# The discovery of novel recessive genetic disorders in dairy cattle.

A thesis presented in partial fulfilment of the requirements

for the degree of

**Doctor of Philosophy**

**in**

**Animal Science**

AL Rae Centre of Genetics and Breeding, Massey University,

Palmerston North, New Zealand

Edwardo G M Reynolds

2022

# Abstract

The selection of desirable characteristics in livestock has resulted in the transmission of advantageous genetic variants for generations. The advent of artificial insemination has accelerated the propagation of these advantageous genetic variants and led to tremendous advances in animal productivity. However, this intensive selection has led to the rapid uptake of deleterious alleles as well. Recently, a recessive mutation in the *GALNT2* gene was identified to dramatically impair growth and production traits in dairy cattle causing small calf syndrome. The research presented here seeks to further investigate the presence and impact of recessive mutations in dairy cattle.

A primary aim of genetics is to identify causal variants and understand how they act to manipulate a phenotype. As datasets have expanded, larger analyses are now possible and statistical methods to discover causal mutations have become commonplace. One such method, the genome-wide association study (GWAS), presents considerable exploratory utility in identifying quantitative trait loci (QTL) and causal mutations. GWAS' have predominantly focused on identifying additive genetic effects assuming that each allele at a locus acts independently of the other, whereas non-additive effects including dominant, recessive, and epistatic effects have been neglected. Here, we developed a single-locus non-additive GWAS model intended for the detection of dominant and recessive genetic mechanisms.

We applied our non-additive GWAS model to growth, developmental, and lactation phenotypes in dairy cattle. We identified several candidate causal mutations that are associated with moderate to large deleterious recessive disorders of animal welfare and production. These mutations included premature-stop (*MUS81*, *ITGAL*, *LRCH4*, *RBM34*), splice disrupting (*FGD4*, *GALNT2*), and missense (*PLCD4*, *MTRF1*, *DPF2*, *DOCK8*, *SLC25A4*, *KIAA0556*, *IL4R*) variants, and these occur at surprisingly high frequencies in cattle. We further investigated these candidates for anatomical, molecular, and metabolic phenotypes to understand how these disorders might manifest. In some cases, these mutations were analogous to disorder-causing mutations in other species, these included: Coffin-Siris syndrome (*DPF2*); Charcot Marie Tooth disease (*FGD4*); a congenital disorder of glycosylation (*GALNT2*); hyper Immunoglobulin-E syndrome (*DOCK8*); Joubert syndrome (*KIAA0556*); and mitochondrial disease (*SLC25A4*). These discoveries demonstrate that deleterious recessive mutations exist in dairy cattle at remarkably high frequencies and we are able to detect these disorders through modern genotyping and phenotyping capabilities. These are important findings that can be used to improve the health and productivity of dairy cattle in New Zealand and internationally.

# Acknowledgements

Firstly, I would like to thank my supervisors Professor Mathew Littlejohn and Professor Dorian Garrick. Thank you for the opportunity to do this research, for the enormous amount of support and feedback I have received over the years, and for guiding me to become a scientist. Matt, thank you for the countless conversations, for challenging me in my research and for your perseverance in getting projects over the line. Dorian, thank you for your attention to detail and your invaluable expertise.

I'd like to thank my fellow PhD students, especially Thomas Lopdell and Swati Jivanji. Thomas, thank you for always being a chair swivel away to answer my queries. Swati, thank you for always having something exciting to say and being the extrovert we needed.

Thank you to everyone in the Research and Development team at LIC for welcoming me into the group, the years of experience I have gained, and the fun and friendly atmosphere. An additional thank you to Kathryn Sanders, Ric Sherlock, Chad Harland, Yu Wang, Tom Johnson, Catherine Neeley, Andrew Wallace, Chris Couldrey, Steve Davis, Anne Winkelman, Bevin Harris, and Richard Spelman who have offered their expertise or cooperated with me on research.

Thank you to everyone at the Al Rae Centre: Tony Pleasants, Rhiannon Handcock, Tiffany Kosch, Melissa Stephens, Jinghang Zhou, Liyuan Liu, and Mike Keehan. Thank you for sharing your experiences from a diversity of backgrounds. Morning and afternoon teas were always a highlight, often featuring stimulating conversations about science and the world.

Personal financial assistance from the Al Rae Centre, MPI, and LIC is greatly appreciated. Thank you to LIC for providing the datasets used in this research. Thank you to NeSI for providing the computational resources for this project.

Thank you to my friends, especially those that made Hamilton a great place to live. Cheers for the games, gins, and adventures.

Finally, thank you to my family. Thank you for the encouragement and the escapes. Without your support, this would not be possible.

# Table of Contents

# List of Tables

# List of Figures

## Extended Data Figures

# List of Abbreviations

| | |
|---|---|
| AI | Artificial Insemination |
| AR2 | Allelic R-squared |
| BW | Breeding worth |
| CMT | Charcot Marie Tooth disease |
| CSS | Coffin-Siris syndrome |
| DNA | Deoxyribonucleic acid |
| DR2 | Dosage R-squared |
| $\delta^2$ | Dominance heritability |
| EL | Embryonic lethal |
| eQTL | Expression QTL |
| GRM | Genomic relationship matrix |
| GWAS | Genome-wide association study |
| $h^2$ | Heritability |
| LD | Linkage disequilibrium |
| LOCO | Leave one chromosome out |
| LOSO | Leave one segment out |
| MAF | Minor allele frequency |
| MCMC | Markov chain Monte Carlo |
| Ne | Effective population size |
| OMIA | Online mendelian inheritance of animals |
| OMIM | Online mendelian inheritance of man |
| PCA | Principal component analysis |
| QTL | Quantitative trait loci |
| RNA | Ribonucleic acid |
| SNP | Single nucleotide polymorphism |

# Chapter 1 Literature Review and Introduction

## 1.1 Introduction to Genetics

### 1.1.1 Foundations

For millennia, favourable genetic variants were unknowingly selected through selective breeding schemes in crops and livestock to improve productivity. It was only in the late 19th century that the genetic basis of simple traits began to be unlocked with Mendel's recognition that a trait may be determined by two “factors” (i.e., alleles at a gene), one inherited from each parent (Mendel 1865). The discovery that DNA encodes genetic information (Avery, Macleod, and McCarty 1944), and the subsequent discovery of the structure of DNA (Watson and Crick 1953) were two further seminal works that set the stage for modern genetics. Since then, genetic variants that influence phenotypes have been discovered, with this information used in medicine and selected for in agriculture and horticulture to accelerate selection response.

The human genome project (Venter et al. 2001) saw a revolution in our ability to derive genetic information cheaply, allowing large scale use of genomic information for many applications. The 1,000 Genomes project began in 2008 with the goal of creating a massive catalogue of global human genetic variation (The 1000 Genomes Project Consortium et al. 2015). Through the success of the 1,000 Genomes project and further advances in sequencing technology, the study of genetic variation in humans, livestock and other species has increased exponentially (Metzker 2010). These advances included international efforts such as the 1,000 Bull Genomes Project (Daetwyler et al. 2014) which now includes thousands of cattle from over 100 breeds.

### 1.1.2 Genetic variants and mechanisms

Understanding the genetic variation in a population provides scope for all further genetic analyses like livestock selection and disease risk assessment. Genetic variants that make up this variation can come in many sizes and through differing mutation mechanisms, but by far the most common class of variant is a single nucleotide polymorphism (SNP) because they are the smallest inherited unit. A SNP is a single base-pair location that differs in nucleotide (A, C, G, or T) within a population of interest. The study of SNPs has been widespread, and many millions have been identified in cattle (Daetwyler et al. 2014).

