants (but depleted for common variants), whereas mCNVs harbours common alleles.

6.2.2. Expansion and contraction of duplications

Our bovine SV catalogue points towards depletion of common biallelic duplication. This phenomenon has been consistently found in multiple human SV catalogues, including large-scale catalogues based on >10,000 deep WGS genomes (Scott et al. 2021; Abel et al. 2020; Collins et al. 2020; Sudmant et al. 2015). On the other hand, human genomes harbour ~1,300 mCNVs, many of which harbour multicopy structural alleles (analogous to the GC CNV) and are strongly associated with gene dosage variation (Handsaker et al. 2015; Scott et al. 2021). This observations, in short, suggest that biallelic duplications remain as rare biallelic variants or expand to mCNV, harbouring multiple highly polymorphic alleles. One possible explanation for this phenomenon is that a *de novo* duplication event (haploid CN1 -> CN2) rarely occurs; however, once several multi-copy alleles segregate, contraction or expansion of copies can occur easily, for instance, via non-allelic homologous recombination (NAHR; Broad Institute and Handsaker 2019). This explanation indicates that CN2/CN2 animals can form gametes with CN1 and CN3, respectively, underlining that the duplication is unstable.

6.2.2.1. An example of unstable duplication in the Bar locus in Drosophila

The mutational instability in duplications is not entirely novel. Among multiple examples showing the contraction and expansion of duplications, here I elaborate on two sex-linked duplication loci reported in Drosophila and chicken. Firstly, unstable duplications giving rise to the revertant wild type allele were reported as early as 1917. In Drosophila, normal flies have round eye shapes with high facet numbers. However, X-linked Bar allele carrying males and homozygous females have low facet numbers with slit-like eyes. In contrast, heterozygous Bar carrying females have an intermediate number of facets, with kidney bean-shaped eyes (Tice 1914). While selecting Drosophila for low facet numbers (thus selection towards slit-like eyes), May (1917) discovered that some progenies are born with round eyes, an indication of reversion. This locus showed a reversion rate of ~0.001 (Zeleny 1919), and later an experimental validation underlined that non-equal crossovers give rise to two unbalanced gametes, where one receives a wild type allele, and the other receives a triplicated allele (Sturtevant and Morgan 1923).

6.2.2.2. An example of unstable duplication in the late feathering trait in chicken

A similar revertant phenomenon was reported in chickens. The late feathering trait is associ-

ated with a 176-kb duplication located on the Z chromosome. The duplicated allele harbours an avian endogenous virus gene (*EV21*) (Elferink et al. 2008). Purebred offspring from the late feathering strain are expected to show a late feathering phenotype; however, often revertant female offspring (in this context, visually distinguishable due to the early feathering phenotype) are born at a frequency of 0.001 (Ron Okimoto, pers. comm.). In this example, a triplicated allele was not reported, and hence a looping-out mechanism which would delete one of two copies within a single allele, was suggested (Levin and Smith 1990).

These two examples allude to repeat sequences (Roo transposon for the Bar locus in Drosophila, and *EV21* for the late feathering locus in chicken) that may cause unequal crossovers, leading to genome instability. NAHR has been widely adopted as one of the key mechanisms inducing CNVs, also resulting in unequal crossovers (Hastings et al. 2009). However, the reports in humans mostly consider low copy repeats (LCR), in other words segmental duplications (>1-kb) as the cause of NAHR. To what degree LCR can induce multiplicated alleles has not been systematically investigated.

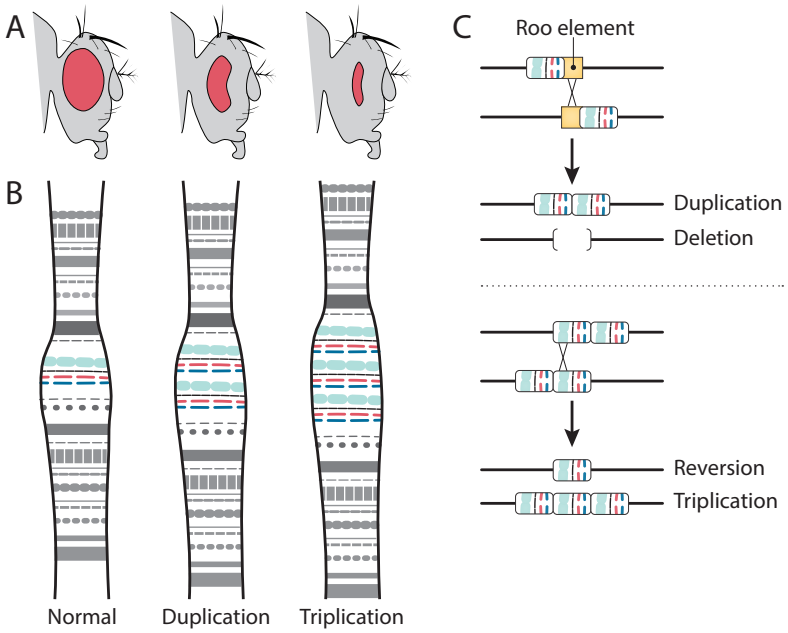


Figure 6.2. Bar alleles and their phenotypes in Drosophila. (**A**) Wild type Drosophila have round eyes. Flies homozygous (or hemizygous) for the Bar mutation have thin, slit-like eyes. Flies homozygous (or hemizygous) for the double-Bar mutation have even smaller eyes. (**B**) Schematic of the Bar region of polytene chromosomes. The Bar mutation is a tandem duplication and double-Bar a tandem triplication of the region. (**C**) The Bar mutation arose by unequal crossing over between two Roo transposable elements (yellow), resulting in a tandem duplication. Reversion and triplication alleles arose from the Bar mutant by unequal cross- over between duplicates that had aligned out of register. Figure courtesy by Wolfer and Miller (2016).

6.2.3. Possible explanations of the biallelic state of the ORM1 duplication

To summarize, duplicated alleles can contract (reversion) or expand (becoming mCNV loci), likely using repeat sequences as substrates. This view posits a question regarding the mutational mechanisms of the ORM1 duplication: Why and how the ORM1 duplication persists as a common biallelic duplication? As duplications segregating at high allele frequency are not common, it is tempting to speculate that the high frequency of the duplicated allele may be associated with adaptive advantages. If that is the case, it may be logical to expect a triplicate allele that may confer even stronger adaptive advantages. However, the ORM1 duplication remains biallelic, not attaining triplicate allele(s) and becoming an mCNV locus (confirmed in the current study population and beef cattle breeds).

There may be two explanations for this phenomenon: (i) the local genomic context lacks substrates to give rise to NAHR (e.g. repeat sequences), hence a triplicated allele does not arise, or (ii) a triplicated allele may be formed via NAHR, as demonstrated in the Bar allele in Drosophila, but does not persist, possibly due to selective disadvantage. For instance, the triplicated allele may become deleterious due to a disrupted gene dosage, and may thus be selected against.

Without excluding that absence of NAHR as a potential explanation for the observed lack of triplicated alleles, a study by Lan and Pritchard (2016) lends further support for the negative selection on a newly emerged triplicate allele. Lan and Pritchard (2016) conducted an empirical study utilizing human and mouse multi-tissue transcriptome data and showed that most young, tandemly duplicated genes are expressed at a lower level than a canonical single gene copy (Lan and Pritchard 2016). This is because these copies are likely governed by the same regulatory elements, prohibiting them from obtaining novel expression profiles. Consequently, the sum of gene expression of the original and the tandemly duplicated copy together would be equivalent to that of single-copy genes. In short, an abrupt increase of gene dosage through copy gain events can disrupt the dosage balance, and in turn reduce chances of survival of the young gene duplicate, making dosage sharing essential for survival (Lan and Pritchard 2016).

Based on the findings by Lan and Pritchard (2016), I speculate that the ORM1 duplication may not obtain more than two copies, possibly because its expression cannot be downregulated sufficiently for a triplicate allele to survive (insufficient negative feedback).

6.3. *De novo* structural variants

In chapter 5, I utilized 127 deeply sequenced bovine pedigrees, and discovered 20 *de novo* SVs (dnSVs). Compared to *de novo* single nucleotide variants (dnSNVs), for which multiple large scale cohorts were investigated regarding the underlying mutagenesis, the investigation on dnSVs has been sparse. Hence, our findings on dnSVs are invaluable in understanding distinctive mutagenesis of SVs. We observed an extreme paternal bias among the germline dnSVs (14:1). Also, our findings point towards a strong bias toward IVF probands, underlining that certain assisted reproduction technologies (ART) can be mutagenic (Table 5.1). Additionally, I discovered five early mosaic dnSVs that account for 21% of all detectable dnSVs. Here, I discuss two different models on the mutagenesis of DNMs, and elaborate on the model that fits best with our findings. Furthermore, I explain why multi-generational extended pedigrees are important in accurate estimations of germline dnSVs. Additionally, I elaborate on the value of bovine pedigrees as a model for future studies. Below, I use the term *de novo* mutations (DNMs) to referred to any type of *de novo* variants, whereas I use dnSVs or dnSNVs to refer to particular type of DNMs.

6.3.1. Source of DNMs

Even before the direct estimation of mutation rate was possible (e.g. using sequencing), classical geneticists considered the male germline as the main contributor of DNMs. Haldane (1947, 1935) estimated the mutation rate for hemophilia, an x-linked recessive disease, and concluded that the mutation rate is much higher in males than in females. With the development of molecular biology, it has become possible to assign the parent-of-origin for autosomal *de novo* diseases, for which the paternal bias was confirmed again. One fairly obvious explanation for the paternal bias is the greater number of cell divisions in the male germline, compared to the female counterpart (Crow 1997). Here, I explain (i) how this view, the so called "replication-driven model", has been both confirmed and rejected by studies in the sequencing era and (ii) why I consider our dnSV findings to not fit well with the replication-driven model, and propose to consider DNA damage as additional source of DNMs.

6.3.1.1. The replication-driven model

An empirical estimate of human germline dnSNV rate from 78 sequenced Icelandic pedigrees showed that ~80% of dnSNVs arose in the male germline. Also, the paternal age was strongly correlated with the number of dnSNVs, indicating that an aging paternal germline accumulated dnSNVs at a rate of one dnSNV per year (Kong et al. 2012). These two factors, sex and age, were subsequently confirmed in multiple human cohorts (Michaelson et al. 2012; Goldmann et al. 2016). These disproportionally high paternal dnSNVs were explained by the difference in male and female gametogenesis. Unlike the female germline that does not undergo postnatal replication, the male germline undergoes continuous replication from puberty onwards. Hence, erroneous replication leading to dnSNVs in spermatogonia stem cells can accrue dnSNVs, as the male germline ages.

6.3.1.2. DNA damage induced dnSNVs

Accumulation of WGS data in humans and other species confirmed the sex and age effect in dnSNVs, supporting the replication driven model described above (Venn et al. 2014; Harland et al. 2017; Thomas et al. 2018). However, a cross-species germline dnSNV comparison cast doubt on the replication-driven model (Wu et al. 2020b). For instance, if dnSNVs accrued in direct relation to the age of the male germline, human paternal bias would be twice as strong as that in baboons. This is because a human male will have undergone many more replications at the age of conception (e.g. from puberty to the moment of conception of offspring) than baboons, which mate at a younger age (Wu et al. 2020a). However, the paternal bias in humans and baboons was similar (4:1). This finding underlines that the number of male germline replications cannot fully explain paternal bias, requiring an additional model.

Other, large-scale human dnSNV studies discovered the maternal age effect (Wong et al. 2016; Goldmann et al. 2016). At a maternal age of > 40, an exponential increase of a maternal age effect explained the data much better than a linear model (Gao et al. 2019). Notably, the emergence of clustered DNMs (cDNMs) was correlated with maternal age, and the maternal cDNMs had a distinctive mutational signature (C>G), suggesting a DNA damage origin instead of a replicative origin (Goldmann et al. 2018). Gao et al. (2019) concluded that the suboptimal DNA repair capacity of ageing oocytes could result in increased dnSNVs in both maternal and paternal genomes. Furthermore, the incidence of dnSV accompanied by small DNMs in physical proximity was also higher in ageing oocytes, likely due to compromised double-strand break (DSB) repair in oocytes (Goldmann et al. 2018). Together, these studies unravelled female germline driven DNMs, which were not caused by replication errors, but by a suboptimal DNA repair function. This novel finding is in contrast to the textbook view that dnSNVs arise from replication errors in the male germline.

6.3.1.3. Weak evidence for replication model in dnSVs

It took long until investigating whether the replication driven model is also applicable to dnSVs, as studying dnSVs requires a much larger cohort than dnSNVs studies, since dnSVs are much rarer than dnSNVs. A dnSV screen in a healthy human cohort (~250 trios) showed a paternal bias of 2.2:1, but without significant age effect (Kloosterman et al. 2015). A more recent and larger study on human dnSVs (~4,000 trios) confirmed these earlier findings: there was no significant paternal age effect, although a paternal bias was shown (2.6:1, Belyeu et al. 2021). Together, these human dnSV studies suggest that (i) the paternal bias indicates that the male germline is more prone to give rise to dnSVs; however, (ii) dnSVs are independent from paternal age, meaning that dnSVs may arise after DNA replication process. Mutations (both dnSVs and dnSNVs) arising after the replication would not bear age effect, as they do not accumulate in the spermatogonial stem cells. Take these together, the replication model lenders weak support for the emergence of dnSVs.

6.3.1.4. Post-meiotic sperm DNA damage leading to paternal bias

The non-age dependent paternal bias shown in human studies points towards post-meiotic

emergence of dnSVs. Thus, I consider post-meiotic sperm DNA damage and its subsequent repair by oocytes as the major source of dnSVs. Male and female gametes undergo distinctive trajectories of differentiation. As for male gametogenesis, spermatogonial stem cells undergo meiosis, resulting in spermatids, which further undergo spermiogenesis (post-meiotic maturation steps, including removal of cytoplasm, compaction of nuclei, stripping of histones and replacement by protamines), resulting in mature sperm cells (Champroux et al. 2016). Although these maturation steps induce DNA lesions, particularly DSBs (Grégoire et al. 2018), these cannot be repaired by sperm cells themselves due to suppressed transcription; hence, these lesions are left unrepaired until fertilization. On the other hand, oocytes can transcribe and accumulate mRNA, which can later repair sperm lesions upon fertilization. Thus, DNA repair in a freshly fertilized zygote is considered a maternal trait (García-Rodríguez et al. 2019). Taken together, post-meiotic DNA damage in sperm cells seems to hold the key to the non-age dependent paternal bias, hinting that the DNA damage model may be more adequate to explain the observation.

6.3.1.5. What explains the IVF effect?

Finally, our dnSVs unraveled a strong IVF effect (Chapter 5). Such elevated dnSVs in IVF probands may be due to (i) increased DSBs in sperm cells, (ii) decreased DNA repair capacity of *in vitro* matured (IVM) oocytes compared to *in vivo* oocytes, or (iii) both. Given that all probands (AI-, FE-, and IVF- produced ones) are fertilized from frozen semen, the amount of spermiogenesis induced DSBs is assumed constant regardless of the ART applied. The fertilization condition in IVF differs from AI and FE, in both of the latter the thawed semen is injected into the vaginal tract. In IVF, sperm cells might accrue more DSBs until fertilization in a petri dish, thus leading to high dnSVs in a proband. Alternatively, IVM oocytes might have a suboptimal capacity to repair DNA damage compared to *in vivo* oocytes (analogous to the ageing oocytes reported in humans (Gao et al. 2019; Goldmann et al. 2018)). IVM oocytes have altered transcriptional activities, suggesting that the accumulation of mRNA for DNA repair is not comparable to *in vivo* oocytes (Erhardt et al. 2003; Katz-Jaffe et al. 2009). To disentangle the true origin of the IVF effect in dnSVs and attribute the source of dnSVs either to sperm cells or IVM oocytes requires follow-up research.

6.3.2. Bovine pedigrees as a model

Thus far, I have discussed the underlying mechanisms of bovine germline dnSVs based on the findings in this thesis (Chapter 5). Here, I bring the discussion to a broader context, by reflecting on the data set used in this thesis, and by discussing why this data set is suitable for studying DNMs. Finally, I end the discussion by proposing bovine pedigrees as a model for DNM studies.

Studies on DNMs can shed light on fundamental evolutionary questions: how, when and where do DNMs arise? Many studies aimed at answering these questions by detecting germline DNMs in two-generational WGS pedigree data (Kong et al. 2012; Michaelson et al. 2012). A large number of two-generational pedigrees have sufficient statistical power to distill mutagenic

factors, such as sex and age effects (Goldmann et al. 2016). However, pedigree structures other than the two-generational trio structure have demonstrated their value in elucidating the timing of DNMs, both of germline and early mosaic DNMs (Harland et al. 2017; Rahbari et al. 2016; Sasani et al. 2019).

We utilized three-generational extended pedigrees, where a trio pedigree, consisting of a proband, sire, and dam, complemented with multiple half-siblings (2nd generation) and offspring (3rd generation) of a proband. This pedigree structure is crucial for confident detection of (i) germline dnSVs, based on transmission to the 3rd generation, and (ii) early mosaic dnSVs, based on imperfect linkage signature. Our dnSV discovery results, consisting of 15 germline dnSVs and 5 early mosaic dnSVs, point towards two distinctive timings of dnSVs: the former arose during gametogenesis of mature animals, whereas the latter arose during early zygotic development. Lack of an intricate pedigree structure (e.g. probands are not complemented with half-siblings) may misclassify early mosaic DNMs in parents into late germline DNMs of the proband, resulting in an inaccurate estimation of the germline mutation rate (Harland et al. 2017).

Furthermore, the 127 probands investigated were obtained from three different assisted reproduction technologies (ARTs): AI, MOET, and IVF. ARTs have been widely adopted both in human and bovine reproduction: globally, ~400K babies are born annually via ART (Duranthon and Chavatte-Palmer 2018), and ~800K bovine embryos were produced *in vitro* in 2019 (Viana 2020)). Despite this wide adoption of ARTs, their impact on DNMs has been difficult to measure in humans. This is because (i) IVF probands were a small fraction of the total cohort, and hence were removed to avoid the introduction of IVF-related bias in a DNM study (Wong et al. 2016), or (ii) DNM study samples were recruited from a disease cohort, which can be confounded with possible IVF effects (e.g. IVF might be used due to illness in parents, hence increased DNMs cannot be confidently attributed to either IVF or disease; Wang et al. 2021).

To summarize, many DNM studies have focused on knowledge gaps regarding endogenous mutagenic factors, such as sex and age effects, exploiting two-generational pedigrees. Multi-generational extended pedigrees have the potential to elucidate DNMs arising in different developmental stages. However, the use of multi-generational pedigrees has been limited in humans, likely due to the long generational interval (Sasani et al. 2019; Jónsson et al. 2018; Rahbari et al. 2016) On the contrary, generating WGS data on such intricate pedigrees is feasible in livestock species: breeding programmes not only have multiple generations of bovine pedigree records, but also generate extensive amounts of phenotype data, potentiating the study on impact of DNMs. Furthermore, ART is routinely practiced in cattle production, particularly on young and healthy animals. Taken together, I propose bovine populations as a powerful model for future studies on DNMs and ART effects on DNMs.

6.4. Trends and perspectives of SV research

In this thesis, I generated population-wide SV catalogues, using the SNP genotyping array and WGS data sets (Chapters 2 and 3). A comprehensive survey of SVs of diverse types and sizes, possibly accurately genotyped, is crucial, as it can potentiate the discovery of high-impact SVs (Chapter 4) and the detection of dnSVs (Chapter 5). Below, I discuss how recent advances in SV detection software packages will expand both array-based and short read WGS-based SV catalogues. Next, I discuss the developments in long-read sequencing (LRS) data.

6.4.1. SV discovery in array data

The BovineHD array-based SV catalogue generated in this thesis contained 1,755 CNV regions. Based on this catalogue, I identified ~32 CNVs on average per bovine genome. The mean length of deletions and duplications was 44.2-kb and 74.6-kb, respectively (Chapter 2). The limitations of the array-based SV catalogue were evident: (i) detectable variants are limited to CNVs, (ii) breakpoints are unresolved, and (iii) detection resolution is low.

Despite these constraints, the BovineHD array-based catalogue contained multiple high-impact SVs, such as a 138-kb deletion located on the HH5 recessive lethal haplotype, ablating the entire *TBF1M* gene (Schütz et al. 2016), as well as a 85-kb *ORM1* duplication. It is worth noting that breeding programmes have genotyped many animals with BovineHD and other arrays. Thus, array-based SV detection will remain a cheap method for high-throughput screening for well-characterized large CNVs, despite the limitations explained above.

Notably, the current array-based catalogue was generated using the PennCNV software package, which exploits B allele frequency and probe hybridization intensity (Wang et al. 2007). A novel computational approach, which takes into account haplotype sharing, additionally to the B allele frequency and the probe hybridization intensity, resulted in a six-fold increased CNV detection sensitivity (Hujoel et al. 2021). Applying the methodology developed by Hujoel et al. (2021) to the already generated large-scale genotyping data available in the bovine breeding programmes will contribute to further expanding the array-based catalogue at no extra cost.

6.4.2. SV discovery in short-read sequencing data

Detection and genotyping of SVs are challenging, particularly in low pass WGS data (Huddleston and Eichler 2016). Nearly ~50% of SVs discovered in the human '1000 Genomes Project' are left with no-to-limited breakpoint resolution (mean coverage=7.4X; Sudmant et al. 2015). Our WGS data, obtained from 266 deeply sequenced genomes from a healthy HF family cohort (127 trios, mean coverage=26X) resulted in a superior SV catalogue, containing ~14,000 SVs with (i) improved detection of smaller variants, (ii) ~70% SVs with single base-pair resolution, and (iii) diverse types of SV (deletions, duplications, mCNVs, processed pseudogenes). Using this catalogue, we identified ~5,000 SVs per bovine genome (Chapter 3).

SV detection in Chapter 3 was performed based on a single hybrid caller, Lumpy (Layer et al. 2014), which takes into account two aberrant mapping signals (discordant read pairs and split

reads). I emphasize that despite the high quality of the SV catalogue generated in the current thesis, more SV can be discovered in the current deeply sequenced bovine genomes using advanced SV detection software packages. For instance, a preprint on SV detection in the deeply sequenced 1000 Genomes Project data discovered ~9,000 SVs per genome (Byrska-Bischop et al. 2021) by (i) combining two ensemble SV calling pipelines (combination of multiple SV callers) used in Abel et al. 2020 and Collins et al. 2020, (ii) a *de novo* assembly based insertion pipeline (github.com/nygenome/absinthe), and (iii) a post-hoc insertion genotyping tool (Chen et al. 2019). Byrska-Bischop et al. (2021)'s catalogue shows that detecting insertions, which are known to be highly challenging to detect in SRS, is possible by combining *de novo* assembly and post-hoc genotyping approaches. As noted earlier, a comprehensive survey of SVs provides a basis of discovery of high impact variants. Hence, applying novel SV detection methods (e.g. ensemble SV calling pipeline and *de novo* assembly based approach) to the already existing WGS data will make the SV catalogue even more comprehensive.

6.4.3. SV discovery in Long-read sequencing data

Long-read sequencing (LRS) technology utilized significantly longer DNA molecules than the conventional short-read sequencing (SRS) technology. Thus, SVs can be discovered even in the SRS-intractable regions (e.g. highly repetitive sequences), resulting in discovery of ~20,000 SVs per genome (Zhao et al. 2021). Also, SV detection from the SRS data mapped to a reference genome inherently has a reference genome bias, which can be resolved using *de novo* assembled LRS data (Ebert et al. 2021; Audano et al. 2019). Despite its comparably high costs (LRS costs ~7x more than SRS data (Zhao et al. 2021), a population-scale LRS genetic analysis was recently conducted on 3,622 Icelanders (Beyter et al. 2021). This LRS study revealed multiallelic SVs located in SRS-intractable regions. One such case was an exonic VNTR in *ACAN* gene, harbouring 11 alleles with a different copy number of a 57-bp motif. The number of motifs was in a linear relationship with the human height in the Icelandic population, pointing out that LRS-based SVs can contribute in expanding our knowledge of functional SVs.

On the other hand, a handful of LRS data can be used to construct a graph-based genome representation. Using previously generated SRS data, Ebert and colleagues discovered ~50% of ~100K SVs in haplotype-resolved LRS assemblies (Ebert et al. 2021). Notably, a SV graph constructed from 15 LRS assemblies to discover SVs in ~30X SRS data revealed SVs under adaptive evolution in humans (Yan et al. 2021). Likewise, multiple reference quality genomes have been published in cattle (Low et al. 2020; Heaton et al. 2021; Oppenheimer et al. 2021). A preprint on a bovine pangenome SV graph suggests that 50 *de novo* LRS assemblies per breed would suffice to capture within-breed SVs (Leonard et al. 2021).

Together, these technological advancements in LRS promise curation of a larger number of SVs, many of which (e.g. mobile genetic variants and VNTRs) are currently under-represented in SRS based SV detection. Future efforts to characterize such variants will better elucidate the roles of SVs in the genome-to-phenome connections.

6.5. Suggestions for the discovery and utilization of functional SVs

Here, I illustrate the workflow and underlying data used in my thesis for identifying functional SVs which can be used as a benchmark for future studies. Furthermore, I elaborate on how to genotype and utilize functional SVs (Figure 6.3).

6.5.1. Workflow and data for discovering functional SVs

Upon completion of the SV catalogues (chapters 2 and 3), I intersected them with an in-house SNP-based GWAS catalogue (Figure 6.3). The GWAS catalogue was generated based on ~4,000 daughter proven bulls, and on >50 traits routinely collected by the Dutch HF dairy cattle breeding programme. These bulls were genotyped with a 16K density array, which was sequentially imputed to 50K, to BovineHD and finally to WGS level. All animals used for construction of imputation panels were from the same HF population. In the absence of a GWAS catalogue generated for the breed of interest, I recommend to use a public GWAS catalogue (Jiang et al. 2019), and the animal QTL database (Hu et al. 2019)

As shown in chapters 2 and 3, the genotyping accuracy of SVs differs, depending on the type of variant. The implication of using accurately genotyped SVs in association analyses is that the association signal may not be as strong or evident as the underlying data ought to show. Thus, I foresee that SNP-based GWAS will remain the key tool in QTL- and fine-mapping, instead of direct association using SV genotypes, given suboptimal genotyping accuracy in SVs compared to SNPs. It is worth noting that the GC CNV and previously reported trait-associated SVs in livestock species were captured well by haplotypes (Derks et al. 2018; Kadri et al. 2014), as was shown in a large-scale haplotype-assisted CNV detection study (Hujoel et al. 2021). A haplotype-based fine-mapping approach (Zhang et al. 2012) can be an additional consideration.

The eQTL data was generated with the liver RNA-seq data of 178 HF cows. These animals were genotyped with a BovineHD array (770K) and then imputed to SVs and SNPs discovered in the WGS data. Imputing rare variants, as for many SVs, is challenging. Often, they are imputed with low accuracy and thus filtered out. Hence, to test for the effects of rare SVs, future eQTL mapping studies might consider generating WGS data to avoid imputation.

In our study, integration of eQTL and an epigenome map generated from a matching tissue type (liver) provided a mechanistic explanation of trait- and expression- association with SVs. Due to the scarcity of publicly available epigenome data, we relied on liver ChIP-seq data generated from bulls of an unknown breed (Villar et al. 2015). However, epigenetic marks and gene expression can differ depending on multiple factors, such as tissue-type, sex, and age (Ardlie et al. 2015; The ENCODE Project Consortium 2012). Future studies should investigate a biologically relevant tissue that underpins the trait of interest.

Additionally, the trait of interest in this thesis, clinical mastitis, can be observed in mammary glands, and thus seemed relevant to investigate. However, GC is not expressed in mammary

glands (Chapter 4). After establishing the causality of the *GC* gene, high *GC* expression in liver and a subsequent increase in clinical mastitis seems logical in retrospect. Thus, I recommend that future studies aiming at establishing molecular causality to investigate gene expression and epigenomic marks, are not only limited to the tissues where traits of interest are observed. Also, future studies should utilize human and bovine expression atlases (Ardlie et al. 2015; Fang et al. 2020) to determine biologically relevant tissues in establishing causality.

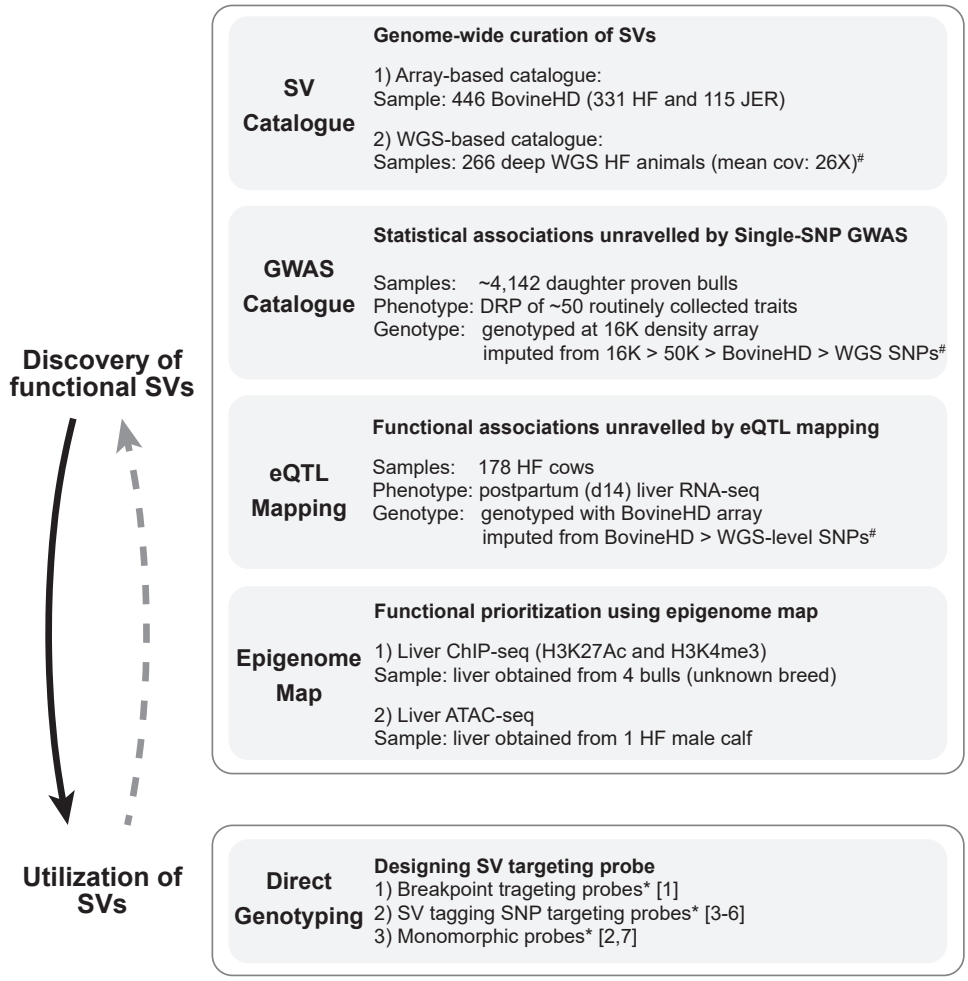


Figure 6.3. Genotyping results of the seven probes targeting GC CNV. We used seven probes (targeting proximal breakpoint, CNV tagging SNPs, and monomorphic probes) to genotype and gain better insights about the GC CNV. The genotyping results obtained based on genotyping an independent cohort are shown in the figure. Our genotyping results indicated that probe 3 and 6 worked equally well, whereas 4 and 5 did not work successfully. Probe 1, which targeted the breakpoint, worked partially, distinguishing CNV carriers from the rest, however it did not clearly cluster the CNV carriers. Two monomorphic probes confirmed two clusters that separated CNV carriers from the rest. However the probe 7 was suboptimal, showing heterozygous animals.

6.5.2. Utilizing functional SVs via direct genotyping

Direct genotyping can be beneficial in two aspects: Firstly, once a functional SV is identified and characterized, a large cohort can be genotyped for the SV to screen and monitor the presence of the SV (black solid downwards arrow in Figure 6.3). For example, I added multiple probes targeting the GC CNV, which can then be used for genomic prediction. Also, in chapter 5, we identified a putative deleterious *de novo* SV, a 50-Kb deletion ablating *CENPC* gene (Chapter 5). I added probes targeting this dnSV, such that the breeding programme can monitor the segregation of the putative risk variant. Of note, based on the genotyping results obtained in Chapter 3, this dnSV was not found in any of the 815 genotyped animals.

Secondly, direct genotyping results can be used to perform a direct association between the SV genotype and phenotype (grey dotted upwards arrow in Figure 6.3). It is worth noting that I relied on a SNP-based GWAS catalogue to discover SVs. The underlying assumption of this approach is that the SNPs are in high LD with the SVs of interest, thus the association signals from SNPs are sufficiently high to guide discovery of functional SVs. However, as elaborated earlier, unstable duplications can revert to wild type alleles. Such reversion involves switching of the background haplotype (e.g. from the haplotype shared among duplicated alleles (CN2), to one shared among wild type alleles (CN1)), resulting in low LD between the duplication and the surrounding SNPs. Thus, this 'bottom-up' approach (first genotype SVs in a large cohort and then associate the phenotypes) may pick up SV-driven association signals which are neglected by the approach used in the current thesis.