Genetic variants can act under many different genetic mechanisms and modes of inheritance. The most commonly studied mechanism is additivity which represents the independent contribution of each allele at a locus. Non-additive mechanisms also exist such as dominance, and epistasis. Dominance mechanisms represent the intra-locus interaction between a pair of alleles, while epistasis, the most complex of these mechanisms, represents the inter-loci interaction between a set of genetic loci (Figure 1.1).

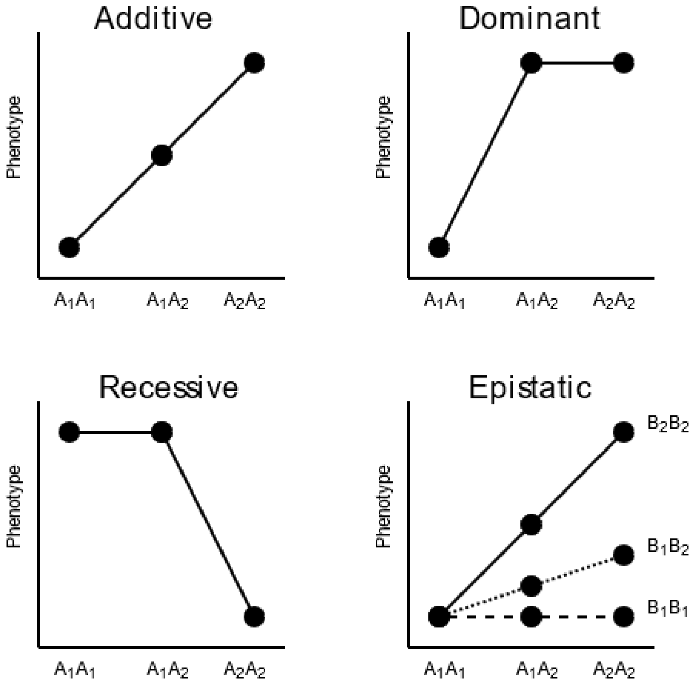


**Figure 1.1 | Examples of genetic mechanisms.**

Theoretical examples of differing genetic effect mechanisms on phenotype. $A_1$ and $A_2$ refer to the alleles at locus A, $B_1$ and $B_2$ refer to the alleles at locus B whose genotype interacts with that of locus A.

Many different examples of genes and variants that operate through these different modes of inheritance have been demonstrated in cattle. The *DGAT1* gene encodes a protein known as acylCoA:diacylglycerol acyltransferase which catalyses the final step in triglyceride synthesis (a major component of milk fat) and as such plays an important role in mammals. Studies in cattle have identified a mutation in this gene with additive effects across milk production traits such that each additional copy of this mutation increases milk-fat percentage by 0.17% and decreases milk yield by 158kg (Grisart et al. 2002).

The polled (hornless) phenotype in cattle manifests through a dominant genetic mechanism where a single mutated copy at the *POLLED* locus results in a hornless animal, regardless of the other allele at the locus (Georges et al. 1993). The most striking example of biologically recessive variants are embryonic lethal mutations. In these cases, a single mutated copy has no effect on embryo viability but in the instance of two mutated copies, the embryo fails to develop normally, usually due to the knock-out of an essential gene. Several embryonic lethal mutations have been detected in cattle populations (Charlier et al. 2016) and many more across humans and other livestock populations have been identified that are recorded in the OMIM and OMIA databases (Amberger et al. 2015; Lenffer et al. 2006).

Epistasis is the most complex of these non-additive mechanisms and while its contribution to phenotypes has been studied (Cockerham 1954), it has been very difficult to discover causal gene interactions in humans and livestock. An exception to this is the discovery of an *ERAP1-HLA* interaction in humans. Cortes et al. (2015) investigated this interaction concluding that the status of multiple HLA alleles influence the effect of *ERAP1* on an individual’s risk to multiple diseases like psoriasis

(Strange et al. 2010), ankylosing spondylitis (Evans et al. 2011), and Bechet's syndrome (Kirino et al. 2013).

An animal's characteristics can be influenced by genetic loci acting under these mechanisms or combinations thereof. A causal mutation or causal variant is a DNA variation that has a causative impact on a trait. Such a mutation may alter the structure of a protein, interfere with the expression of a gene, or act through some other regulatory pathway to cause a change in an animal's characteristic. Often it can be difficult to identify the precise nucleotide change that is the causal mutation, however quantitative trait loci (QTL) or variants that are genetically linked with the causal mutation (tag variants) can be more easily identified. A QTL is a genomic position or region with a quantitative impact on a quantitative trait although the biological mechanism underpinning this impact is unknown (Lynch and Walsh 1998).

The minor allele frequency (MAF) of a bi-allelic variant is the frequency of the less-common allele in the population of interest and indicates the prevalence of the variant in the population. Ancient variants occur at varying frequencies due to the effects of long-term natural and artificial selection and are often classed as common (MAF > 5%), low frequency (1 – 5%), or rare (<1%), although the threshold of these classes vary from study to study (Yang et al. 2010; Qianqian Zhang et al. 2018). Common variants make up the majority of studies focussed on genetic variation due to the relative ease of detection and improved associative power of common variants (assuming the underlying causal mutation is also common). Despite genetic variation contributing to 80% of the variation of human height, only 60% of this can be explained by common or low frequency variants (Wood et al. 2014), suggesting the importance of rare causal mutations.

Rare and low frequency variants are harder to detect and characterise, however as sample sizes increase, variants with lower allele frequencies can be more readily investigated. Rare variants can have large effects on characteristics, where, for example, in an Icelandic human population rare variants were found to be highly associated with type II diabetes risk, height and body mass index (Steinthorsdottir et al. 2014). In a study of human height in 711,428 individuals, 83 rare or low frequency coding variants were identified, some with large effect (2 cm), demonstrating that with a large enough sample size these variants can be accurately associated with traits (Marouli et al. 2017). An attempt to investigate the use of rare and low frequency variants in genomic prediction of fertility and welfare traits in cattle showed that while on simulated datasets rare variants could increase prediction reliabilities, on real data they did not result in the same improvement., Assuming this discrepancy derives from the difficulty in pinpointing rare variants that are actually causal, this finding may highlight the importance of that endeavour to utilising rare variant information (Qianqian Zhang et al. 2018).

### 1.1.3 Genetic Architecture

The genetic architecture of a trait represents the number, mechanism, frequency, and effect size of all the causal mutations that contribute to variation in a given trait (Mackay 2001). Simple Mendelian traits have a relatively simple genetic architecture where discrete characteristics are apparent from variation at a single genetic locus. Complex traits, however, are those that are influenced by many causal loci of varying effect sizes, mechanisms, and frequencies. Through discovery of a trait's genetic architecture, we improve our understanding of the fundamental biology underlying these effects, how organismal characteristics come to be, and provide scope for more accurate selection of desirable attributes.

Simple Mendelian traits are often the result of dominant or recessive mechanisms and are mostly characterised by Mendelian diseases such as cystic fibrosis in humans or embryonic lethal mutations in any diploid species (Gasparini et al. 1990; Charlier et al. 2016). Mendelian diseases in humans have been of great interest in medical genetics (McKusick 2007), with over 6,000 mutations identified that mostly cause disease-related phenotypes (Amberger et al. 2015). In animals, Mendelian disease has attracted less attention but nevertheless over 150 causal mutations have been identified for various Mendelian traits (Lenffer et al. 2006). Aside from inherited diseases, cattle present several striking phenotypic phenomenon which comprise simple Mendelian traits such as the red or black coat colour characteristic (*MC1R* gene, (Klungland et al. 1995)), the polled phenotype (Georges et al. 1993), the belted coat colour seen in Galloway cattle (*TWIST2* gene; (Mishra et al. 2017)), double muscling seen in Belgian Blue cattle (*MSTN* gene (Grobet et al. 1997)), and the slick coat of Senepol cattle (*PRLR* gene (Littlejohn, Henty, et al. 2014)). However, many traits including those of interest to breeders such as growth rate, fertility, and milk production are complex and are influenced by many loci.

Complex traits are those that are influenced by variation at multiple loci. These traits are often quantitative, like human height and milk productivity, and can be affected by many genetic mechanisms such as additivity, dominance, epistasis as well as exhibiting gene by environment interactions. In humans, a prominent example is height where over 400 independent genetic loci have been identified to contribute to the variation of human height (Wood et al. 2014). Risks to complex psychiatric disorders such as schizophrenia are impacted by over 100 QTL across the human genome (Gratten et al. 2014). In cattle, 1,000's of SNPs have been identified with significant effects on milk production traits (Jiang et al. 2019). Discovering the causal variants at these genetic loci and understanding how they influence complex traits can lead to better disease mitigation and selection opportunities, and is therefore of great interest to geneticists, breeders, and other biologists alike.

## 1.2 Dairy Cattle – History & Technology

Cattle, descendant from wild aurochs, have been an important socio-economic resource to humans for over 10,000 years (Larson and Fuller 2014), and have been used as a source of milk for at least 8,000 years (Evershed et al. 2008). While geographical breed variation has been noted since Ancient Roman times (MacKinnon 2010), it is only in the past few centuries that selective breeding to change and maintain animal characteristics has been effective and widespread (Felius et al. 2014). Through phenotype recording and artificial selection, some breeds have become more homogeneous in performance, and are easily distinguished by distinct coat colours and other breed-specific characteristics.

Artificial insemination (AI) is a procedure by which semen is collected from bulls, diluted, and then artificially inseminated into cows without the bulls' presence being required. The advent of AI has allowed superior bulls to be dispersed throughout an industry (Foote 2002). Through increased uptake and improved technology, some sires have generated millions of semen straws, and through dissemination across national herds, great advances in the genetic gain of selected traits in dairy cattle has been possible internationally.

The extensive recording of pedigree by breeders and the desire to select the best sires for AI led to the development of formal approaches to selection. Genetic selection comprised new statistical methodologies such as selection indices (Hazel 1943) to combine information from different sources, and mixed model equations that can also simultaneously adjust for non-genetic effects (Henderson 1953). Selection indices such as Breeding Worth (BW; (Harris, Clark, and Jackson 1996)) have been developed to increase on-farm production and longevity, and recent additions such as fertility and somatic cell score allow for the selection of a well-balanced, healthy cow (DairyNZ 2021). The mixed model equations developed by Henderson have been instrumental through the accurate ranking of sires based on the performance of their daughters. Through the use of these methods, milk production per Holstein cow doubled over 40 years (1960 – 2000), more than half of which was attributable to improved genetics (Dekkers and Hospital 2002).

Genomic selection is a new technology through which selection can be based on statistics derived from the genetic makeup of an animal (Meuwissen, Hayes, and Goddard 2001; Hayes et al. 2009). This technique aims to use DNA genotyping to capture the variation from causal mutations across the genome and predict phenotypic variation of quantitative traits to rank animals (Georges, Charlier, and Hayes 2018). Through this approach, the breeding values of unproven young sires can be more reliable than those estimated with pedigree without performance measures on offspring. Through the use of young sires, the generation interval is reduced, and in some cases genetic gain can be accelerated (Falconer 1960; Meuwissen, Hayes, and Goddard 2001).

In New Zealand, our pasture-based dairy cattle population consists of 4.92 million cows across 11,179 herds (Livestock Improvement Corporation 2020). The national herd has a unique breed composition of Holstein-Friesian (HF; 32.7%), Jersey (J; 8.4%), crossbred HFxJ (49.1%), and other breeds (9.8%). These animals, fed primarily ryegrass and clover, produced 21.1 billion litres of milk in the 2019/20 season (Livestock Improvement Corporation 2020). The export focus of dairy in New Zealand means farmers are rewarded for milk solids (milkfat and milk-protein yields) and penalised for milk volume. These economic rewards and penalties pose a case for using a selection index. Through the development of indices like breeding worth (Harris, Clark, and Jackson 1996) along with AI, farmers can balance their systems to improve profit.

The artificial selection of cattle has led to a highly related population and can rapidly increase the frequencies of alleles carried by superior sires. Despite a population of over 4 million, the effective population size estimates of New Zealand HF and J are much smaller at approximately 100 (de Roos et al. 2008). This unique population structure  and the widespread availability of phenotypes position cattle as an interesting model to detect causal mutations and provides scope to elucidate the additive and non-additive genetic architecture of complex traits which may benefit selection.

## 1.3 Tools – Phenotypic and genetic data generation

### 1.3.1 Phenotypic datasets

Data on lactating dairy cattle is routinely gathered via herd testing as a way for farmers to make more informed selection decisions. Herd visits typically gather lactation data, but may also record live weight, reproductive events, and traits other than production (TOP traits).  Some 73.5% of herds were herd tested in the 2019-2020 season (Livestock Improvement Corporation 2020), this volume of tests results in a wealth of data that can be exploited for discovery and has been a cornerstone to the advancements in genetic gain and farm production for the past 50 years.

Lactation traits including milk volume, milk fat percentage, milk protein percentage, and somatic cell score are routinely measured via herd testing. Milk fat yield and milk protein yield are generated as the product of milk volume and their respective percentages. Owing to their economic importance and availability, the genetic architecture of lactation traits has been the focus of much study. Through these studies several causal genes have been identified to have a significant impact on lactation traits such as *DGAT1* (Grisart et al. 2002), *GHR* (Blott et al. 2003), *ABCG2* (Cohen-Zinder et al. 2005), *AGPAT6* (Littlejohn, Tiplady, et al. 2014), and *MGST1* (Littlejohn et al. 2016).

Growth and developmental traits include liveweight, stature, and body condition score (BCS). Liveweight and BCS are included in the breeding worth index as these traits are important indicators of maintenance feed intake, carcass weight, reproductive success, survival and animal health. The genes *PLAG1* (Karim et al. 2011; Fink et al. 2017) and *CCND2* (Bouwman et al. 2018) are two candidates with impacts on liveweight and stature in cattle. With over 400 QTL identified to influence human height (Wood et al. 2014) this would suggest a similar trait in cattle may be influenced by many more undiscovered genetic effects.

Traits other than production are management and conformation characteristics which are important to farmers but do not have a direct link to production. Management characteristics focus on how an animal behaves while being milked such as adaptability to milking, shed temperament, and overall farmer opinion. Conformation characteristics, scored by a TOP inspector, include measures of the form of an animal such as chest capacity, and udder and teat conformation (Advisory Committee on Traits Other than Production 2020). Wu et al. (2013) identified QTL across 26 body conformation traits similar to TOP traits and highlighted several candidate genes underlying these effects.