6.5.2.1. Challenges in direct genotyping of SVs

In chapter 3, I added 372 probes directly targeting CNVs, and ~ 80% were confirmed in an independent HF cohort. Most probes targeted breakpoints of biallelic CNVs. Here, I briefly discuss the CNV genotyping results, taking the results obtained from the GC CNV as an example. I further highlight some caveats of the direct genotyping approach.

To thoroughly genotype the GC CNV, we added seven probes, of which (i) one probe targeted the distal breakpoint (1), (ii) four probes targeted polymorphic SNPs tagging the GC CNV (3-6), and (iii) two probes targeted monomorphic positions that aimed to distinguish (CN1) individuals from CNV carriers (CNs4-6), and which could have enabled us to distinguish CN4 from CNs5-6 (2,7; marked with green arrows; Figure 6.4).

Genotyping results in this thesis confirmed that two polymorphic probes (3 and 6) and one monomorphic probe (2) worked successfully among the seven probes. This result demonstrates that genotyping CNVs requires multiple probes to determine which probe works best.

Lastly, it is worth noting that CNV genotyping results should be processed with caution. The SNP genotyping probes are designed to hybridize diploid DNA without structural complexity (e.g. gain or loss of DNA). However, most CNV probes used in this thesis targeted the breakpoints sequences where the gain or loss of DNA started (Chapter 3). Hence, default genotype clustering performed by the Illumina Genome Studio software may lead to suboptimal clustering for CNV targeting probes, leading to low genotyping accuracy. Likewise, I retrospectively

learned that some CNV probes were filtered out due to a stringent call rate threshold (0.99), despite a moderate-to-high call rate obtained (~0.98). Thus, blindly applying stringent filtering criteria set for SNP targeting probes may inadequately penalize the CNV probes. Consequently, it remains essential to conduct manual curation and inspection for clustering of CNV targeting probes.

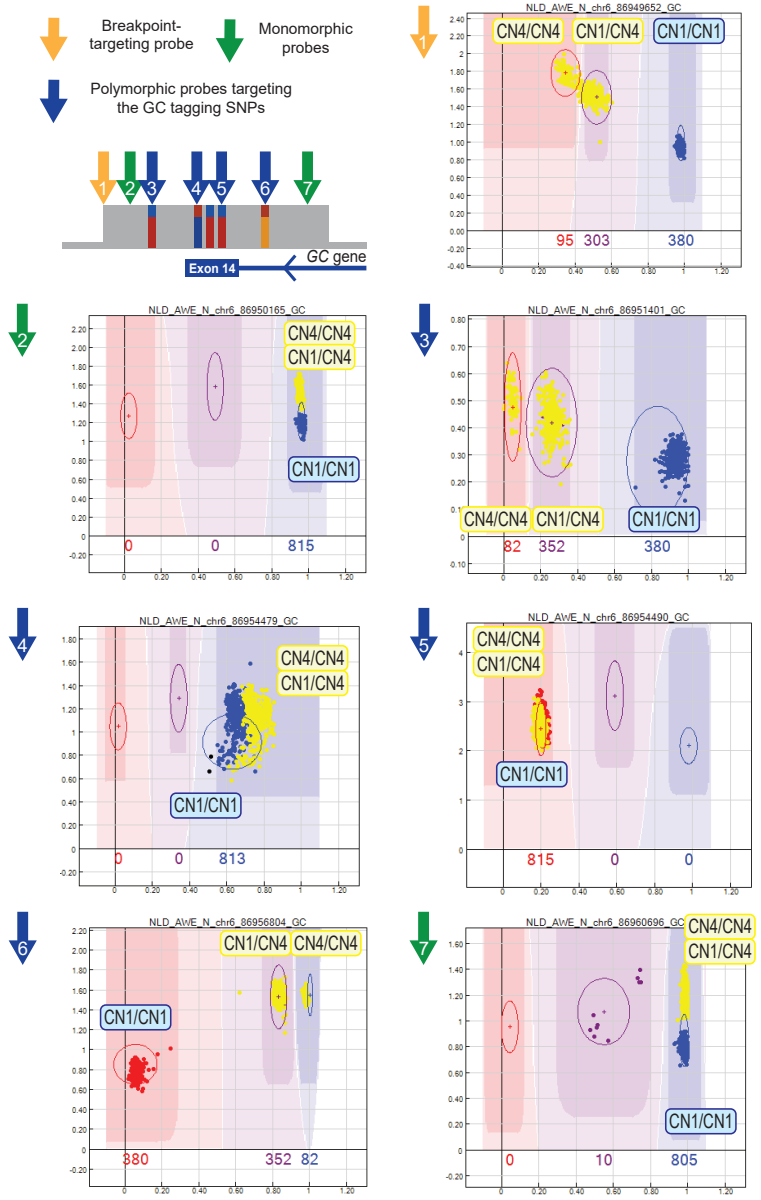


Figure 6.4. Genotyping results of the seven probes targeting GC CNV. We used seven probes (targeting proximal breakpoint, CNV tagging SNPs, and monomorphic probes) to genotype and gain better insights into the GC CNV. The genotyping results obtained from an independent cohort are shown in the figure.

6.5.3. Utilizing of SVs through though phasing and imputation

An ultimate aim of SV catalogue construction is the future inclusion of SVs into routine genetic analyses. One approach to include SVs in to routine genetic analyses is to exploit phasing and imputation, as done for SNPs (Hayes and Daetwyler 2019). Phasing and imputation of SVs has been shown feasible, as long as this is limited to simple variants (e.g. biallelic CNVs). Most simple SVs can be imputed with high accuracy, however, some achieved low imputation accuracy compared to SNPs and indels, likely reflecting the low genotyping accuracy of SVs (Hehir-Kwa et al. 2016; Chen et al. 2021). Likewise, phasing and imputation of complex multiallelic SVs are possible, but require extensive characterization of the underlying structural alleles (Boettger et al. 2012). Also, mCNVs are often formed via recurrent mutations, thus accurately imputing exact copy numbers is not trivial (Handsaker et al. 2015). As such, attempts to impute variable number tandem repeats (VNTRs), hypermutable SVs, have been lacking. Yet, phasing and imputing VNTRs is possible, by exploiting extensive amounts of sibling IBD haplotypes (~500K genotyped samples were imputed, using the reference panel constructed from ~50K UK Biobank whole exome sequenced samples; Mukamel et al. 2021).

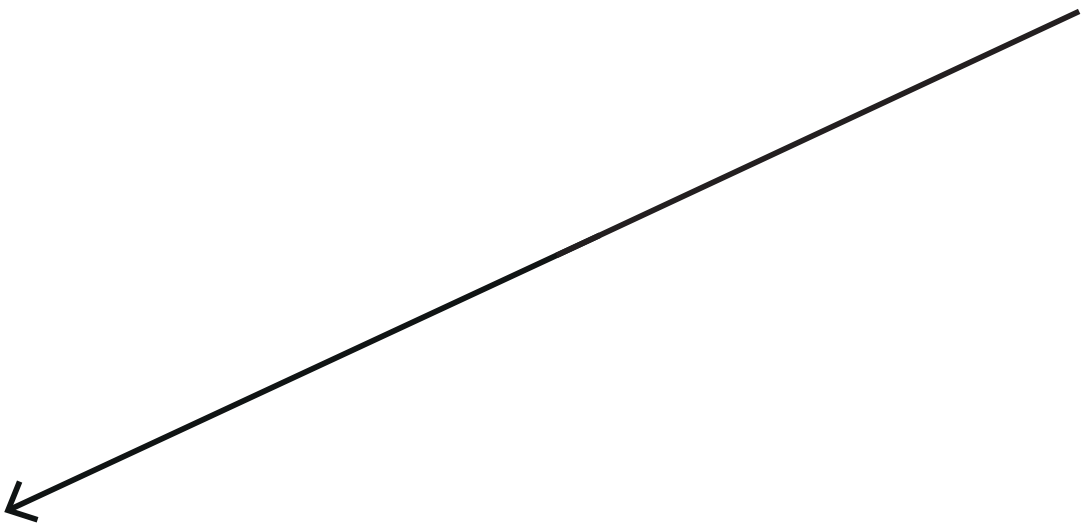
To summarize, phasing of imputation of SVs are dependent on (i) the complexity of SVs and (ii) the power of the reference panel. In my experience, imputing the copy number of the GC CNV, particularly to impute rare alleles (CNs 5-6), was not feasible using the 266 WGS samples as the reference panel (results not shown). Handsaker et al (2015) similarly noted challenges in accurately imputing recurrently mutated structural alleles, based on the reference panel of ~2,000 samples. On the other hand, simple biallelic duplications could be phased and imputed well, and hence used for eQTL mapping (Chapter 3).

Based on this experience, I propose that the two SV call sets from this thesis can be phased and imputed and further utilized in future bovine genetics studies. The first set of SVs consists of ~4,000 accurately genotyped SVs (the stringent call set in Chapter 3). These SVs are simple biallelic variants, and can be merged with small genetic variants and phased simultaneous, as done in the 1000 Genomes Project (Sudmant et al. 2015). The second set of SVs consists of ~300 CNVs that are added in the EuroGenomics genotyping array (Chapter 3). Given that this array will be used to genotype a large number of animals managed by breeding programmes, massive SNP and CNV genotyping results can be simultaneously phased and imputed, and further utilized in downstream analyses.

6.6. Concluding remarks

This thesis highlighted bovine SVs of a broad spectrum: from the emergence of dnSV to dissection of a likely causative SV of economically important traits. This work clearly demonstrated that SVs have an impact at molecular and phenotypic levels, with high relevance to livestock breeding. Suggestions for discovering and utilizing functional SVs presented in this discussion may serve as a benchmark for future studies. Further efforts to characterize population-wide SVs curated in the current thesis, and novel SVs that will be discovered using advanced technologies, will contribute to a complete understanding of bovine genetic variation.

6




REFERENCES

References

Abdel-Shafy H, Bortfeldt RH, Reissmann M, Brockmann GA. 2014. Short communication: Validation of somatic cell score-associated loci identified in a genome-wide association study in German Holstein cattle. *J Dairy Sci* **97**: 2481–2486. http://dx.doi.org/10.3168/jds.2013-7149.

Abel HJ, Larson DE, Regier AA, Chiang C, Das I, Kanchi KL, Layer RM, Neale BM, Salerno WJ, Reeves C, et al. 2020. Mapping and characterization of structural variation in 17,795 human genomes. *Nature* **583**: 83–89.

Abo-Ismail MK, Brito LF, Miller SP, Sargolzaei M, Grossi DA, Moore SS, Plastow G, Stothard P, Nayeri S, Schenkel FS. 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**: 1–29.

Abyzov A, Urban AE, Snyder M, Gerstein M. 2011. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* **21**: 974–984.

Alkan C, Coe BP, Eichler EE. 2011. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**: 363–376.

Allais-Bonnet A, Hintermann A, Deloche M, Cornette R, Bardou P, Naval-sanchez M, Pinton A, Haruda A, Zakany J, Bigi D, et al. 2021. Analysis of Polycerate Mutants Reveals the Evolutionary Co-option of HOXD1 for Horn Patterning in Bovidae. *Mol Biol Evol* **38**: 2260–2272.

Allen HL, Estrada K, Lettres G, Berndt SI, Michael N W, Rivadeneira F, CJ W, AU J, S V, S R, et al. 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**: 832–838.

Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**: 648–660.

ArrayGen. Custom array design process. https://www.arraygen.com/single-nucleotide-polymorphism.php (Accessed January 22, 2020).

Audano PA, Sulovari A, Graves-lindsay TA, Li YI, Wilson RK, Eichler EE, Audano PA, Sulovari A, Graves-lindsay TA, Cantsilieris S, et al. 2019. Characterizing the Major Structural Variant Alleles of the Human Genome Resource. *Cell* **176**: 663–675.

Autier P, Boniol M, Pizot C, Mullie P. 2014. Vitamin D status and ill health : a systematic review. *Lancet Diabetes Endocrinol* **2**: 76–89.

Auwera GA Van der, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline*.

Axelsson E, Ratnakumar A, Arendt ML, Maqbool K, Webster MT, Perloski M, Liberg O, Arnemo JM, Hedhammar Å, Lindblad-Toh K. 2013. The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**: 360–364.

Bae JS, Cheong HS, Kim LH, NamGung S, Park TJ, Chun JY, Kim JY, Pasaje CFA, Lee JS, Shin HD. 2010. Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics* **11**.

Barre L, Fournel-gigleux S, Finel M, Netter P, Magdalou J, Ouzzine M. 2007. Substrate specificity of the human UDP-glucuronosyltransferase UGT2B4 and UGT2B7. *FEBS J* **274**: 1256–1264.

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**: 263–265.

Belyeu JR, Brand H, Wang H, Zhao X, Pedersen BS, Feusier J, Gupta M, Nicholas TJ, Baird L, Devlin B, et al. 2021. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *Am J Hum Genet* **108**: 1–11. https://doi.org/10.1016/j.ajhg.2021.02.012.

Ben Sassi N, González-Recio Ó, de Paz-del Río R, Rodríguez-Ramilo ST, Fernández AI. 2016. Associated effects of copy number variants on economically important traits in Spanish Holstein dairy cattle. *J Dairy Sci* **99**: 6371–6380. http://linkinghub.elsevier.com/retrieve/pii/S0022030216302740.

Bergero R, Ellis P, Haerty W, Larcombe L, Macaulay I, Mehta T, Mogensen M, Murray D, Nash W, Neale MJ, et al. 2021. Meiosis and beyond – understanding the mechanistic and evolutionary processes shaping the germline genome. *Biol Rev* **96**: 822–841.

Berry D, Buckley F, Dillon P, Evans R, Rath M, Veerkamp R. 2003. Genetic relationships among body condition score, body weight, milk yield, and fertility in dairy cows. *J Dairy Sci* **86**: 2193–2204.

Bertolotti AC, Layer RM, Gundappa MK, Gallagher MD, Pehlivanoglu E, Nome T, Robledo D, Kent MP, Røsæg LL, Holen MM, et al. 2020. The structural variation landscape in 492 Atlantic salmon genomes. *Nat Commun* **11**.

Bertrand AR, Flori L, Druet T. 2019. RZooRoH : An R package to characterize individual genomic autozygosity and identify homozygous-by-descent segments. *Methods Ecol Evol* **2019**: 860–866.

Beyter D, Ingimundardottir H, Oddsson A, Eggertsson HP, Bjornsson E, Jonsson H, Atlason BA, Kristmundsdottir S,

Mehringer S, Hardarson MT, et al. 2021. Long-read sequencing of 3,622 Icelanders provides insight into the role of structural variants in human diseases and other traits. *Nat Genet* **53**: 779–786. http://dx.doi.org/10.1038/s41588-021-00865-4.

Bickhart DM, Liu GE. 2014. The challenges and importance of structural variation detection in livestock. *Front Genet* **5**: 1–14.

Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG, Song J, Garcia JF, Sonstegard TS, Van Tassell CP, et al. 2016. Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. *DNA Res* **23**: 253–262.

Bikle DD, Schwartz J. 2019. Vitamin D binding protein, total and free Vitamin D levels in different physiological and pathophysiological conditions. *Front Endocrinol (Lausanne)* **10**: 1–12.

Binsbergen R Van, Bink MCAM, Calus MPL, Eeuwijk FA Van, Hayes BJ, Hulsegge I, Veerkamp RF. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet Sel Evol* **46**: 1–13.

Bloemhof S, Jong G De, Haas Y De. 2009. Genetic parameters for clinical mastitis in the first three lactations of Dutch Holstein cattle. *Vet Microbiol* **134**: 165–171.

Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, Blaszczyk K, Saeed S, Hamilton-Shield J, Clayton-Smith J, O'Rahilly S, et al. 2010. Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* **463**: 666–670.

Boettger LM, Handsaker RE, Zody MC, Mccarroll SA. 2012. Structural haplotypes and recent evolution of the human 17q21.31 region. *Nat Genet* **44**: 881–885. http://dx.doi.org/10.1038/ng.2334.