The economic importance of these phenotypes means many are incorporated in national genetic evaluations to produce breeding values (BVs), measures of genetic worth for given traits. These evaluations fit several non-genetic effects to account for confounding such as contemporary group, age, and stage of lactation. The weighted average of a cow's records less these non-genetic effects can be calculated for downstream analyses and are called yield deviations (VanRaden and Wiggans 1991).

### 1.3.2 Genotyping, Sequencing and Imputation

SNP-chip genotyping is a popular and efficient way to generate genetic data across many individuals for many loci simultaneously. Through SNP-chip genotyping a predetermined set of SNPs, ranging in number from a few hundred markers upwards to over a million, can be investigated across samples. In 2005, the first commercially available bovine SNP-chip was designed (Bovine 10K; (Affymetrix Inc. 2005)) and subsequent commercial designs for medium (Bovine SNP50; Illumina Inc., San Diego, CA) and high (BovineHD; Illumina Inc., San Diego, CA) density panels were released. The BovineSNP50 platform was designed with the intention of spacing at least 50,000 SNPs evenly throughout the genome and across the minor allele frequency spectrum such that at least 70% of markers were polymorphic in 21 cattle breeds (Matukumalli et al. 2009). Using this platform and a variety of others, millions of cattle have been genotyped in recent years (Wiggans et al. 2017; Georges, Charlier, and Hayes 2018).

Whole genome sequencing involves not just interrogating an individual's genotype at a predetermined set of variant locations, but instead at every position across their whole genome, using fundamentally different technology. Large scale initiatives such as the 1,000 Genomes Project (The 1000 Genomes Project Consortium et al. 2015) and 1,000 bull genomes project (Daetwyler et al. 2014) have sequenced thousands of individuals aiming to characterise the genetic variation in each species. Through sequencing technology advances, reference genome assemblies have been improved for cattle in terms of chromosomal continuity, accuracy, and completeness (The

Bovine Genome Sequencing and Analysis Consortium et al. 2009; Rosen et al. 2020). The accumulation of these advances in the quality and quantity of data enables the efficient prediction of millions of genotypes across a population through a statistical approach known as imputation.

Imputation is a method that uses haplotype phasing and linkage patterns to infer missing genotype status based on a reference panel and allele frequency distributions ( Browning and Browning 2007; Howie et al. 2012). Various imputation methods can be used to project SNP chip genotypes to higher density panels, or to sequence resolution, and in doing so imputation can increase the opportunity to find QTL and causal variants. Jivanji et al. (2019) showed through statistical techniques that new QTL and candidate causal mutations could be discovered for a variety of coat colour characteristics in cattle based on over 18 million variants imputed from a variety of medium to high density SNP-chips to sequence resolution.

## 1.4 Tools - Statistics

### 1.4.1 Quantitative Genetics

The study of complex traits is often through the lens of quantitative genetics. Quantitative genetics is a statistical framework for the study of characteristics that are affected by multiple genetic loci (Mackay 2001). Through quantitative genetics we can understand the genetic mechanisms that contribute to the genetic variation of traits. Genetic variance, that part of the phenotypic variance attributable to genetics, gives an indication of what genetic mechanisms may be assumed when assessing the trait. Heritability (Lush 1940) is the ratio of genetic variance to phenotypic variance which represents the opportunity for genetic change by natural and artificial selection (Hill, Goddard, and Visscher 2008). Narrow sense heritability ($h^2$) is the proportion of phenotypic variance attributable to additive genetic variance. Broad sense heritability ($H^2$) is the proportion of phenotypic variance explained by the total genetic variance, that is additive, dominance and epistatic genetic variance (Visscher, Hill, and Wray 2008). Understanding the genetic variance and heritabilities of phenotypes can be used in selection decisions, however, it doesn't clearly reveal the underlying biological mechanisms causing trait variation (Huang and Mackay 2016).

### 1.4.2 Genome-wide association studies

When investigating quantitative traits, we are primarily interested in what effect certain variation has on an animal's characteristics, and identifying which variants are causal. Linkage analysis was a traditional approach for mapping causal regions, where knowledge of family structures and recombination maps were combined to localise their physical location on the genome. Several well-known genes influencing milk production in cattle were mapped this way including *DGAT1* (Grisart et al. 2002), *GHR* (Blott et al. 2003), and *ABCG2* (Cohen-Zinder et al. 2005).

As SNP genotyping has become more common, there has been a rise in the popularity of association studies owing to their increased power when investigating complex traits (Risch and Merikangas 1996). Association studies attempt to estimate the effect and significance of a variant on a phenotype of interest and often consider variants spanning the entire genome, these are termed Genome-Wide Association Studies (GWAS).

A genome wide association study is a statistical analysis aimed at detecting an association between the allele distributions of a set of genetic markers (typically 20,000 to 20,000,000+) and the phenotype(s) of interest. The results of these analyses can be summarised into statistics indicating to the researcher the significance of variant association and whether the implicated genomic regions warrant further investigation. GWAS can be used as an exploratory tool to improve our biological understanding of phenotypes by identifying novel genes and enabling localised examination of the underlying genetic mechanisms. Subsequent investigation into genomic regions may include higher density genotyping, whole genome sequencing of individuals, or a multi-omics approach (Visscher et al. 2017). GWAS has led to a very large number of discoveries across species. It has fast tracked exploration of the genetic architecture of complex traits and promises opportunity in the prevention and treatment of disease as well as improvement of selection in livestock.

Typical GWAS methods involve fitting a set of variants one-at-a-time in a linear mixed model to test for association between the variant and a phenotype (Yu et al. 2006; Zhou and Stephens 2012). Linear mixed models contain both fixed and random effects and can be used to test the effects of variants while accounting for possible confounding effects. We can fit a GWAS model to estimate the effects of variants on a trait and improve our understanding of the genetic architecture of complex traits. This model equation is often written as;

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{e} \tag{1}$$

where $y$ is a vector of phenotypes, $X$ is a design matrix relating individuals to fixed effects, $b$ is a vector of fixed effects, $Z$ is a design matrix relating individuals to random effects, $u$ is a vector of random effects, and $e$ is a vector of random residual errors. In the case of fitting variants one-at-a-time, the variant is typically fitted as a fixed effect (Yu et al. 2006; Zhou and Stephens 2012), however when fitting multiple variants at a time they are typically fitted as random effects (Meuwissen, Hayes, and Goddard 2001; Fernando and Garrick 2013).

In testing for association at a bi-allelic genomic locus, typically using a SNP, an additive genetic model is typically adopted. This model aims to detect the effect of substituting the reference allele out for the alternate allele at a locus of interest, in the next generation. To do so, a variable representing each locus is created with genotype classes (G11, G12, and G22) encoded (0, 1, and 2).

A basic association analysis can be performed using a simple linear model testing the marker variable against the phenotype (omitting the *Zu* term in (1)), however such a model can result in spurious associations due to confounding effects. Confounding effects include factors that can impact the independent and dependent variables in a statistical model. Confounding effects might be attributes like age, contemporary group, and diet, but can also have a genetic basis through family structure and population stratification. It's important to account for confounding effects where possible to avoid spurious associations which can mislead researchers and impact the interpretation of results (Yu et al. 2006).

Population structure reflects the complex and entangled genetic relationships between individuals in a population. It is influenced by geography, group aggregations such as breed, family structure, or the effects of natural and artificial selection. Population structure can lead to spurious associations in GWAS by finding an association with a genetic locus due to a sub-group's ancestry rather than a causal effect (Lander and Schork 1994). The New Zealand dairy herd has a highly admixed, closely related population structure due to intense artificial selection and AI, as such accounting for this relatedness is important to avoid false discoveries. Methods for accounting for this confounding effect often include the use of relationship matrices to describe the covariation between phenotyped individuals.

Relationship matrices for *n* individuals are *n x n* symmetric matrices characterising the relationships between all pairs of individuals in a sample. Expected relationships can be constructed using pedigree information or realised relationships can be constructed using genetic marker data (Vanraden 2008). Animal geneticists have used pedigree-based relationship matrices (often denoted A) for a long time in best linear unbiased prediction (BLUP) methods (Henderson 1976) where the A matrix comprises one plus inbreeding coefficients on the diagonal and coefficients of relatedness between pairs of individuals on the off-diagonals. Increased uptake of SNP genotyping in human, model organism, and agricultural populations has led to an increase in the use of genomic relationship matrices (GRMs). GRMs offer the ability to account for unobserved pedigree and unexpected (cryptic) relatedness by defining relationships through genotypes that are identical by state (Vanraden 2008). Relationship matrices can be incorporated in linear models as fixed effects based on some principal components, or as an approach to define the variance-covariance of random genetic effects in a linear mixed model.

Principal components analysis (PCA) is a technique for reducing the dimensionality of a matrix into its most variable components represented by eigenvectors (Hotelling 1933). This allows one to add a relatively small number of additional fixed effects to a linear model to account for most of the population structure without the complexity that a mixed effects model brings. PCA has been used in human studies and has been shown to be useful in detecting novel loci (Akiyama et al. 2017). Although PCA is sometimes a valid approach to account for cryptic relatedness, linear mixed models have been shown to be more effective at finding markers of interest and avoiding spurious associations (K. Wang, Hu, and Peng 2013).

Methodological developments have made linear mixed models which incorporate a relationship matrix feasible in many studies (K. Wang, Hu, and Peng 2013). These have been shown to reduce both type I and type II error rates when faced with family-based data, or structured population data (Yu et al. 2006), or even when no population structure is present (Yang et al. 2014). Fitting a relationship matrix as a random effect in a linear mixed model is now a common procedure in GWAS analysis in many species.

Pedigree-based relationship matrices can also be used when pedigree information is available. Lopdell et al. (2017) used pedigree to investigate milk lactose components in cattle, highlighting several candidate causal variants and expression-based regulatory QTL. Pedigree has been used in a linear mixed models to detect parent-of-origin effects on mouse body mass index (Hu, Rosa, and Gianola 2015), among other applications.

When pedigree data is incomplete or unavailable, a better option might be to use genetic markers to generate a GRM. Examples include a study by Yu et al. (2006) who fitted a GRM in place of a pedigree relationship matrix in a linear mixed model detecting genetic effects in Maize. Yang et al. (2010) fitted a genetic relationship matrix as a random effect to detect QTL associated with human height. A drawback of GRMs is they can lead to double fitting, by including markers being tested for association in the GRM, a situation referred to as proximal contamination. Recently, approaches such as 'Leave one chromosome out' (LOCO), or 'Leave one segment out' (LOSO) have been adopted to avoid proximal contamination (Listgarten et al. 2012; Eu-ahsunthornwattana et al. 2014) where the marker being tested (or tag SNPs thereof) is not included in the GRM.

### 1.4.3 Frequentist Approaches

In most approaches to GWAS the genetic architecture of a complex trait is assumed to be made up of a large number of small effect loci spread across the allele frequency spectrum, termed the infinitesimal model (Fisher 1930; Barton, Etheridge, and Véber 2017). A commonly-used frequentist approach to GWAS involves fitting all variants of interest, one at a time, to compute a p-value for that variant indicating the strength of association.

Frequentist approaches to GWAS have undergone iterative development and competition to improve computational efficiency, while accounting for population structure more effectively. These methods began with fixed effect models including principal components, and progressed onto efficient mixed model techniques such as TASSEL (Yu et al. 2006) and EMMA (Kang et al. 2008) which better account for population structure (Eu-ahsunthornwattana et al. 2014). Through differing approximations of variance components, modified relationship matrices, approximated correction factors, and other improved algorithmic efficiencies, a host of different methods have been developed including EMMAX (Kang et al. 2010), FaST-LMM (Lippert et al. 2011), GCTA (Yang et al. 2011), GEMMA (Zhou and Stephens 2012), and BOLT-LMM (Loh et al. 2015). Eu-ahsunthornwattana et al. (2014) showed strong concordance between results from different mixed model methods (EMMAX, FaST-LMM, GCTA, and GEMMA) and suggested prioritising by ease of use and computational efficiency. In 2015, BOLT-LMM was released and is orders of magnitude more efficient than these previous methods both in CPU time, and memory use (Loh et al. 2015), and has been successfully applied to large GWAS of cattle (Tiplady et al. 2021).

These GWAS approaches have seen success across species and models, providing means of detecting QTL in a vast number of studies (Visscher et al. 2017). The GCTA software was used in a study of the genetic architecture of amyotrophic lateral sclerosis (ALS) in the human genome (van Rheenen et al. 2016). The authors conducted GWAS across over 35,000 individuals and over 8 million variants, identifying novel QTL associated with the disease. In the study, LOCO methodology was used to remove proximal contamination, by creating 23 GRMs each omitting markers on the chromosome being tested.

Pausch et al. (2016) used EMMAX in a GWAS study on Fleckvieh cattle to identify QTL associated with udder conformation traits. Seven morphology traits were investigated in 10,000 animals across 20 million imputed sequence variants revealing 12 QTL located in possible regulatory regions. The authors estimated a GRM using a high density SNP chip (634,109 autosomal SNPs) to account for population structure, a common requirement of studies in livestock populations. Another study used BOLT-LMM-INF to detect height and BMI QTL in humans (Yengo et al. 2018). The study investigated over 700,000 individuals through meta-analysis and identified thousands of additive QTL (including hundreds of novel associations) across the human genome. That research reflects the numbers of individuals and variants often used in modern GWAS and emphasises the need for efficient and effective software.

### 1.4.4 Bayesian Approaches

Bayesian inference considers not just the data itself, but also prior beliefs about the data to assess a hypothesis and calculate a posterior probability. In the animal breeding industry, a Bayesian regression approach is sometimes used in genomic prediction and similar methods can be applied to GWAS as well (Fan et al. 2011). Most Bayesian analyses avoid the infinitesimal model assumption made by frequentist models as well as the restrictions to inference that come with p-values and instead estimate posterior probability distributions (Stephens and Balding 2009).

The Bayes Alphabet is a series of Bayesian models that can be applied to GWAS and genomic prediction models (Meuwissen, Hayes, and Goddard 2001; Fernando and Garrick 2013). These linear mixed models describe methods for fitting multiple markers at once and allow for differing assumptions of the underlying genetic architecture of traits. These differing assumptions are reflected in the priors used for marker effects, often through mixture models.  Rather than relationship matrices used to account for population structure, the Bayes Alphabet models fit marker effects as random covariates to make inference on each SNP while accounting for all others at the same time (Meuwissen, Hayes, and Goddard 2001; Habier et al. 2011; Erbe et al. 2012). This technique has been shown to be robust in accounting for population structure (Toosi, Fernando, and Dekkers 2018).