Boichard D, Boussaha M, Capitan A, Rocha D, Sanchez MP, Tribout T, Letaief R, Croiseau P, Grohs C, Li W, et al. 2018. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. In *11th World Congress on Genetics Applied to Livestock Production*, pp. 1–6.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bortoluzzi C, Megens H, Bosse M, Derks MFL, Dibbits B, Laport K, Weigend S, Groenen MAM, Crooijmans RPMA. 2020. Parallel Genetic Origin of Foot Feathering in Birds. *Mol Biol Evol* **37**: 2465–2476.

Bouillon R, Schuit F, Antonio L, Rastinejad F. 2020. Vitamin D Binding Protein: A Historic Overview. *Front Endocrinol (Lausanne)* **10**: 1–21.

Boussaha M, Esquerré D, Barbieri J, Djari A, Pinton A, Letaief R, Salin G, Escudié F, Roulet A. 2015. Genome-Wide Study of Structural Variants in Bovine Holstein , Montbéliarde and Normande Dairy Breeds. *PLoS One* **10**: 1–21.

Bouwman AC, Daetwyler HD, Chamberlain AJ, Ponce CH, Sargolzaei M, Schenkel FS, Sahana G, Govignon-gion A, Boitard S, Dolezal M, et al. 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.

Brand B, Scheinhardt MO, Friedrich J, Zimmer D, Reinsch N, Ponsuksili S, Schwerin M, Ziegler A. 2016. Adrenal cortex expression quantitative trait loci in a German Holstein ⊠ Charolais cross. *BMC Genet* **17**: 1–11.

Brandler WM, Antaki D, Gujral M, Kleiber ML, Whitney J, Maile MS, Hong O, Chapman TR, Tan S, Tandon P, et al. 2018. Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360**: 327–331.

Brandler WM, Antaki D, Gujral M, Noor A, Rosanio G, Chapman TR, Barrera DJ, Lin GN, Malhotra D, Watts AC, et al. 2016. Frequency and Complexity of de Novo Structural Mutation in Autism. *Am J Hum Genet* **98**: 667–679. http://dx.doi.org/10.1016/j.ajhg.2016.02.018.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Britt JH, Cushman RA, Dechow CD, Dobson H, Humblot P, Hutjens MF, Jones GA, Mitloehner FM, Ruegg PL, Sheldon IM, et al. 2021. Review: Perspective on high-performing dairy cows and herds. *Animal* 100298. https://doi.org/10.1016/j.animal.2021.100298.

Broad Institute, Handsaker RE. 2019. MPG Primer: Structural Variation (2019).

Brown WE, Garcia M, Mamedova LK, Christman KR, Zenobi MG, Staples CR, Leno BM, Overton TR, Whitlock BK, Daniel JA, et al. 2021. Acute-phase protein α-1-acid glycoprotein is negatively associated with feed intake in postpartum dairy cows. *J Dairy Sci* **104**: 806–817. http://dx.doi.org/10.3168/jds.2020-19025.

Browning BL, Zhou Y, Browning SR. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet* **103**: 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015.

Buitenhuis B, Poulsen NA, Gebreyesus G, Larsen LB. 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**: 1–12.

Butty AM, Chud TCS, Cardoso DF, Lopes LSF, Miglior F, Schenkel FS, Cánovas A, Häfliger IM, Drögemüller C, Stothard P, et al. 2021. Genome-wide association study between copy number variants and hoof health traits in Holstein dairy cattle. *J Dairy Sci* **104**: 8050–8061.

Byrska-Bischop M, Evani US, Zhao X, Basile AO. 2021. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios . *BioRxiv*.

Cai Z, Dusza M, Guldbrandtsen B, Lund MS, Sahana G. 2020. Distinguishing pleiotropy from linked QTL between milk production traits and mastitis resistance in Nordic Holstein cattle. *Genet Sel Evol* **52**: 19. https://gsejournal.biomedcentral.com/articles/10.1186/s12711-020-00538-6.

Cai Z, Guldbrandtsen B, Lund MS, Sahana G. 2018. Prioritizing candidate genes post-GWAS using multiple sources of data for mastitis resistance in dairy cattle. *BMC Genomics* **19**: 1–11.

Cameron DL, Di Stefano L, Papenfuss AT. 2019. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. *Nat Commun* **10**: 1–11. http://dx.doi.org/10.1038/s41467-019-11146-4.

Campbell CD, Eichler EE. 2013. Properties and rates of germline mutations in humans. *Trends Genet* **29**: 575–584. http://dx.doi.org/10.1016/j.tig.2013.04.005.

Cao Y, Chen G, Wu G, Zhang X, McDermott J, Chen X, Xu C, Jiang Q, Chen Z, Zeng Y, et al. 2019. Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res* **29**: 40–52.

Champroux A, Gharagozloo P, Drevet JR, Kocer A. 2016. Mammalian sperm nuclear organization : resiliencies and vulnerabilities. *Basic Clin Androl* 1–22. http://dx.doi.org/10.1186/s12610-016-0044-5.

Charlier C, Agerholm JS, Coppieters W, Karlskov-mortensen P, Li W, Jong G De, Fasquelle C, Karim L, Cirera S, Cambisano N, et al. 2012. A Deletion in the Bovine FANCI Gene Compromises Fertility by Causing Fetal Death and Brachyspina. *plos* **7**: 2–8.

Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, Cambisano N, Carta E, Dardano S, Dive M, Fasquelle C, et al. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. **40**: 449–454.

Chen L, Abel HJ, Das I, Larson DE, Ganel L, Kanchi KL, Regier AA, Young EP, Kang CJ, Scott AJ, et al. 2021. Association of structural variation with cardiometabolic traits in Finns. *Am J Hum Genet* **108**: 583–596. https://doi.org/10.1016/j.ajhg.2021.03.008.

Chen S, Krusche P, Dolzhenko E, Sherman RM, Petrovski R, Schlesinger F, Kirsche M, Bentley DR, Schatz MC, Sedlazeck FJ, et al. 2019. Paragraph: A graph-based structural variant genotyper for short-read sequence data. *GenomeBiology* **20**: 1–13.

Chiang C, Scott AJ, Davis JR, Tsang EK, Li X, Kim Y, Hadzic T, Damani FN, Ganel L, Montgomery SB, et al. 2017. The impact of structural variation on human gene expression. *Nat Genet* **49**: 692–699.

Chun RF, Peercy BE, Orwoll ES, Nielson CM, Adams JS, Hewison M. 2014. Vitamin D and DBP: The free hormone hypothesis revisited. *J Steroid Biochem Mol Biol* **144**: 132–137. http://dx.doi.org/10.1016/j.jsbmb.2013.09.012.

Clop A, Vidal O, Amills M. 2012. Copy number variation in the genomes of domestic animals. *Anim Genet* **43**: 503–517.

Coe BP, Witherspoon K, Rosenfeld JA, van Bon BWM, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LELM, et al. 2014. Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat Genet* **46**: 1063–1071. http://www.nature.com/doifinder/10.1038/ng.3092.

Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Francioli LC, Khera A V., Lowther C, Gauthier LD, Wang H, et al. 2020. A structural variation reference for medical and population genetics. *Nature* **581**: 444–451.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nat Rev Genet* **9**: 938–950.

Conrad DF, Hurles ME. 2007. The population genetics of structural variation. *Nat Genet* **39**: s30–s36.

Conrad DF, Keebler JEM, Depristo MA, Lindsay SJ, Zhang Y, Casals F, Idaghdour Y, Hartl CL, Torroja C, Garimella K V., et al. 2011. Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**: 712–714.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**: 704–712.

Cooper GM, Nickerson DA, Eichler EE. 2007. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* **39**: S22–S29.

Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. 2008. Systematic assessment of copy number variant detection. *Nat Genet* **40**: 1199–1203.

Coorens THH, Moore L, Robinson PS, Sanghvi R, Christopher J, Hewinson J, Przybilla MJ, Lawson ARJ, Spencer Chapman M, Cagan A, et al. 2021. Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**: 387–392. http://dx.doi.org/10.1038/s41586-021-03790-y.

Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, et al. 2010. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**: 713–720. http://dx.doi.org/10.1038/nature08979.

Crow JF. 1997. The high spontaneous mutation rate: Is it a health risk? *Proc Natl Acad Sci U S A* **94**: 8380–8386.

CRV. 2020. Breeding value Udder Health (Manual Quality, Chapter E-27). https://cooperatiecrv-be6.kxcdn.com/wp-content/uploads/2020/04/E_27-Uiergezondheid-April-2020-Engels.pdf (Accessed August 25, 2020).

Daiger SP, Mel S, Cavalli-Sforza LL. 1975. Group-Specific Component (Gc) Proteins Bind Vitamin D and 25-Hydroxyvitamin D. *Proc Nati Acad Sci* **72**: 2076–2080.

de Souza MM, Zerlotini A, Geistlinger L, Tizioto PC, Taylor JF, Rocha MIP, Diniz WJS, Coutinho LL, Regitano LCA. 2018. A comprehensive manually-curated compendium of bovine transcription factors. *Sci Rep* **8**: 1–12. http://dx.doi.org/10.1038/s41598-018-32146-2.

Dechow CD, Rogers GW, Sander-Nielsen U, Klei L, Lawlor TJ, Clay JS, Freeman AE, Abdel-Azim G, Kuck A, Schnell S. 2004. Correlations among body condition scores from various sources, dairy form, and cow health from the United States and Denmark. *J Dairy Sci* **87**: 3526–3533. http://dx.doi.org/10.3168/jds.S0022-0302(04)73489-X.

Delaneau O, Marchini J, Zagury J. 2012. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**: 179–181.

Demars J, Cano M, Drouilhet L, Plisson-Petit F, Bardou P, Fabre S, Servin B, Sarry J, Woloszyn F, Mulsant P, et al. 2017. Genome-Wide Identification of the Mutation Underlying Fleece Variation and Discriminating Ancestral Hairy Species from Modern Woolly Sheep. *Mol Biol Evol* **34**: 1722–1729.

DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–8. http://dx.doi.org/10.1038/ng.806.

Derks MFL, Lopes MS, Bosse M, Madsen O, Dibbits B, Harlizius B, Groenen MAM, Megens HJ. 2018. Balancing selection on a recessive lethal deletion with pleiotropic effects on two neighboring genes in the porcine genome. *PLoS Genet* **14**: 1–20. http://dx.doi.org/10.1371/journal.pgen.1007661.

Derks MFL, Steensma M. 2021. Review : Balancing Selection for Deleterious Alleles in Livestock. *Front Genet* **12**: 1–13.

Destito MCS, Souza MM, Cirillo CA, Ledda M, Zamboni A, Martin N, Morya E, Sameshima K, Beckmann JS, Coutre J, et al. 2014. GWAS of human bitter taste perception identifies new loci and reveals additional complexity of bitter taste genetics. *Hum Mol Genet* **23**: 259–267.

Dickinson ME, Flenniken AM, Ji X, Teboul L, Wong MD, White JK, Meehan TF, Weninger WJ, Westerberg H, Adissu H, et al. 2016. High-throughput discovery of novel developmental phenotypes. *Nature* **537**: 508–514.

Druet T, Gautier M. 2017. A model-based approach to characterize individual inbreeding at both global and local genomic scales. *Mol Ecol* **26**: 5820–5841.

Druet T, Georges M. 2015. LINKPHASE3: An improved pedigree-based phasing algorithm robust to genotyping and map errors. *Bioinformatics* **31**: 1677–1679.

Durán Aguilar M, Román Ponce SI, Ruiz López FJ, González Padilla E, Vásquez Peláez CG, Bagnato A, Strillacci MG. 2017. Genome-wide association study for milk somatic cell score in holstein cattle using copy number variation as markers. *J Anim Breed Genet* **134**: 49–59.

Duranthon V, Chavatte-Palmer P. 2018. Long term effects of ART: What do animals tell us? *Mol Reprod Dev* **85**: 348–368.

Durkin K, Coppieters W, Drögüller C, Ahariz N, Cambisano N, Druet T, Fasquelle C, Haile A, Horin P, Huang L, et al. 2012. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* **482**: 81–84.

Ebert P, Audano PA, Zhu Q, Rodriguez-martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Mari RS, et al. 2021. Haplotype-resolved diverse human genomes andintegrated analysis of structural variation. *Science* **372**.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* **11**: 446–450. http://dx.doi.org/10.1038/nrg2809.

Elferink MG, Vallée AAA, Jungerius AP, Crooijmans RPMA, Groenen MAM. 2008. Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. *BMC Genomics* **9**: 1–9.

Erhardt S, Su I, Schneider R, Barton S, Bannister AJ, Perez-burgos L. 2003. Consequences of the depletion of zygotic and embryonic enhancer of zeste 2 during preimplantation mouse development. *Development* **130**: 4235–4248.

Eslami Rasekh M, Hernández Y, Drinan SD, Fuxman Bass JI, Benson G. 2021. Genome-wide characterization of human minisatellite VNTRs: Population-specific alleles and gene expression differences. *Nucleic Acids Res* **49**: 4308–4324.

Ewing AD, Ballinger TJ, Earl D, Harris CC, Ding L, Wilson RK, Haussler D. 2013. Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol* **14**: 1–14.

Fagerberg L, Hallstrom BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, Habuka M, Tahmasebpoor S, Danielsson A, Edlund K, et al. 2014. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* **13**: 397–406.

Fang L, Cai W, Liu S, Canela-Xandri O, Gao Y, Jiang J, Rawlik K, Li B, Schroeder SG, Rosen BD, et al. 2020. Com-

prehensive analyses of 723 transcriptomes enhance genetic and biological interpretations for complex traits in cattle. *Genome Res* **30**: 790–801.

Fang L, Sahana G, Su G, Yu Y, Zhang S. 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**: 1–16.

Faust GG, Hall IM. 2014. SAMBLASTER : fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**: 2503–2505.

Feng X, Jiang J, Padhi A, Ning C, Fu J, Wang A, Mrode R, Liu JF. 2017. Characterization of genome-wide segmental duplications reveals a common genomic feature of association with immunity among domestic animals. *BMC Genomics* **18**: 1–11.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet* **7**: 85–97.

Feusier J, Watkins WS, Thomas J, Farrell A, Witherspoon DJ, Baird L, Ha H, Xing J, Jorde LB. 2019. Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res* **29**: 1567–1577.

Ford DE, Jones KW, Polani PE, de Almeida JC, Briggs JH. 1959. A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet* **1**: 711–713.

Freebern E, Santos DJA, Fang L, Jiang J, Parker Gaddis KL, Liu GE, Vanraden PM, Maltecca C, Cole JB, Ma L. 2020. GWAS and fine-mapping of livability and six disease traits in Holstein cattle. *BMC Genomics* **21**: 1–11.

Frischknecht M, Flury C, Leeb T, Rieder S, Neuditschko M. 2016. Selection signatures in Shetland ponies. *Anim Genet* **47**: 370–372.

Gallagher MD, Chen-plotkin AS. 2018. The Post-GWAS Era : From Association to Function. *Am J Hum Genet* **102**: 717–730. https://doi.org/10.1016/j.ajhg.2018.04.002.

Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, Amster G, Przeworski M. 2019. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc Nati Acad Sci* **116**: 9491–9500.

García-Rodríguez A, Gosálvez J, Agarwal A, Roy R, Johnston S. 2019. DNA damage and repair in human reproductive cells. *Int J Mol Sci* **20**: 1–22.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. 1–9. http://arxiv.org/abs/1207.3907.

Gautier M, Klassmann A, Vitalis R. 2017. rehh 2.0: a reimplementation of the R package rehh to detect positive selection from haplotype structure. *Mol Ecol Resour* **17**: 78–90.

Gel B, Dıez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. 2016. regioneR : an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **32**: 289–291.

Georges M. 2007. Mapping , Fine Mapping , and Molecular Dissection of Quantitative Trait Loci in Domestic Animals. *Annu Rev Genomics Hum Genet* **8**: 131–62.

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

Giuffra E, Evans G, Törnsten A, Wales R, Day A, Looft H, Plastow G, Andersson L. 1999. The Belt mutation in pigs is an allele at the Dominant white (I/KIT) locus. *Mamm Genome* **10**: 1132–1136.

Giuffra E, Tornsten A, Marklund S, Bongcam-rudlo E, Chardon P, Kijas MHJ, Anderson SI, Archibald AL, Andersson L. 2002. A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. *Mamm Genome* **13**: 569–577.

Giuffra E, Tuggle CK. 2019. Functional Annotation of Animal Genomes (FAANG): Current Achievements and Roadmap. *Annu Rev Anim Biosci* **7**: 65–88.

Goddard ME. 2011. Genetic Architecture of Complex Traits. *Proc Assoc Adv Anim Breed Genet* **19**: 1–6. http://www.aaabg.org/proceedings/2011/goddard001.pdf.