In many Bayesian approaches, closed forms for characterising the posterior probability are not available. To overcome this problem, Markov chain Monte Carlo (MCMC) techniques such as Gibbs samplers are commonly used.  Such MCMC techniques are methods to draw plausible samples from the target distribution, such that at a sufficient chain length, inferences made on the samples in the chain converges to the inferences that would be obtained from the posterior distribution (Metropolis et al. n.d.; Hastings 1970). While some MCMC techniques update all parameters of interest at once, under the single-site Gibbs Sampler technique these parameters are updated individually (Geman and Geman 1984).  Advancements in computing efficiency have seen single-site Gibbs sampling techniques implemented across disciplines. In genetic contexts, this includes software such as Gensel (Fernando and

Garrick 2013), BLGR (Pérez and de Los Campos 2014) and JWAS (Cheng, Garrick, and Fernando 2016).

Similar to frequentist techniques, Bayesian approaches to GWAS have successfully led to the discovery of QTL across many species. Fan et al. (2011) used Bayesian techniques to investigate commercially important pig phenotypes on a medium density SNP chip, where all markers were fitted at the same time. BayesC, part of the Bayes Alphabet, was used to apply a mixture model to that data where 99.5% of markers were assumed to have zero effect at each iteration of the Markov chain. In doing so, the study identified a number of novel candidate genes affecting traits such as body size and loin muscle area.

Moser et al. (2015) describes the use of Bayesian mixture models to model different genetic architectures between inherited diseases, and estimate risk of disease in a human population. This study showed how for some architectures, BayesR (which assumes variants can come from four non-zero effect distributions) can outperform disease prediction using frequentist and Bayesian infinitesimal models like those used in GCTA. The work demonstrates how the differing genetic architectures between complex traits requires different model assumptions to best detect associated loci.

In dairy cattle, Littlejohn et al (2016) investigated milk characteristics in 42,000 cattle using a BayesB model. The BayesB model fitted over 400,000 SNPs at once assuming 0.2% had a non-zero effect in each iteration. This technique localised a 2Mbp region on Chromosome 5 further investigated at sequence resolution to identify variants near the MGST1 gene influencing milk fat percentage, likely through *cis* modulation of expression of that gene. This work demonstrates Bayesian approaches can be effective with large samples sizes and in highly structured populations.

## 1.5 Challenges – Factors influencing the power of GWAS

While heritabilities and variance components can provide an overview to the genetic architecture of complex traits, detecting QTL or better yet causal variants can elucidate how and why a phenotype varies in the population. There are several factors that influence our ability to detect QTL in GWAS including the genetic architecture of the trait itself, experimental costs and constraints involved in the study, and the models applied. These are often intertwined and present the importance of understanding their interactions when designing an experiment.

### 1.5.1 Genetic Architecture

The genetic architecture of a trait reflects the characteristics of the underlying causal variants themselves, i.e., their minor allele frequencies, effect sizes, effect mechanisms, and the linkage disequilibrium of the genomic regions around them. The additive effect size (β) and minor allele frequency (MAF) of a causal variant influence our ability to detect association signal as the contribution an allelic variant can have on genetic variance is influenced by its allele frequency. Under an additive model, a causal variant with a larger β and/or larger MAF is easier to detect owing to the increased variance contributed by the locus ($2*MAF*(1-MAF)*\beta^2$) (Falconer 1960). Studies may omit variants with MAF lower than 1 or 5% due to the lack of power to detect effects at those loci and their potential to produce spurious associations (Yang et al. 2010).

The effect mechanism of a causal variant impacts the power to detect it. Most GWAS methods search for additive effects and are successful at identifying additive QTL, however these models can struggle to detect other genetic mechanisms such as distinguishing dominance from additivity or finding epistasis. Therefore, selecting a model best suited for detecting the mechanism of interest can increase power.

Through selection and genetic drift, allele frequencies in a population can fluctuate over time. Linkage disequilibrium (LD) represents the non-random fluctuation of these frequencies, where alleles physically closer to each other are more likely to be inherited together as they are less likely to have been subject to a recombination event between them. LD is often represented as the square of the correlation coefficient between pairs of loci, $R^2$. It can be exploited in QTL studies, such that tagging a variant in linkage with a causal variant may be enough to detect the QTL. Using this approach, much of the additive genetic variance can be accounted for without mapping causal mutations (Wray 2005; Wood et al. 2014).

### 1.5.2 Experimental constraints

Experimental constraints will impact the power a study has to detect QTL. These include sample size, marker density, and imputation quality. Sample size is a major contributing factor to power when detecting causal variants, especially for low frequency and rare frequency variants. Over the last decade sample studies in human populations have grown rapidly through major consortia-based efforts such as the UK Biobank project (Sudlow et al. 2015; Bycroft et al. 2018). Studies in cattle are also presenting ever larger sample sizes ranging in the 10,000s (Littlejohn et al. 2016), to 100,000s (Jiang et al. 2019).

Marker density indicates the concentration of variants across the genome or in a genomic region. SNP chips are often designed to be low (e.g., 10,000 SNPs), medium (e.g., 50,000 SNPs), or high (e.g., 700,000 SNPs) density, and whole genome sequencing yields variants at sequence resolution (e.g., 20,000,000+ variants). The higher the density of markers, the more likely the variant set is to include or be in high LD with the causal variants and therefore more likely to detect their effect. It is thus difficult to use common variants in a single locus analysis to tag rare variants, since their differing minor allele frequencies result in reduced LD (Wray 2005). When markers are fitted simultaneously, through their linear combinations, these markers may tag a causal variant better than when fitted alone. With that in mind, it is also useful to have a marker set with a diverse minor allele frequency distribution in order to tag more of the genome.

Many studies use imputation to increase the average marker density of their sample, so high imputation quality is crucial to improve power especially for low frequency variants.  Often measured through allelic $R^2$ or dosage $R^2$, these indicators of imputation accuracy can help researchers decide which variants to test and which to avoid due to large inaccuracies (Browning and Browning 2007; Browning, Zhou, and Browning 2018). Through improvements in the accuracy and accessibility of genotype and sequence imputation, a researcher's ability to detect causal loci increases.

Through understanding the factors that influence the power to detect association, studies can be better designed to improve our knowledge of the genetic architecture of complex traits. A further consideration, alluded to above, is the mechanism by which causal variants may influence traits. Many studies investigate genetic variation through an additive model and ignore non-additive variation. This has likely been due to the assumption that the contribution of non-additive effects to quantitative traits is negligible (Hill, Goddard, and Visscher 2008), however further research has shown this is not always the case (Huang and Mackay 2016).

## 1.6 Non-additive analysis

### 1.6.1 Non-additive variance

The majority of GWAS methods and studies to date describe additive GWAS and ignore potential non-additive effects, such that the extent to which non-additive variation explains quantitative trait variation is still largely unknown. The main arguments for this omission include the apparent lack of non-additive variance which contributes to quantitative trait variation, the relative lack of software for the exhaustive computations required for mapping non-additive effects, and the unavailability of desired datasets (Varona et al. 2018). Hill et al. (Hill, Goddard, and Visscher 2008) showed that additive variance captures most if not all of the genetic variation in a population for most quantitative traits, even if there are interaction effects present in biological pathways and gene networks. This is despite reports that the inclusion of dominance effects can increase accuracy in breeding value prediction (Toro and Varona 2010).