Goldmann JM, Seplyarskiy VB, Wong WSW, Vilboux T, Neerincx PB, Bodian DL, Solomon BD, Veltman JA, Deeken JF. 2018. Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat Genet* **50**: 487–492. http://dx.doi.org/10.1038/s41588-018-0071-6.

Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LELM, Hoischen A, Roach JC, et al. 2016. Parent-of-origin-specific signatures of de novo mutations. *Nat Genet* **48**: 935–939.

Gomme PT, Bertolini J. 2004. Therapeutic potential of vitamin D-binding protein. *TRENDS Biotechnol Biotechnol* **22**.

Gondo Y, Gardner JM, Nakatsu Y, Durham-pierre D, Deveaut SA, Kuper C, Brilliant MH. 1993. High-frequency genetic reversion mediated by a DNA duplication : The mouse pink-eyed unstable mutation. *Proc Nati Acad Sci* **90**: 297–301.

Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, et al. 2005. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* **307**: 1434–1440.

Gopalakrishnan S, Sullivan BA, Trazzi S, Della Valle G, Robertson KD. 2009. DNMT3B interacts with constitutive

centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions. *Hum Mol Genet* **18**: 3178–3193.

Grégoire MC, Leduc F, Morin MH, Cavé T, Arguin M, Richter M, Jacques PÉ, Boissonneault G. 2018. The DNA double-strand "breakome" of mouse spermatids. *Cell Mol Life Sci* **75**: 2859–2872. https://doi.org/10.1007/s00018-018-2769-0.

Grobet L, Jose L, Martin R, Poncelet D, Pirottin D, Brouwers B, Riquet J, Schoeberlein A, Dunner S, Massabanda J, et al. 1997. A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nat Genet* **17**: 71–74.

Guryev V, Saar K, Adamovic T, Verheul M, Van Heesch SAAC, Cook S, Pravenec M, Aitman T, Jacob H, Shull JD, et al. 2008. Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet* **40**: 538–545.

Halasa T, Huijps K, Østerås O, Hogeveen H. 2007. Economic effects of bovine mastitis and mastitis management: A review. *Vet Q* **29**: 18–31.

Haldane JBS. 1947. The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann Eugen* **13**: 262–271.

Haldane JBS. 1935. The rate of spontaneous mutation of a human gene. *J Genet* 317–326.

Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, Mccarroll SA. 2015. Large multiallelic copy number variations in humans. *Nat Genet* **47**: 296–303. http://dx.doi.org/10.1038/ng.3200.

Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mni M, Mullaart E, Coppieters W, Georges M. 2017. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. *BioRxiv* 1–27.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. **10**.

Heaton MP, Smith TPL, Bickhart DM, Vander Ley BL, Kuehn LA, Oppenheimer J, Shafer WR, Schuetze FT, Stroud B, Mcclure JC, et al. 2021. A Reference Genome Assembly of Simmental Cattle, Bos taurus taurus. *J Hered* **112**: 184–191.

Hedrick PW. 2015. Heterozygote Advantage : The Effect of Artificial Selection in Livestock and Pets. *J Hered* **106**: 141–154.

Hehir-Kwa JY, Marschall T, Kloosterman WP, Francioli LC, Baaijens JA, Dijkstra LJ, Abdellaoui A, Koval V, Thung DT, Wardenaar R, et al. 2016. A high-quality human reference panel reveals the complexity and distribution of genomic structural variants. *Nat Commun* **7**: 1–10.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38**: 576–589. http://dx.doi.org/10.1016/j.molcel.2010.05.004.

Hekselman I, Yeger-Lotem E. 2020. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nat Rev Genet*. http://www.ncbi.nlm.nih.gov/pubmed/31913361.

Henkel J, Saif R, Jagannathan V, Schmocker C, Zeindler F, Bangerter E, Herren U, Posantzis D, Bulut Z, Ammann P, et al. 2019. Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. *PLoS Genet* **15**: 1–18.

Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. 2006. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat Genet* **38**: 82–85.

Ho SS, Urban AE, Mills RE. 2019. Structural variation in the sequencing era. *Nat Rev Genet*. http://www.ncbi.nlm.nih.gov/pubmed/31729472.

Horst RL, Reinhardt TA, Reddy GS. 2005. Vitamin D Metabolism. In *Vitamin D* (eds. D. Feldman, J.W. Pike, and J.S. Adams), pp. 15–36.

Hou Y, Bickhart DM, Hvinden ML, Li C, Song J, Boichard DA, Fritz S, Eggen A, DeNise S, Wiggans GR, et al. 2012. Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array. *BMC Genomics* **13**: 376.

Hu Z-L, Park C, Reecy J. 2016. Developmental progress and current status of the Animal QTLdb. *Nucleic Acids Res* **44**: 827–833.

Hu ZL, Park CA, Reecy JM. 2019. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. *Nucleic Acids Res* **47**: D701–D710.

Huddleston J, Eichler EE. 2016. An incomplete understanding of human genetic variation. *Genetics* **202**: 1251–1254.

Hujoel MLA, Sherman MA, Barton AR, Mukamel RE, Vijay GS, Loh P. Influences of rare copy number variation on human complex traits. BioRxiv. 2021;1–25.

Hunt SE, Mclaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, et al. 2018. Ensembl variation resources. 1–12.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36**: 949–951. http://www.ncbi.nlm.nih.gov/pubmed/15286789.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: Classifying and distinguishing between models. *Nat*

Rev Genet **11**: 97–108.

Irani M, Merhi Z, D M. 2014. Role of vitamin D in ovarian physiology and its implication in reproduction : a systematic review. *Fertil Steril* **102**: 460-468.e3. http://dx.doi.org/10.1016/j.fertnstert.2014.04.046.

Itsara A, Wu H, Smith JD, Nickerson DA, Romieu I, London SJ, Eichler EE. 2010. De novo rates and selection of large copy number variation. *Genome Res* **20**: 1469–1481.

Jacobs PA, Strong JA. 1959. A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature* **183**: 302–303.

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, Vanliere JM, Fung H, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, et al. 2008. Genotype , haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998–1003.

Jamrozik J, Koeck A, Miglior F, Kistemaker GJ, Schenkel FS, Kelton DF, Doormaal BJ Van. 2013. Genetic and Genomic Evaluation of Mastitis Resistance in Canada. In *Interbull Bulletin*, pp. 43–51.

Jiang J, Cole JB, Freebern E, Da Y, VanRaden PM, Ma L. 2019a. Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls. *Commun Biol* **2**: 1–12.

Jiang J, Ma L, Prakapenka D, VanRaden PM, Cole JB, Da Y. 2019b. A large-scale genome-wide association study in U.S. Holstein cattle. *Front Genet* **10**.

Jiang L, Jiang J, Yang J, Liu X, Wang J, Wang H, Ding X, Liu J, Zhang Q. 2013. Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. *BMC Genomics* **14**.

Jolliffe DA, Walton RT, Grif CJ, Martineau AR. 2016. Single nucleotide polymorphisms in the vitamin D pathway associating with circulating concentrations of vitamin D metabolites and non-skeletal health outcomes : Review of genetic association studies. *J Steroid Biochem Mol Biol* **164**: 18–29.

Jonsson H, Magnusdottir E, Eggertsson HP, Stefansson OA, Arnadottir GA, Eiriksson O, Zink F, Helgason EA, Jonsdottir I, Gylfason A, et al. 2021. Differences between germline genomes of monozygotic twins. *Nat Genet* **53**. http://dx.doi.org/10.1038/s41588-020-00755-1.

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Gudjonsson SA, Ward LD, Hardarson MT, Hjorleifsson KE, et al. 2017. Parental influenec on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**: 519–522. http://dx.doi.org/10.1038/nature24018.

Kadri N, Charlier C, Cambisano N, Deckers M, Mullaart E. 2018. High resolution mapping of cross-over events in cattle using NGS data. In *Proceedings of the World Congress on Genetics Applied to Livestock Production*, p. 11.808.

Kadri NK, Harland C, Faux P, Cambisano N, Karim L, Coppieters W, Fritz S, Mullaart E, Baurain D, Boichard D, et al. 2016. Coding and noncoding variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B affect recombination rate in cattle. *Genome Res* 1323–1332.

Kadri NK, Sahana G, Charlier C, Iso-Touru T, Guldbrandtsen B, Karim L, Nielsen US, Panitz F, Aamand GP, Schulman N, et al. 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**.

Kalitsis P, Fowler KJ, Earle E, Hill J, Choo KHA. 1998. Targeted disruption of mouse centromere protein C gene leads to mitotic disarray and early embryo death. *Proc Natl Acad Sci U S A* **95**: 1136–1141.

Kato M, Kawaguchi T, Ishikawa S, Umeda T, Nakamichi R, Shapero MH, Jones KW, Nakamura Y, Aburatani H, Tsunoda T. 2010. Population-genetic nature of copy number variations in the human genome. *Hum Mol Genet* **19**: 761–773.

Kemper KE, Littlejohn MD, Lopdell T, Hayes BJ, Bennett LE, Williams RP, Xu XQ, Visscher PM, Carrick MJ, Goddard ME. 2016. Leveraging genetically simple traits to identify small-effect variants for complex phenotypes. *BMC Genomics* **17**: 1–9.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler a. D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006.

Kim-Hellmuth S, Aguet F, Oliva M, Muñoz-Aguirre M, Wucher V, Kasela S, Castel S, Hamel A, Viñuela A, Roberts A, et al. 2019. Cell type specific genetic regulation of gene expression across human tissues. *Science* **1332**.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**. http://dx.doi.org/10.1038/s41587-019-0201-4.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* **217**: 624–626.

Kloosterman WP, Francioli LC, Hormozdiari F, Marschall T, Hehir-Kwa JY, Abdellaoui A, Lameijer EW, Moed MH, Koval V, Renkens I, et al. 2015. Characteristics of de novo structural changes in the human genome. *Genome Res* **25**: 792–801.

Koivula M, Mäntysaari EA, Negussie E, Serenius T. 2005. Genetic and phenotypic relationships among milk yield and somatic cell count before and after clinical mastitis. *J Dairy Sci* **88**: 827–833.

Koltai H, Weingarten-baror C. 2008. Specificity of DNA microarray hybridization : characterization , effectors and approaches for data correction. *Nucleic Acids Res* **36**: 2395–2405.

Kommadath A, Grant JR, Krivushin K, Butty AM, Baes CF, Carthy TR, Berry DP, Stothard P. 2019. A large interactive visual database of copy number variants discovered in taurine cattle. *Gigascience* **8**: 1–12.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**: 471–475. http://dx.doi.org/10.1038/nature11396.

Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol* **20**: 8–11.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos : An information aesthetic for comparative genomics. 1639–1645.

Lan X, Pritchard JK. 2016. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **352**: 1009–1013.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Larson DE, Abel HJ, Chiang C, Badve A, Das I, Eldred JM, Layer RM, Hall IM. 2019. Genome analysis svtools : population-scale analysis of structural variation. *Bioinformatics* **35**: 4782–4787.

Laugsch M, Bartusel M, Rehimi R, Alirzayeva H, Karaolidou A, Crispatzu G, Zentis P, Nikolic M, Bleckwehl T, Kolovos P, et al. 2019. Modeling the Pathological Long-Range Regulatory Effects of Human Structural Variation with Patient-Specific hiPSCs. *Cell Stem Cell* **24**: 736-752.e12.

Lauridsen AL, Vestergaard P, Hermann AP, Brot C, Heickendorff L, Mosekilde L, Nexo E. 2005. Plasma concentrations of 25-Hydroxy-Vitamin D and 1,25-Dihydroxy-Vitamin D are Related to the Phenotype of Gc (Vitamin D-Binding Protein): A Cross-sectional Study on 595 Early Postmenopausal Women. *Calcif Tissue Int* **25**: 15–22.

Layer RM, Chiang C, Quinlan AR, Hall IM. 2014. LUMPY : a probabilistic framework for structural variant discovery. *Genome Biol* **15**: 1–19.

Leal-Gutiérrez JD, Elzo MA, Mateescu RG. 2020. Identification of eQTLs and sQTLs associated with meat quality in beef. *BMC Genomics* **21**: 1–15.

Lee Y-L, Bosse M, Mullaart E, Groenen MAM, Veerkamp RF, Bouwman AC. 2020. Functional and population genetic features of copy number variations in two dairy cattle populations. *BMC Genomics* **21**: 1–15.

Lee Y-L, Takeda H, Moreira GCM, Karim L, Mullaart E, Coppieters W, Appeltant R, Veerkamp RF, Groenen MAM, Georges M, et al. 2021. A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle. *PLoS Genet* 1–27. http://dx.doi.org/10.1371/journal.pgen.1009331.

Lejeune J, Gautier M, R. TM. 1959. Etude des chromosomes somatiques de neuf enfants mongoliens. *C R Acad Sci* **248**: 1721–1722.

Lemoine S, Combes F, Crom S Le. 2009. An evaluation of custom microarray applications : the oligonucleotide design challenge. *Nucleic Acids Res* **37**: 1726–1739.

Leonard AS, Crysnanto D, Fang Z, Heaton MP, Brian, L Ley V, Herrera C, Bollwein H, Bickhart DM, Kuhn KL, Smith TP, et al. 2021. Bovine pangenome reveals trait- associated structural variation from diverse assembly inputs. *bioRxiv*.

Levin I, Smith EJ. 1990. Molecular analysis of endogenous virus ev21-slow feathering complex of chickens. 1. Cloning of proviral-cell junction fragment and unoccupied integration site. *Poult Sci* **69**: 2017–2026.

Li B, Fang L, Null DJ, Hutchison JL, Connor EE, VanRaden PM, VandeHaar MJ, Tempelman RJ, Weigel KA, Cole JB. 2019. High-density genome-wide association study for residual feed intake in Holstein dairy cattle. *J Dairy Sci* **102**: 11067–11080.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr arXiv* **00**: 3. http://arxiv.org/abs/1303.3997.

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li J, Lee M, Davis BW, Lamichhaney S, Dorshorst BJ, Siegel PB, Andersson L. 2020. Mutations Upstream of the TBX5 and PITX1 Transcription Factor Genes Are Associated with Feathered Legs in the Domestic Chicken. *Mol Biol Evol* **37**: 2477–2486.

Lippolis JD, Reinhardt TA, Sacco RA, Nonnecke BJ, Nelson CD. 2011. Treatment of an Intramammary Bacterial Infection with 25-Hydroxyvitamin D 3. *PLoS One* **6**: 1–7.

Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, Couldrey C, Keehan M, Sherlock RG, Harland C, et al. 2016. Sequence-based Association Analysis Reveals an MGST1 eQTL with Pleiotropic Effects on Bovine Milk Composition. *Sci Rep* 1–14.

Littlejohn MD, Tiplady K, Lopdell T, Law TA, Scott A, Harland C, Sherlock R, Henty K, Obolonkin V, Lehnert K, et

al. 2014. Expression variants of the lipogenic AGPAT6 gene affect diverse milk composition phenotypes in Bos taurus. *PLoS One* **9**: 1–12.

Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. 2019. Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet* **51**. http://dx.doi.org/10.1038/s41588-019-0404-0.

Liu X, Yu X, Zack DJ, Zhu H, Qian J. 2008. TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics* **9**: 1–7.

Lizio M, Abugessaisa I, Noguchi S, Kondo A, Hasegawa A, Hon CC, De Hoon M, Severin J, Oki S, Hayashizaki Y, et al. 2019. Update of the FANTOM web resource: Expansion to provide additional transcriptome atlases. *Nucleic Acids Res* **47**: D752–D758.

Locke DP, Sharp AJ, Mccarroll SA, Mcgrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM, et al. 2006. Linkage Disequilibrium and Heritability of Copy-Number Polymorphisms within Duplicated Regions of the Human Genome. *Am J Hum Genet* **79**: 275–290.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**: 1170–1187. http://dx.doi.org/10.1016/j.cell.2016.09.018.

Lopdell TJ, Tiplady K, Struchalin M, Johnson TJJ, Keehan M, Sherlock R, Couldrey C, Davis SR, Snell RG, Spelman RJ, et al. 2017. DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. *BMC Genomics* **18**: 1–18.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 1–21.

Low WY, Tearle R, Liu R, Koren S, Rhie A, Bickhart DM, Rosen BD, Kronenberg ZN, Kingan SB, Tseng E, et al. 2020. Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. *Nat Commun* **11**. http://dx.doi.org/10.1038/s41467-020-15848-y.

Lupski JR, Stankiewicz P. 2005. Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genet* **1**: 0627–0633.

Macé A, Tuke MA, Deelen P, Kristiansson K, Mattsson H, Nõukas M, Sapkota Y, Schick U, Porcu E, Rüeger S, et al. 2017. CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nat Commun* **8**: 1–11.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* **461**: 747–753. http://dx.doi.org/10.1038/nature08494.

Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, Antaki D, Consortium C and SWG of the PG. 2017. Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nat Genet* **49**: 27–35.

May. 1917. Selection for Higher and Lower Facet Numbers in the Bar-Eyed Race of Drosophila and the Appearance of Reverse Mutations. *Biol Bull* **33**: 361–395.

Mccarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, et al. 2006. Common deletion polymorphisms in the human genome. *Nat Commun* **38**: 86–92.

McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, De Bakker PIW, Maller JB, Kirby A, et al. 2008. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* **40**: 1166–1174.

McClintock B. 1931. Cytological Observations of Deficiencies Involving Known Genes , Translocations and an Inversion in Zea mays. *Agric Exp Stn* **3**.

McClintock B. 1950. The origin and behavior of mutable loci in Maize. *Proc Nati Acad Sci* **36**.

McGuckin MM, Giesy SL, Davis AN, Abyeta MA, Horst EA, Saed Samii S, Zang Y, Butler WR, Baumgard LH, McFadden JW, et al. 2020. The acute phase protein orosomucoid 1 is upregulated in early lactation but does not trigger appetite-suppressing STAT3 signaling via the leptin receptor. *J Dairy Sci* **103**: 4765–4776.

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

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *bioRxiv* 042374. http://biorxiv.org/content/early/2016/03/04/042374.abstract.

McPherron AC, Lee S-J. 1997. Double muscling in cattle due to mutations in the myostatin gene. *Proc Nati Acad Sci* **94**: 12457–12461.

Merriman KE, Powell JL, Santos JEP, Nelson CD. 2018. Intramammary 25-hydroxyvitamin D3 treatment modulates innate immune responses to endotoxin-induced mastitis. *J Dairy Sci* **101**: 7593–7607. http://dx.doi.org/10.3168/jds.2017-14143.

Mesbah-Uddin M, Guldbrandtsen B, Iso-Touru T, Vilkki J, De Koning D-J, Boichard D, Lund MS, Sahana G. 2017. Genome-wide mapping of large deletions and their population-genetic properties in dairy cattle. *DNA Res* **25**:

49–59.

Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al. 2012a. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**: 1431–1442.

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, et al. 2012b. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151**: 1431–1442.

Miglior F, Fleming A, Malchiodi F, Brito LF, Martin P, Baes CF. 2017. A 100-Year Review: Identification and genetic selection of economically important traits in dairy cattle. *J Dairy Sci* **100**: 10251–10271. http://dx.doi.org/10.3168/jds.2017-12968.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**: 59–65.

Mishra NA, Drögemüller C, Jagannathan V, Keller I, Wüthrich D, Bruggmann R, Beck J, Schütz E, Brenig B, Demmel S, et al. 2017. A structural variant in the 5'-flanking region of the TWIST2 gene affects melanocyte development in belted cattle. *PLoS One* **12**: 1–16.

Mukamel RE, Handsaker RE, Sherman MA, Barton AR, Zheng Y, McCarroll SA, Loh PR. 2021. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. *Science* **373**: 1499–1505.

Mulder HA, Lee SH, Clark S, Hayes BJ, Werf JHJ Van Der. 2019. The Impact of Genomic and Traditional Selection on the Contribution of Mutational Variance to Long-Term. *Genetics* **213**: 361–378.

Nakano K, Shiroma A, Shimoji M, Tamotsu H. 2017. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Hum Cell* **30**: 149–161.

Nandolo W, Utsunomiya YT, Mészáros G, Wurzinger M, Khayadzadeh N, Torrecilha RBP, Mulindwa HA, Gondwe TN, Waldmann P, Ferenčaković M, et al. 2018. Misidentification of runs of homozygosity islands in cattle caused by interference with copy number variation or large intermarker distances. *Genet Sel Evol* **50**: 1–13.

Nayeri S, Sargolzaei M, Abo-Ismail M, Miller S, Schenkel F, Moore S, Stothard P. 2017. Genome-wide association study for lactation persistency, female fertility, longevity, and lifetime profit index traits in Holstein dairy cattle. *J Dairy Sci* **100**: 1246–1258. http://dx.doi.org/10.3168/jds.2016-11770.

Negussie E, Lidauer M, Nielsen US, Aamand GP. 2010. Combining Test Day SCS with Clinical Mastitis and Udder Type Traits: A Random Regression Model for Joint Genetic Evaluation of Udder Health in Denmark, Finland and Sweden. In *Interbull Bulletin*, pp. 25–32.

Ngcungcu T, Oti M, Sitek JC, Haukanes BI, Linghu B, Bruccoleri R, Stokowy T, Oakeley EJ, Yang F, Zhu J, et al. 2017. Duplicated Enhancer Region Increases Expression of CTSB and Segregates with Keratolytic Winter Erythema in South African and Norwegian Families. *Am J Hum Genet* **100**: 737–750. http://dx.doi.org/10.1016/j.ajhg.2017.03.012.

Nguyen D, Webber C, Ponting CP. 2006. Bias of Selection on Human Copy-Number Variants. *PLoS Genet* **2**: 198–207.

Ohno S. 1970. Why Gene Duplication? , Duplication for the Sake of Producing More of the Same. In *Evolution by Gene Duplication* http://coleoguy.github.io/reading.group/Ohno - 10-12.pdf.

Olsen HG, Knutsen TM, Lewandowska-Sabat AM, Grove H, Nome T, Svendsen M, Arnyasi M, Sodeland M, Sundsaasen KK, Dahl SR, et al. 2016. Fine mapping of a QTL on bovine chromosome 6 using imputed full sequence data suggests a key role for the group-specific component (GC) gene in clinical mastitis and milk production. *Genet Sel Evol* **48**: 1–16.

Oppenheimer J, Rosen BD, Heaton MP, Vander Ley BL, Shafer WR, Schuetze FT, Stroud B, Kuehn LA, Mcclure JC, Barfield JP, et al. 2021. A Reference Genome Assembly of American Bison, Bison bison bison. *J Hered* **112**: 174–183.

Osterwalder M, Barozzi I, Tissières V, Fukuda-yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-frick I, Pickle CS, et al. 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*.

Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, Burke M, Füllgrabe A, Fuentes AMP, George N, et al. 2018. Expression Atlas: Gene and protein expression across multiple studies and organisms. *Nucleic Acids Res* **46**: D246–D251.

Pausch H, Emmerling R, Schwarzenbacher H, Fries R. 2016. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. *Genet Sel Evol* **48**: 1–9.

Pedersen BS, Quinlan AR. 2019. Duphold : scalable , depth-based annotation and curation of high-confidence structural variant calls. *Gigascience* 1–5.

Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**: 1256–1260.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2016. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotech* **33**: 290–295.

Pierce MD, Dzama K, Muchadeyi FC. 2018. Genetic Diversity of Seven Cattle Breeds Inferred Using Copy Number

Variations. *Front Genet* **9**: 1–14.

Poindexter MB, Kweh MF, Zimpel R, Zuniga J, Lopera C, Zenobi MG, Jiang Y, Engstrom M, Celi P, Santos JEP, et al. 2020. Feeding supplemental 25-hydroxyvitamin D 3 increases serum mineral concentrations and alters mammary immunity of lactating dairy cows. *J Dairy Sci* **103**: 805–822.

Prinsen RTMM, Rossoni A, Gredler B, Bieber A, Bagnato A, Strillacci MG. 2017. A genome wide association study between CNVs and quantitative traits in Brown Swiss cattle. *Livest Sci* **202**: 7–12. http://dx.doi.org/10.1016/j.livsci.2017.05.011.

Prinsen RTMM, Strillacci MG, Schiavini F, Santus E, Rossoni A, Maurer V, Bieber A, Gredler B, Dolezal M, Bagnato A. 2016. A genome-wide scan of copy number variants using high-density SNPs in Brown Swiss dairy cattle. *Livest Sci* **191**: 153–160. http://linkinghub.elsevier.com/retrieve/pii/S1871141316301779.

Pritchard T, Coffey M, Mrode R, Wall E. 2013. Genetic parameters for production, health, fertility and longevity traits in dairy cows. *Animal* **7**: 34–46.

Pryce JE, Coffey MP, Brotherstone S. 2000. The genetic relationship between calving interval, body condition score and linear type and management traits in registered Holsteins. *J Dairy Sci* **83**: 2664–2671. http://dx.doi.org/10.3168/jds.S0022-0302(00)75160-5.

Purcell S, Neale B, Todd-brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, Bakker PIW De, Daly MJ, et al. 2007. PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* **81**: 559–575.

Qanbari S, Pimentel ECG, Tetens J, Thaller G, Lichtner P, Sharifi AR, Simianer H. 2010. A genome-wide scan for signatures of recent selection in Holstein cattle. *Anim Genet* **41**: 377–389.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, Dominiczak A, Morris A, Porteous D, Smith B, et al. 2016. Timing, rates and spectra of human germline mutation. *Nat Genet* **48**: 126–133.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al. 2006. Global variation in copy number in the human genome. *Nature* **444**: 444–54.

Richardson SR, Gerdes P, Gerhardt DJ, Sanchez-Luque FJ, Bodea GO, Muñoz-Lopez M, Jesuadian JS, Kempen MJHC, Carreira PE, Jeddeloh JA, et al. 2017. Heritable L1 retrotransposition in the mouse primordial germline and early embryo. *Genome Res* **27**: 1395–1405.

Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, Rowan TN, Low WY, Zimin A, Couldrey C, et al. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. *Gigascience* **9**: 1–9.

Rubin C-J, Megens H-J, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg O, Jern P, Jorgensen CB, et al. 2012a. Strong signatures of selection in the domestic pig genome. *Proc Natl Acad Sci* **109**: 19529–19536.

Rubin C-JJ, Megens H-JJ, Barrio AM, Maqbool K, Sayyab S, Schwochow D, Wang C, Carlborg Ö, Jern P, Jørgensen CB, et al. 2012b. Strong signatures of selection in the domestic pig genome. *Proc Natl Acad Sci* **109**: 19529–19536. http://www.ncbi.nlm.nih.gov/pubmed/23151514.

Sahana G, Guldbrandtsen B, Thomsen B, Holm LE, Panitz F, Brøndum RF, Bendixen C, Lund MS. 2014. Genome-wide association study using high-density single nucleotide polymorphism arrays and whole-genome sequences for clinical mastitis traits in dairy cattle. *J Dairy Sci* **97**: 7258–7275. http://dx.doi.org/10.3168/jds.2014-8141.

Sahana G, Guldbrandtsen B, Thomsen B, Lund MS. 2013. Confirmation and fine-mapping of clinical mastitis and somatic cell score QTL in Nordic Holstein cattle. *Anim Genet* **44**: 620–626.

Santos JEP, Bisinotto RS, Ribeiro ES. 2016. Mechanisms underlying reduced fertility in anovular dairy cows. *Theriogenology* **86**: 254–262. http://dx.doi.org/10.1016/j.theriogenology.2016.04.038.

Sasaki S, Watanabe T, Nishimura S, Sugimoto Y. 2016. Genome-wide identification of copy number variation using high-density single-nucleotide polymorphism array in Japanese Black cattle. *BMC Genet* **17**: 1–9.

Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, Quinlan AR. 2019. Large , three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife* **8**: 1–24.

Scholes C, Biette KM, Harden TT, DePace AH. 2019. Signal Integration by Shadow Enhancers and Enhancer Duplications Varies across the Drosophila Embryo. *Cell Rep* **26**: 2407-2418.e5. https://doi.org/10.1016/j.celrep.2019.01.115.

Schrider DR, Hahn MW. 2010. Lower Linkage Disequilibrium at CNVs is due to Both Recurrent Mutation and Transposing Duplications. *Mol Biol Evol* **27**: 103–111.

Schütz E, Wehrhahn C, Wanjek M, Bortfeld R, Wemheuer WE, Beck J, Brenig B. 2016a. 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**: 1–15.

Schütz E, Wehrhahn C, Wanjek M, Bortfeld R, Wemheuer WE, Beck J, Brenig B. 2016b. 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**: 1–15.

Scott AJ, Chiang C, Hall IM. 2021. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Biol*.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**: 525–8. http://www.ncbi.nlm.nih.gov/pubmed/15273396.

Shabalin AA. 2012. Matrix eQTL : ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**: 1353–1358.

Sharma A, Lee JS, Dang CG, Sudrajad P, Kim HC, Yeon SH, Kang HS, Lee S-H. 2015. Stories and Challenges of Genome Wide Association Studies in Livestock — A Review. *Asian-Australas J Anima Sci* **28**: 1371–1379.

Sinotte M, Diorio C, Berube S, Pollak M, Brisson J. 2009. Genetic polymorphisms of the vitamin D binding protein and plasma concentrations of 25-hydroxyvitamin D in premenopausal women. *Am J Clin Nutr* **25**: 634–640.

Smit M, Segers K, Carrascosa LG, Shay T, Baraldi F, Gyapay G, Snowder G, Georges M, Cockett N, Charlier C. 2003. Mosaicism of solid gold supports the causality of a noncoding A-to-G transition in the determinism of the callipyge phenotype. *Genetics* **163**: 453–456.

Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ. 2018. Mouse Genome Database (MGD)-2018 : knowledgebase for the laboratory mouse. *Nucleic Acids Res* **46**: 836–842.

Sodeland M, Kent MP, Olsen HG, Opsal MA, Svendsen M, Sehested E, Hayes BJ, Lien S. 2011. Quantitative trait loci for clinical mastitis on chromosomes 2, 6, 14 and 20 in Norwegian Red cattle. *Anim Genet* **42**: 457–465.

Stegle O, Parts L, Piipari M, Winn J, Durbin R. 2012. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. **7**: 500–507.

Strillacci MG, Gorla E, Cozzi MC, Vevey M, Genova F, Scienski K, Longeri M, Bagnato A. 2018. A copy number variant scan in the autochthonous Valdostana Red Pied cattle breed and comparison with specialized dairy populations. *PLoS One* **13**: 1–18.

Sturtevant AH. 1913. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *J Exp Zool* **14**: 43–59.

Sturtevant AH, Morgan TH. 1923. Reverse Mutation of the Bar Gene Correlated with Crossing over. *Science* **57**: 746–747.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MHY, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**: 75–81.

Sun Y, Yang Y, Qin Z, Cai J, Guo X, Tang Y, Wan J, Su DF, Liu X. 2016. The acute-phase protein orosomucoid regulates food intake and energy homeostasis via leptin receptor signaling pathway. *Diabetes* **65**: 1630–1641.

Swamy N, Dutta A, Ray R. 1997. Roles of the structure and orientation of ligands and ligand mimics inside the ligand-binding pocket of the vitamin D-binding protein. *Biochemistry* **36**: 7432–7436.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba : fast processing of NGS alignment formats. *Bioinfo* **31**: 2032–2034.

The Bovine Hapmap Consortium. 2009. Genome-Wide Survery of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science* **324**: 528–532.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

The FAANG Consortium, Andersson L, Archibald AL, Bottema CD, Brauning R, Burgess SC, Burt DW, Casas E, Cheng HH, Clarke L, et al. 2015. Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol* **16**: 4–9.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437**: 1299–1320.

Thomas GWC, Wang RJ, Nguyen J, Harris RA, Raveendran M, Rogers J, Hahn MW. 2020. Origins and long-term patterns of copy-number variation in rhesus macaques. *Mol Biol Evol* 1–12.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Res* **13**: 2129–2141. http://www.genome.org/cgi/doi/10.1101/gr.772403.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief Bioinform* **14**: 178–192.

Tice SC. 1914. A new sex-linked character in Drosophila. *Biol Bull* **26**: 221–230.

Trask BJ. 2002. Human cytogenetics: 46 Chromosomes, 46 years and counting. *Nat Rev Genet* **3**: 769–778.

Tribout T, Croiseau P, Lefebvre R, Barbat A, Boussaha M, Fritz S, Boichard D, Hoze C, Sanchez MP. 2020. Confirmed effects of candidate variants for milk production, udder health, and udder morphology in dairy cattle. *Genet Sel Evol* **52**: 1–26. https://doi.org/10.1186/s12711-020-00575-1.

Tšuiko O, Catteeuw M, Esteki MZ, Destouni A, Pascottini OB, Besenfelder U, Havlicek V, Smits K, Kurg A, Salumets A, et al. 2017. Genome stability of bovine in vivo-conceived cleavage-stage embryos is higher compared to in

vitro-produced embryos. *Hum Reprod* **32**: 2348–2357.

Turner TN, Coe BP, Dickel DE, Pennacchio LA, Darnell RB, Eichler EE, Turner TN, Coe BP, Dickel DE, Hoekzema K, et al. 2017. Genomic Patterns of De Novo Mutation in Simplex Resource Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**: 710-715.e12. https://doi.org/10.1016/j.cell.2017.08.047.

Upadhyay M, da Silva VH, Megens HJ, Visker MHPW, Ajmone-Marsan P, Bâlteanu VA, Dunner S, Garcia JF, Ginja C, Kantanen J, et al. 2017. Distribution and functionality of copy number variation across European cattle populations. *Front Genet* **8**: 1–12.

USDA ARS. 2018. Bovine reference genome ARS-UCD1.2. https://www.ncbi.nlm.nih.gov/assembly/GCA_002263795.2 (Accessed March 23, 2018).

Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, Cao H, Moon JE, Kashin S, Fuchsberger C, et al. 2015. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. *Nat Genet* **47**: 921–925.

Van Den Berg I, Hayes BJ, Chamberlain AJ, Goddard ME. 2019. Overlap between eQTL and QTL associated with production traits and fertility in dairy cattle. *BMC Genomics* **20**: 1–18.

Veerkamp RF, Bouwman AC, Schrooten C, Calus MPL. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. *Genet Sel Evol* **48**: 1–14.

Veerkamp RF, Calus MPL, De Jong G, Linde R van der, Haas Y De. 2014. Breeding Value for Dry Matter Intake for Dutch Bulls based on DGV for DMI and BV for Predictors. In *10th World Congress of Genetics Applied to Livestock Production*.

Viana J. 2020. *2019 Statistics of Embryo Collection and Transfer in Domestic Farm Animals*.

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* **160**: 554–566. http://dx.doi.org/10.1016/j.cell.2015.01.006.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS Discovery : Biology, Function, and, Translation. *Am J Hum Genet* **101**: 5–22.

Voet T, Vanneste E, Vermeesch JR. 2011. The Human Cleavage Stage Embryo Is a Cradle of Chromosomal Rearrangements. *Cytogenet Genome Res*.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol* **4**: e72. http://www.ncbi.nlm.nih.gov/pubmed/16494531.

Wade TD, Gordon S, Medland S, Bulik CM, Heath AC, Montgomery GW, Martin NG. 2014. Genetic variants associated with disordered eating. *Int J Eat Disord* **46**: 594–608.

Wang C, Lv H, Ling X, Li H, Diao F, Dai J. 2021. Association of assisted reproductive technology, germline de novo mutations and congenital heart defects in a prospective birth cohort study. *Cell Res* **148**: 148–162.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, Hakonarson H, Bucan M. 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* **17**: 1665–1674.

Wathes DC, Cheng Z, Salavati M, Buggiotti L, Takeda H, Tang L, Becker F, Ingvartsen KI, Ferris C, Hostens M, et al. 2021. Relationships between metabolic profiles and gene expression in liver and leukocytes of dairy cows in early lactation. *J Dairy Sci* **104**: 3596–3616. http://dx.doi.org/10.3168/jds.2020-19165.

Weischenfeldt J, Symmons O, Spitz F, Korbel JO. 2013. Phenotypic impact of genomic structural variation : insights from and for human disease. *Nat Rev Genet* **14**: 125–138.

Weissensteiner MH, Bunikis I, Catalán A, Francoijs KJ, Knief U, Heim W, Peona V, Pophaly SD, Sedlazeck FJ, Suh A, et al. 2020. Discovery and population genomics of structural variation in a songbird genus. *Nat Commun* **11**: 1–11. http://dx.doi.org/10.1038/s41467-020-17195-4.

Welch JL. 1940. Famous Individuals in the History of the Jersey and Holstein-Friesian Breeds. *Iowa State Univ Vet* **2**: 111–140.

Wellenreuther M, Bernatchez L. 2018. Eco-Evolutionary Genomics of Chromosomal Inversions. *Trends Ecol Evol* **33**: 427–440.

Werling DM, Brand H, An J, Stone MR, Zhu L, Glessner JT, Collins RL, Dong S, Layer RM, Markenscoff-papadimitriou E, et al. 2018. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* **50**. http://dx.doi.org/10.1038/s41588-018-0107-y.

Wolfer MF, Miller DE. 2016. Alfred Sturtevant Walks into a Bar : Gene Dosage, Gene Position, and Unequal Crossing Over in Drosophila. *Genetics* **204**: 833–835.

Wong WSW, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, Baker R, Thach DC, Iyer RK, Vockley JG, et al. 2016. New observations on maternal age effect on germline de novo mutations. *Nat Commun* **7**: 1–10.

Wray NR. 2005. Allele Frequencies and the r2 Measure of Linkage Disequilibrium : Impact on Design and Interpretation of Association Studies. *Twin Res Hum Genet* **8**: 87–94.

Wright S. 1950. Genetical structure of populations. *Nature* **166**: 247–9.

Wu FL, Przeworski M, Moorjani P, Przeworski M, Strand AI, Cox LA, Cox LA, Ober C, Wall JD, Strand AI, et al. 2020. *A comparison of humans and baboons suggests germline mutation rates do not track cell divisions*. http://dx.doi.org/10.1371/journal.pbio.3000838.

Xiang R, van den Berg I, MacLeod IM, Daetwyler HD, Goddard ME. 2020. Effect direction meta-analysis of GWAS identifies extreme, prevalent and shared pleiotropy in a large mammal. *Commun Biol* **3**: 1–14. http://dx.doi.org/10.1038/s42003-020-0823-6.

Xu L, Cole JB, Bickhart DM, Hou Y, Song J, VanRaden PM, Sonstegard TS, Van Tassell CP, Liu GE. 2014. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC Genomics* **15**: 1–10.

Xu L, Hou Y, Bickhart DM, Zhou Y, Hay EHA, Song J, Sonstegard TS, Van Tassell CP, Liu GE. 2016. Population-genetic properties of differentiated copy number variations in cattle. *Sci Rep* **6**: 1–8. http://dx.doi.org/10.1038/srep23161.

Xue Y, Sun D, Daly A, Yang F, Zhou X, Zhao M, Huang N, Zerjal T, Lee C, Carter NP, et al. 2008. Adaptive Evolution of UGT2B17 Copy-Number Variation. *Am J Hum Genet* **83**: 337–346.

Yamamoto N, Kumashiro R. 1993. Conversion of vitamin D3 binding protein (group-specific component) to a macrophage activaating factor by the stepwise action of beta-galactosidase of B cells and sialidase of T cells. *J Immunol* **151**: 2794–2802.

Yan SM, Sherman RM, Taylor DJ, Nair DR, Bortvin AN, Schatz MC, McCoy RC. 2021. Local adaptation and archaic introgression shape global diversity at human structural variant loci. *Elife* **10**: 1–29.

Yang J, Lee SH, Goddard ME, Visscher PM. 2011. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**: 76–82. http://dx.doi.org/10.1016/j.ajhg.2010.11.011.

Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL. 2014. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**.

Zamani Esteki M, Viltrop T, Tšuiko O, Tiirats A, Koel M, Nõukas M, Žilina O, Teearu K, Marjonen H, Kahila H, et al. 2019. In vitro fertilization does not increase the incidence of de novo copy number alterations in fetal and placental lineages. *Nat Med* **25**: 1699–1705.

Zeleny. 1919. A change in the Bar gene of Drosophila involving further descrease in facet number and increase in dominance. *J Gen Physiol* 69–71.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* R137.

Zhang Z, Guillaume F, Sartelet A, Charlier C, Georges M, Farnir F, Druet T. 2012. Ancestral haplotype-based association mapping with generalized linear mixed models accounting for stratification. *Bioinformatics* **28**: 2467–2473.

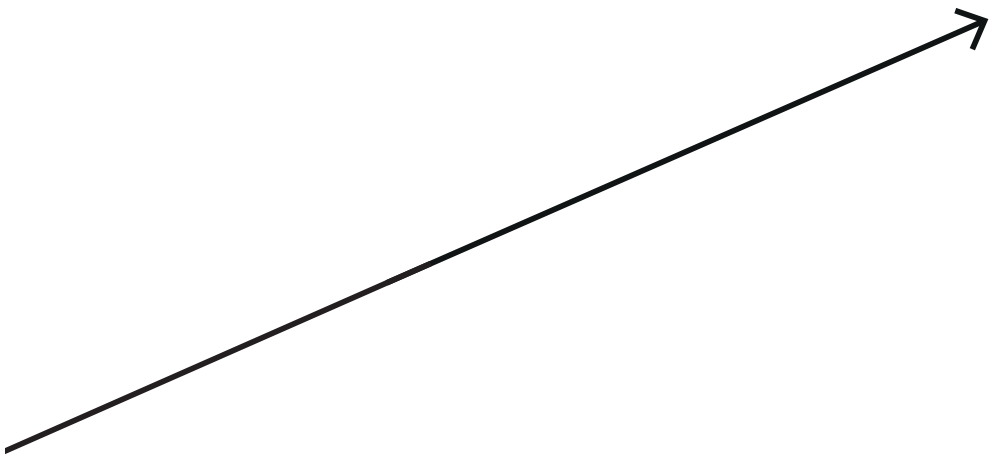
Zhao X, Collins RL, Lee WP, Weber AM, Jun Y, Zhu Q, Weisburd B, Huang Y, Audano PA, Wang H, et al. 2021. Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies. *Am J Hum Genet* **108**: 919–928. https://doi.org/10.1016/j.ajhg.2021.03.014.

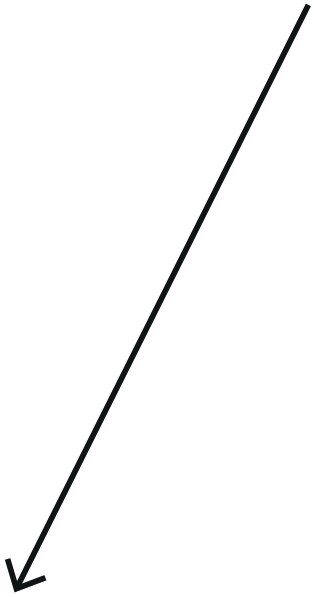
Zhou Y, Connor EE, Wiggans GR, Lu Y, Tempelman RJ, Schroeder SG, Chen H, Liu GE. 2018. Genome-wide copy number variant analysis reveals variants associated with 10 diverse production traits in Holstein cattle. 1–9.

Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM, et al. 2016. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**.

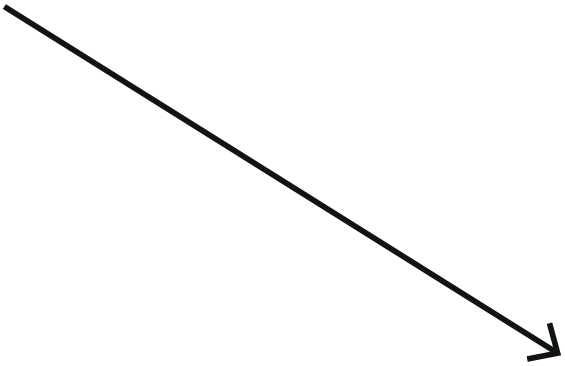
Zimin A V., Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, et al. 2009. A whole-genome assembly of the domestic cow, Bos taurus. *Genome Biol* **10**.

Zwald NR, Weigel KA, Chang YM, Welper RD, Clay JS. 2004. Genetic Selection for Health Traits Using Producer-Recorded Data . I . Incidence Rates , Heritability Estimates , and Sire Breeding Values. *J Dairy Sci* **87**: 4287–4294. http://dx.doi.org/10.3168/jds.S0022-0302(04)73573-0.





APPENDICES



A

Summary

The DNA is hereditary material that harbours numerous genetic variants inherited from an individual's ancestors. Of these variants, a small fraction are spontaneously occurring de novo mutations that can persist in the population variants pool, if not lost by drift. Cattle is a livestock species with high economic significance; thus, deciphering how its genome and genomic information relates to its phenotype is crucial. However, investigation on the mutational processes and population variants with high impact in cattle, thus far, has been limited to SNPs. As a results, less tractable and complex variants, such as structural variants, despite affecting more bases than small varaints, have not been deeply investigated. To close this gap, I analysed bovine structural variants, using data from genotyping arrays and multi-generational deeply sequenced genomes. This thesis generated catalogues of SVs segregating in dairy cattle populations, and demonstrates that SVs can have molecular and phenotypic impacts. In particular, I disected one largest QTL for clinical mastitis and unravelling a 12-Kb multiallelic CNV as the casative variant. Furthermore, my work showed that the mutational mechanisms of SVs are likely inherently different compared to SNPs, highlighting the importance of a comprehensive survey of mutational processes.

In **Chapter 2**, I made a catalogue of CNVs, based on high-density SNP-array data generated from two dairy cattle breeds. I showed that CNV discovery results could vary, depending on the quality of reference genomes. Exploiting the allele frequencies of the CNVs, I highlighted that some CNVs likely differentiated between the two breeds, which might undergo recent selection. Furthermore, linkage disequilibrium between SNP-CNV pairs was generally low compared to SNP-SNP pairs.

Chapter 3 describes an improved SV catalogue exploiting deeply sequenced bovine genomes. This catalogue contains many small SVs undiscovered in the previous catalogue (chapter 2), many of which have sequenced resolved breakpoints. Using a direct genotyping approach, I confirmed that nearly 80% of the SVs were present in an independent cohort of animals. Using sequenced level variants (SNPs and SVs), I showed that most SVs have tagging SNPs; however, findings were discrepant when using SNP-array data (50K SNPs and directly genotyped CNVs). This finding indicated that the variation arising from CNVs might not be fully captured based on SNP array data alone. Finally, I investigated high-impact SVs and mapped two SV-eQTL which alter gene expression.

In **Chapter 4**, I dissected a major QTL for clinical mastitis located on chromosome 6. By fine-mapping this region, I discovered the lead variants downstream of the GC gene within a 12-kb CNV. This CNV encompasses the 3' alternative exon of the GC gene. By exploiting the pedigree structure in the data set, I delineated the multi-allelic nature of the CNV, of which the multiplicated allele underwent recent positive selection. The liver eQTL mapping results showed that the CNV is a lead variant for GC expression at both gene and transcript levels.

Furthermore, the epigenome data confirm the presence of an enhancer signal within the CNV. Thus, the GC CNV likely alters the enhancer activity and subsequently upregulates GC.

In **Chapter 5**, I investigated the *de novo* structural variants (dnSVs) arising in the bovine germline by exploiting the multi-generational pedigrees. Of the 20 dnSVs detected, 15 occurred during gametogenesis, whereas the rest arose during embryonic development. The 15 germline dnSVs, detected in 127 pedigrees, correspond to 1 dnSV per 8.5 births, similar to the rates estimated in human cohorts. Of these 15 germline dnSVs, 14 were of paternal origin, unravelling an extreme paternal bias of 14:1. Furthermore, the majority of dnSVs were found in IVF produced probands, suggesting that the application of reproductive technology can be mutagenic.

Finally, in **Chapter 6**, I brought the findings from Chapters 2-4 together and discussed them in a broader perspective. I discussed two functional SVs from Chapter 3 (ORM1 duplication) and Chapter 4 (GC CNV) from an evolutionary point of view. Subsequently, I discussed models accounting for de novo mutations and explained why the DNA damage model fits better with our observations for the bovine germline dnSVs. Next, I elaborated on future research trends in SVs that will advance our knowledge. Lastly, I ended the discussion with a practical guideline for detecting and utilising functional SVs.

Glossary

Abbreviation	Explanation
ART	Assisted reproductive technology
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
CDS	Coding sequences
ChIP-seq	Chromatin immunoprecipitation sequencing
CM	Clinical mastitis
CN	Copy number
CNV	Copy number variants
CNVR	Copy number variant regions
dnCNV	*de novo* copy number variants
DNM	*de novo* mutations
dnSNV	*de novo* single nucleotide polymorphisms
dnSV	*de novo* structural variants
DSB	Double strand breaks
EBV	Estimated breeding value
GS	Genomic selection
GWAS	Genome-wide association study
LD	Linkage disequilibrium
MAF	Minor allele frequency
NGS / WGS	Next generation sequencing / Whole genome sequencing
QTL	Quantitative trait loci
RD	Read depth
SD	Segmental duplications
SV	Structural variants

Acknowledgements

My PhD journey ends here. I want to thank people who have been indispensable in reaching this milestone.