Additivity and non-additivity draw an interesting divide between estimating variance components using quantitative genetic techniques and understanding the biological mechanism of causal mutations. In variance component estimation, additive genetic variance represents transmissible variation which includes contributions from dominance and epistasis (Crow 2010; Huang and Mackay 2016), while dominance genetic variance represents additional dominance which hasn't been attributable to additivity. While this method may not reflect the true biological nature of underlying genes, capturing the transmissible variation has been effective in selection (Crow 2010; Hill, Goddard, and Visscher 2008).  Huang & Mackay (Huang and Mackay 2016) argue that the parameterisation of variance component estimation can be manipulated to prioritise whichever variation component is most interesting, and therefore while variance estimates are useful in some situations they should not be used alone to infer the genetic architecture of complex trait.  As such, despite much of the non-additive variation of quantitative traits being accounted for by additive variation, searching for dominance and epistatic QTL using specialised models may be more powerful for the detection of causal loci.

Another reason for the paucity of non-additive studies is that the ability to detect non-additive variance is limited by the marker density of the datasets used to investigate it. While the additive variance attributed to an observed marker tagging a causal variant will reduce at the rate of LD ($R^2$) (Wray 2005), the rate of decrease is $R^4$ for dominance variance and additive x additive epistatic variance and exacerbated further for more complex multi-locus epistasis (Wei, Hemani, and Haley 2014). This means that for the same number of markers and genotyped animals we have more power to detect additive effects than non-additive effects, particularly when LD is moderate. This concern presents the importance of genotyping causal or strong tag variants through sufficient marker density to accurately detect non-additive variation. As datasets continue to increase in size, however, these studies become more tractable.

### 1.6.2 Non-additive GWAS

Non-additive models are typically extensions of the additive models described earlier. In the case of single locus models these are parameterised by two equivalent models, the 'biological' or the 'statistical' model (Vitezica, Varona, and Legarra 2013). The biological model fits a genotypic additive and genotypic dominance effect (encoded [0, 1, 2], and [0, 1, 0], respectively), an important note is that this genotypic additive effect is not the same as the additive (substitution) effect fitted in the original additive-only GWAS models. The statistical model extends the additive-only models and fits a breeding value and a dominance deviation effect (encoded [0, 1, 2], and [0, 2p, 4p-2], respectively, where p is the minor allele frequency) (Vitezica, Varona, and Legarra 2013). These are equivalent models such that either of the specifications can be fitted and their estimated effects transformed into the estimated effects that would be obtained by the other (Henderson 1985). To this end, model selection depends on desired interpretation - more details on this comparison are found in Chapter 2 – Supplementary Note.

In multi-marker GWAS there have been further attempts to extend Bayes Alphabet models to also represent the non-additive genetic architecture of complex traits. These methods aim to model the joint distribution of biological additive and dominance effects of causal mutations (Bennewitz et al. 2017). The method tends to have increased power to map QTL, however due to doubling the number of random covariates fitted, the effect of a causal mutation may spread over several closely linked markers (Bennewitz et al. 2017). The authors stress the need for large datasets to map small to medium effect mutations with this approach.

Zhu et al. (2015) investigated 79 phenotypes in humans for dominance genetic variance and dominance effects in GWAS. While, on average, traits presented relatively low, insignificant dominance variance component estimates (0.03), eight traits related to obesity, bloody pressure and heart rate had significant estimates where dominance variation explained up to 16% of the phenotypic variation in systolic blood pressure. The study comprised GWAS on 79 traits across up to 13,000 individuals on 1.1 million SNP, but only identified a single locus (adjacent the *ABO* gene) with a significant dominance effect. The researchers concluded that dominance variance is typically negligible compared to additive variance, consistent with Hill et al (Hill, Goddard, and Visscher 2008).

Powell et al. (2013) investigated additive and non-additive contributions to gene expression in a human study. Through variance component analysis the RNA expression level of 14,753 probes were identified to have significant non-zero additive components and 960 had significant dominance components. 208 expression QTL (eQTL) had significant dominance effects (32 of which could be replicated in an independent population) including 7 over-dominance effects. These findings suggest while dominance variance contributes less than additive variance, significant dominance effects can still be identified.

## 1.7 Applications in cattle

### 1.7.1 Non-additive effects in cattle

Cattle have also been investigated for non-additive variation. Jiang et al. (2017) investigated 42,000 cow records for additive, dominance, and imprinting effects. The authors estimated small, but not-negligible dominance variance components across milk production traits. The study identified a novel dominance QTL on milk yield near the *RUNX2* gene through a GWAS using an imputed 50K SNP panel.  Two years later, the same research group published a follow up study in what appears to be the largest cattle GWAS to date, an analysis encompassing over 290,000 genotyped animals (Jiang et al. 2019). Through these analyses, the study identified significant dominance SNP effects across milk production traits including dominance components near previously identified additive QTL such as *DGAT1* (Grisart et al. 2002), and *GHR* (Blott et al. 2003). These studies show there are dominance contributions to production traits in cattle and SNP effects underlying these contributions are detectable as sample sizes increase.

The above studies have quantified dominance variance across a variety of phenotypes and detected QTL presenting dominance effects. Most dominance QTL identified, however, have been minor components of otherwise additive QTL, and few have presented complete dominance or recessive mechanisms.  Charlier et al. (2016) screened cattle populations for embryonic lethal (EL) mutations and detected complete recessive mechanisms resulting in the termination of a calf. The study validated 9 low frequency loss-of-function mutations across 3 cattle breeds (NZ Jersey, NZ Holstein Friesian, and Belgian Blue cattle), where carrier x carrier crosses resulted in zero offspring homozygous for a loss of function allele.

The EL study also estimated the expected number of lethal mutations and the frequency at which they likely segregate in populations with variable effective population sizes (Ne). This work reports that while human populations (Ne = 10000) have many more lethal mutations segregating than cattle populations (Ne = 100 (de Roos et al. 2008); 1,925 variants vs 11)), the average frequency of these mutations is much greater in cattle than humans (~2.41% vs 0.13%) (Charlier et al. 2016). This suggests that while there might be fewer discrete mutations to discover in populations with small Ne, the power to discover ELs in these populations will be much greater, suggesting cattle as an interesting model that may afford greater opportunity to investigate non-additive genetic variation.

While embryonic lethality is a relatively easy to detect ‘phenotypic’ outcome, semi-lethal or reduced viability phenotypes may also have genetic origins, though have less obvious presentation. Jenko et al. (2019) investigated haplotype depletion in Irish beef breeds (Simmental, Aberdeen Angus, and Charolais) as a proxy for early termination (through lethality, or reduced viability). Through this approach, three haplotypes were identified that carried recessive lethal or semi-lethal alleles at common frequencies (8.8 – 15.2%). That study presents another example that major-effect non-additive effects exist and may segregate at alarming frequencies in highly selected populations (Jenko et al. 2019).

### 1.7.2 Small calf syndrome

Small calf syndrome (Cronshaw 2013) is a Mendelian recessive genetic disorder in Holstein-Friesian cattle resulting in animals which are much smaller than their contemporaries (Figure 1.2, (Reynolds et al. 2021)). Though dairy farmers may have known about the syndrome for 40 years prior, the discovery of the splice-site mutation in the *GALNT2* gene presents opportunity for modern genetic tools to be used to identify causal mutations of non-additive effect in cattle (Charlier et al. 2016). Although dominance has received less attention than additivity, with increasing sample sizes and increasing marker resolution the power to detect causal dominance or recessive variants may represent a new opportunity to uncover novel, major effect genetic loci.