Roel and Martien, thank you very much for allowing me to be an experimental unit of this ambitious project aiming to connect quantitative genetics and genomics **Roel**, thank you for your pragmatic supervision, which helped me focus on the research goals. **Martien**, thank you for trusting me and giving me opportunities to explore research topics. Admittedly, my job required me to sit in front of a monitor all the time, but I felt like I was on a rollercoaster ride. My point is that it was that dynamic, with many surprises and sparsely with some sweet successes on the way. Thank you for guiding me through the journey. Will you consider having biological replicates for this type of experimental project? :-)

This journey would be incomplete without the help from my daily supervisors, Aniek and Mirte. Your willingness to support me has laid the foundation for my project. **Aniek**, you have generously taught me many things, both in- and outside of the science domain. I hope to become a manager like you one day, with all the things I learned from you. **Mirte**, having you in the supervision team was very special! I really appreciate that you guided me patiently towards the end of my PhD, and at the same time, you made sure that I am scientifically happy in my journey. Your gentle remarks about my struggle were enlightening and made me reflect on myself. Your advice was like a compass to me – keeping me from losing sight. Thank you!

I spent a few months at the Georges lab in Liege, Belgium, for research collaboration. **Michel**, thanks a lot for inviting me to your team and generously sharing Damona, the divine cows that I have been obsessed with. You have shown me the beauty in science, which will eventually bloom with patience, time, and a lot of suffering. ;-) Thank you for always bringing me back to normal Lim (from suffering Lim) with silly jokes! **Carole**, you are a special one! I wholeheartedly cherish the Charlier group's wonderful and unprecedented tradition: dr Charlier will bring cakes when PhDs give presentations. You enabled me to reach beyond my capabilities. Thank you for taking me on as a postdoctoral researcher, and I look forward to further advancing science together with you. It will be an exciting ride! **Tom**, thank you for kindly helping me whenever I struggled. You are not only an intelligent scientist but also one with high integrity. Having met you during my PhD will have a lasting impact during the rest of my career.

My project was supported by the Breed4Food consortium, which gave me the opportunity to interact and learn from scientists from the breeding industry. **Erik**, thanks for supporting my visit to Michel's group. I appreciate your interest in my project. **Chris**, thanks for your help in getting to know the Jersey:Holstein ratio of the samples. It was a bit of detective work! :-) Although my work was done for the bovine genome, I was also connected with and learned from non-cattle scientists: **Barbara**, **Rachel**, **Ron**, **Marco**, and **Katrijn**. Thanks for your genuine interest in my work. Your comments during the project meetings made me critically reflect on my work and gave me new inspiration.

My thanks goes further to my ABG colleagues, who provided me with a stimulating environment during my PhD. **Ole**, you will always be my very special ex-supervisor, handing out pieces of wisdom whenever I am looking for one. **Richard**, thank you for being in the (background of) the best selfie of my life. I hope to renew this record with you once again. Any plans for upcoming conferences? **Sipke-Joost**, thank you for enduring my endless questions about Dutch traditions and society during our ride to Davos. I will pay you back your favour next time! Several people ran this PhD Marathon with me: **Marieke**, **Malou**, **Ibrahim**, **Henri**, and **Renzo**, we started our PhD around the same time, and now we are finding our places as early career scientists. Trots op ons, for marking this special moment of our lives together!

I have claimed to be an honorary member of office E.0201. **Maria**, you were the trademark of the office E.0201. When you enter a room, it becomes 10 times brighter (...and louder!!). Thank you for your cheerful energy! **Langqing**, my special ex-supervisor (II), wie geht's? while you were gone, Lijing (see below) has been supplying amazing Chinese food to me. My best Chinese chef ranking may permanently change if you are away for too long. Don't stay too far from Wageningen! **Zhoucitta**, the bird genome enthusiast! I have to be honest with you; I think you sound way better with a Scottish than a Californian accent; please refrain from using the latter. **Harmen**, thank you for teaching me how to write a motivation letter (to secure a house – it worked!); I see a great talent as a teacher in you ;-).

Genomics meetings every Monday was like a lookout where I could see everyone was doing well. My old colleagues **Vinicius** and **Maulik**, I miss you guys and your looooooooong presentations! **Rayner**, **Marta**, **Yun**, **Xiaofei**, **Chiara**, **Jani**, **Carolina**, and **Gibbs**, interacting with you guys always motivated me to progress. Thanks a lot! **Martijn**, thanks a lot for helping me with so many things. In my defence, my endless requests for help prepared you for supervising nasty students, no? :-) **Pauline** and **Roy**, you guys are the wizards of food revenge, the most shocking of which was the Thai BBQ. Will you please come to Wageningen again for nice long evenings together? :-) **Chrissy**, despite the distance, we managed to supervise Marije together, and I am very proud of the outcome. See you soon in Rotterdam! **Marije**, it was fun to work with you, I already look forward to our paper together! **Siyuan** and **Qitong**, hanging out with you guys in Beijing was great! Will you bring me to that crazy hot pot place again? Next time, please choose less spicy food (T_T).

Last but not least, there are colleagues that I cannot put in a particular category but that nevertheless are highly valued. **Tom**, thank you for sharing a survival guide for postdoctoral researchers. I know that it will be you who owns 10-Gb of drunk Lim pictures... you are obsessed with pictures! **Pascal**, my desk pal, you know that I have had a bad conscience due to my messy desk, right? Your tidy desk has been my role model, but unfortunately, I failed to reach your tidiness (sorry). **Biaty**, the animal geneticist gone to plant genetics. I know why you are doing that - Elias keeps telling me, "Come to the green side" :-). I wish you lots of fun with crop breeding! Buongiorno **Haibo**, how is life in Italy? I am sure you will not miss the Netherlands in terms of food. Please enjoy on my behalf!

My time in Liege would have been boring without the amazing colleagues I came across. Lijing

and Gabriel, you two are my research half-sibs (biologically not related at all!!). **Lijing**, you seem very serious, but once you start talking, we could easily spend hours chatting over loads of different things. Thank you for supplying me with amazing food and papers. **Gabriel**, you are the hidden hero who enabled the discovery of the GC CNV. The first trio you gave me had Mendelian inconsistency! I will always think of you when I think about the GC CNV paper, Obrigado! **Can**, are you still sitting alone in Tom's office? :) Thanks for the GWAS analyses! **Haru**, thanks a lot for generously sharing your unpublished data. I love your portrait-based document archiving system. It is unheard of, and I am sure nobody tried it either! ;-) **Miyako**, I really like that we've shared many things (cakes, beers, cheese, information about opening hours of garden centers, etc.). I admire your capability for handling Michel's "*ready-to-explode*" agenda. You always manage to put my meetings magically. Thank you so much! **Wouter**, thanks for answering all my random questions – notably, your response rate went up in relation to the amount of chocolates supplied by dr Charlier ;-).

Thank you, my friends, who gave me endless courage to stay positive and keep going! **Michi *et al.,*** you made Parma my "pseudo" hometown, and yes, I often feel homesick. The sunshine, food, beauty in everything. I never get enough of Parma! Thank you for being there and listening to my complaints and all. And, you know that my thanks expand to all the people and animals (*Equus caballus* and *Canis lupus familiaris*) I met via Michi, right? :-) **Marielle** and **Archi**, my very special friends. We evolved from Wageningse lunch mates to life-long friends. I find it so strange that we are so different but also very similar – which makes us great friends! Hartelijk bedankt voor sharing many things with me. **Yuqi**, thank you for making my time in China unforgettable. Beijing would not be the same without you! **Yongran**, there are many things I have to thank you for, but among them, your tip where I can find mountain garlic plants in Wageningen is the top one. Xie xie! **Alders family**, you are to be blamed for why I stayed so long in the Netherlands. The beautiful memories we shared anchored me here. Thank you for sharing your news every now and then. I appreciate them.

There are a few legendary Koreans who deserve to be mentioned here. **Je-seung**, without you, Wageningen is only half fun. Thank you so much for making my time in Wageningen so flavourful! My dear ladies: **Suyeon**, **Dayoung**, and **Dawn**, how lucky I am to have friends like you with whom I can share sweet and sour PhD life? Thank you so much for our quality time. Not to forget, my ancient Wageningse friends, **Sejong**, **Sun**, and **Baekhyun** – who ever imagined that I would stay in this land of wooden shoes this long? The beautiful memories we shared in Wageningen are always in my heart! **KT Lee**, thank you for the warm encouragement. I wanted to prove that you are right (that I am doing good!).

Being far from home was made far less lonely thanks to people I call "Kaiserliche Menschen". Thank you **Christoph** and **Annette**, for being so warm and kind to me. **Sarah**, **Hannes**, and **Lene**, you are my surrogate sisters and brother. Spending time with you guys reminds me of nice coffee, cakes, and happy moments. Thank you so much for welcoming me every time! My thanks goes to the other side of the globe, where my base camp is located. Papa, mama, Mimi, and Uk-Jae – I cannot thank you enough for everything I achieved. **Papa** and **mama**, thank you for making me who I am. My stubbornness and determination annoyed a fair share

of people around me, but I say, genetics is genetics: I know where I got it from. **Mimi**, I will never be able to thank you enough, especially for the past few years. Thank you for staying strong! **Uk-Jae**, I always want to be a kinder "nunaya" to you. I will support you whatever you do, wherever you are!! Dear **Elias**, dr Kaiser, the first doctor in our small household. Do you realize how many things we have done together past few years, one being this PhD dissertation? :-) Thank you for being the single student in my linkage disequilibrium lecture. Your eagerness to learn genetics is much appreciated. By now, your understanding of LD is as solid as my understanding of stomatal conductance. I will thank you eternally for making me brave and fearless. Let us be further blissed by our GxE interaction!

About the author

Young-Lim (Lim) Lee

Research Experience

- **March 2022 to present**, *Characterizing de novo mutations in the bovine germline*, postdoctorate research, Unit of Animal Genomics, GIGA-R, Veterinary faculty, University of Liege, Belgium
- **January 2018 to February 2022**, *Structural variants in the bovine genome*, PhD, Animal Breeding and Genomics, Wageningen University & Research, the Netherlands
- **May 2017 to October 2017**, *Genetics of resilience indicators in sows and their relation to sow longevity*, master thesis, Animal Breeding and Genomics, Wageningen University & Research, the Netherlands
- **June 2016 to May 2017**, *Detection of selection signatures and phylogenetic analysis of the endangered Pygmy hog (Porcula Salvania)*, master thesis, Animal Breeding and Genomics, Wageningen University & Research, the Netherlands
- **September 2014 to December 2014**, *Understanding boar libido issues - genetic analyses on binary traits measured in boars*, Bachelor thesis, TOPIGS Research Centre IPG, the Netherlands
- **February 2014 to June 2014**, *Variation in individual feed intake during finishing period as a predictor of PRRS outbreaks*, Bachelor internship, TOPIGS Research Centre IPG, the Netherlands

Education

- **September 2015 to October 2017**, Research Master, *Animal Science (Specialization in Animal Breeding and Genetics)*, Wageningen University & Research, Wageningen, the Netherlands (honors: cum laude)
- **September 2011 to March 2015**, Bachelor of Science *Animal Husbandry*, Van Hall Larenstein University of Applied Sciences, Wageningen, the Netherlands
- **March 2007 to February 2011**, Bachelor of Science *Business Management*, Hanyang University, Seoul, South Korea (honors: cum laude)

Peer-reviewed Publications

- **A 12 kb multi-allelic copy number variation encompassing a *GC* gene enhancer is associated with mastitis resistance in dairy cattle**; Lee YL, Takeda H, Costa Monteiro Moreira G, Karim L, Mullaart E, Coppieters W, The GplusE consortium, Appeltant R, Veerkamp RF, Groenen MAM, Georges M, Bosse M, Druet T, Bouwman AC, Charlier C; **PLoS Genetics, 2021**

- **Functional and population genetic features of copy number variations in two dairy cattle populations**; Lee YL, Bosse M, Mullaart E, Groenen MAM, Veerkamp RF, Bouwman AC (2020); **BMC Genomics, 2020**

- **Genomic analysis on pygmy hog reveals extensive interbreeding during wild boar expansion**; Liu L, Bosse M, Megens HJ, Frantz L, Lee YL, Irving-Pease E, Narayan G, Groenen MAM, Madsen O (2019); **Nature Communications, 2019**

Manuscripts in preparation

- **High-resolution structural variation catalogue in deeply sequenced cattle genomes**; Lee YL, Bosse M, Takeda H, Costa Monteiro Moreira G, Karim L, Coppieters W, Veerkamp RF, Groenen MAM, Georges M, Bouwman AC, Charlier C

- **Extreme paternal bias in bovine *de novo* structural mutations in *in vitro* produced embryos including a high proportion of post-fertilization events**; Lee YL, Bouwman AC, Bosse M, Costa Monteiro Moreira G, Harland C, Karim L, Veerkmap RF, Groenen MAM, Mullart E, Coppieters W, Georges M, Charlier C

Contributions to conferences

- **131 deeply sequenced cattle trios reveal parent-of-origin effects in *de novo* structural variations**; Lee YL, Costa Monteiro Moreira G, Karim L, Mullaart E, Coppieters W, Veerkamp RF, Groenen MAM, Bosse M, Bouwman AC, Georges M, Charlier C (2021) 72nd Annual Meeting of the European Federation of Animal Science, 30 Aug - 3 Sep, Davos, Switzerland

- **Pedigree-based *de novo* rate estimation for structural variation in the cattle germline**; Lee YL, Costa Monteiro Moreira G, Karim L, Mullaart E, Coppieters W, Veerkamp RF, Groenen MAM, Bosse M, Bouwman AC, Georges M, Charlier C (2020) 32nd The Biology of Genomes, 11-14 May, Virtual meeting

- **Fine resolution CNV catalogue from deeply sequenced cattle genomes**; Lee YL, Takeda H, Costa Monteiro Moreira G, Karim L, Bosse M, The GplusE consortium, Bouwman AC, Mullaart E, Coppieters W, Georges M, Druet T, Charlier C (2020) 71st Annual Meeting of the European Federation of Animal Science, 1 -4 Dec, Virtual meeting

- **High resolution copy number variation analysis using two cattle genome assemblies**; Lee YL, Bosse M, Veerkamp RF, Mullaart E, Groenen MAM, Bouwman AC (2019)., 70th Annual Meeting of the European Federation of Animal Science, 26 - 30 Aug, Ghent, Belgium

- **Copy number variants reveal traces of recent selection in two dairy cattle breeds**; Lee YL, Bouwman AC, Groenen MAM, Mullaart E, Veerkamp RF, Bosse M (2019) International Society for Animal Genetics Conference, 7-12 July, Lleida, Spain

- **Detection of selection signatures and phylogenetic analysis of the highly endangered Pygmy hog (*Porcula salvania*)**; Lee YL, Liu L, Bosse M, Frantz LAF, Narayan G, Megens HJ, Groenen MAM, Madsen O (2018) Netherlands Society for Evolutionary Biology, 11 April, Ede, the Netherlands

Training and Supervision Plan

EDUCATION AND TRAINING	Year(s)	ECTS
A. The Basic Package		
WIAS Introduction Day	2018	0.3
Course on philosophy of science and/or ethics	2018	1.5
B. Disciplinary Competences		
Writing research proposal	2018	6.0
Population Genomics: background and tools (Exilir) - Naples, Italy	2018	2.0
Linear models in animal breeding (NOVA) - Orsa Grönklitt, Sweden	2018	3.0
ChIP-seq (wet-lab) and basic functional animal genome analysis	2019	1.5
IMAGE Genetic diversity course	2019	1.5
Rmarkdown	2020	0.6
Research visit to University of Liege, Belgium	2019-2020	2.0
Quantitativ Genetics Discussion group	2018-2021	2.0
C. Professional Competences		
High Impact Writing in Science	2019	1.3
Survival Guide to Peer Review	2019	0.3
Scientific artwork, data visualization, and infographics with Adobe Illustrator	2020	0.6
The final touch	2020	0.6
Project and time management	2021	1.5
Career Orientation	2021	1.5
Adobe InDesign	2021	0.6
D. Presentation Skills		
1st The Netherlands Society for Evolutionary Biology Meeting#	2018	1.0
37th International Society of Animal Genetics Conference*	2019	1.0
70th Annual Meeting of the European Federation of Animal Science*	2019	1.0
71st Annual Meeting of the European Federation of Animal Science*	2020	1.0
72nd Annual Meeting of the European Federation of Animal Science*	2021	
Cold Spring Harbor Laboratory meeting "The Biology of Genomes"#	2021	
E. Teaching competences		
Research Master Cluster: reviewing the proposals	2019	0.5
Practical assistant (Genomics)	2019, 2021	2.0
Supervising MSc thesis student	2021	2.0
TOTAL		35.3

*Oral presentation; #Poster presentation

Colophon

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. The use of the HPC cluster was made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR). The research visit of Young-Lim Lee to the Unit of Animal Genomics at the University of Liege (Belgium) was financially supported by WIAS.

The cover was designed by Stefan van den Heuvel.

Printed by DigiForce | Proefschriftmaken.nl