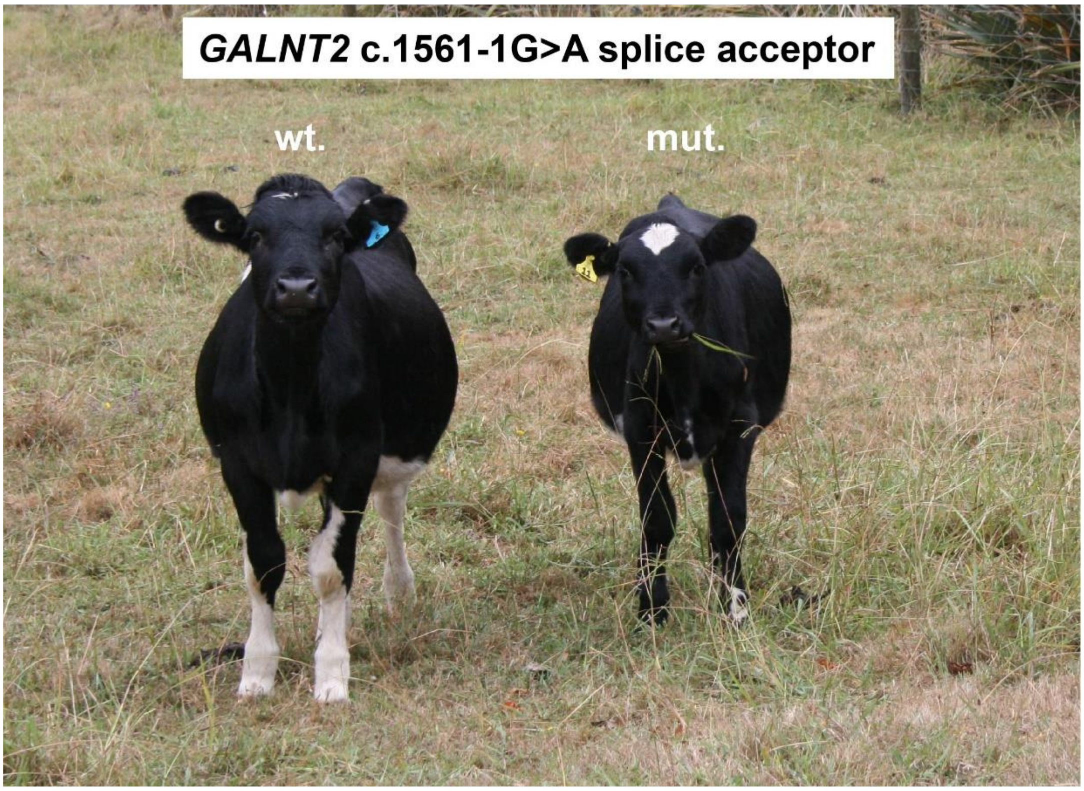


**Figure 1.2 | Photographs of *GALNT2* mutant and control animals.**
Photograph of GALNT2 control and mutant individuals, contrasting a homozygous reference (wt.) animal and a homozygous mutant (mut.) animal for the *GALNT2* c.1561-1G>A splice acceptor mutation. Animals represent individuals from research farm studies that were neither the smallest nor largest animals within each of their genotype classes. The photo has not been standardized and is provided for qualitative purposes (Reynolds et al. 2021).

## 1.8 Aims and Objectives

Artificial insemination and genetic selection have shaped the population structure of the national herd and the genetic architecture of traits important to animal production and welfare. The nature of breeding schemes presents the opportunity for allele frequencies to fluctuate rapidly in successive generations through the use of superior bulls. For the most part, this has been positive for the industry with increases in the genetic gain of selected traits, but rare deleterious variants may also be transmitted at surprising frequencies.

Non-additive variation and non-additive effects have seen relatively little study across species; however, these do exist and may be useful for selection and elucidating the genetic architecture of complex traits. Some of the difficulty in studying this architecture may be alleviated by the ever-expanding genomic datasets available for cattle, allowing variants at rarer frequencies to be investigated. The discovery and identification of the *GALNT2* small calf syndrome mutation on farm as well as several embryonic lethal mutations suggests other large effect recessive QTL may exist in the population. To this end, the main objectives of the work in this thesis were:

- Develop and implement a non-additive GWAS model to detect additive and dominance effects.
- Investigate non-additive effects and non-additive variation across quantitative traits.
- Identify causal mutations for such effects, thereby enabling new selection opportunities and options for minimising the animal welfare and health consequence of genetic disorders

Through the discovery of candidate causal mutations with recessive mechanisms across growth and developmental traits, we aimed to further investigate their impacts across other traits. Follow up objectives included:

- Validation of disease status of several candidate causal mutations in a farm trial.
- Investigate pleiotropic impacts of candidate causal variants.
- Further examine lactation traits for non-additive effects.

## 1.9 Thesis structure

Chapter 2. This chapter presents the article published in Nature Genetics detailing the discovery of several novel recessive mutations in growth and developmental traits. The Supplementary Note and Supplementary Methods have been included in the chapter as these sections provide substantial additional information. Supplementary Tables have not been included in the thesis but can be readily obtained at https://doi.org/10.1038/s41588-021-00872-5.

Chapter 3. This chapter presents un-published work to supplement the work presented in Chapter 2. It details the development and implementation of the GWAS model used to investigate non-additive effects for quantitative traits.

Chapter 4. This chapter presents un-published work to supplement the work presented in Chapter 2. It details the estimation of recombination rate in cattle and aims to investigate whether a mutation in the *MUS81* gene has an effect.

Chapter 5. This chapter presents an article published in Genetics Selection Evolution. This article details the discovery of several recessive mutations in lactation traits. Additional File 1: Table S1 has not been included in the thesis but can be obtained at https://doi.org/10.1186/s12711-021-00694-3.

Chapter 6. This chapter concludes the work presented here, with discussion on the research contributions made, and possible future work.

Appendix: The statement of contribution forms (DRC16) can be found here.

## 1.10 Related publications, conference talks, and patents

The following is a list of first author journal articles, conferences for which I was a speaker, and patents for which I am a named inventor as a result of this research.

*Journal Articles*

Reynolds, Edwardo G. M., Catherine Neeley, Thomas J. Lopdell, Michael Keehan, Keren Dittmer, Chad S. Harland, Christine Couldrey, et al. 2021. "Non-Additive Association Analysis Using Proxy Phenotypes Identifies Novel Cattle Syndromes." Nature Genetics 53 (7): 949–54. https://doi.org/10.1038/s41588-021-00872-5.

Reynolds, Edwardo G.M., Thomas Lopdell, Yu Wang, Kathryn M. Tiplady, Chad S. Harland, Thomas J. J. Johnson, Catherine Neeley, et al. 2022. "Non-Additive QTL Mapping of Lactation Traits in 124,000 Cattle Reveals Novel Recessive Loci". Genet Sel Evol 54 (5). https://doi.org/10.1186/s12711-021-00694-3

*Conferences*

Reynolds, Edwardo GM. Discovery of breed-specific genomic content in beef and dairy breeds. World Congress of Genetics Applied to Livestock Production, Auckland, New Zealand, January 2018.

Reynolds, Edwardo GM. Genome-wide association studies in cattle. School of Agriculture and Environment 2018 Symposium, Palmerston North, New Zealand, November 2018.

Reynolds, Edwardo GM. Mapping non-additive effects in a large dairy cattle population. Quantitative Genetics and Genomics (GRS), Barga, Italy, February 2019.

Reynolds, Edwardo GM. Genome wide association studies in cattle. School of Agriculture and Environment 2019 Symposium, Palmerston North, New Zealand, November 2019.

Reynolds, Edwardo GM. Non-additive effects in dairy cattle. International Conference of Quantitative Genetics 6, Brisbane, Australia, November 2020 (ONLINE).

*Patents*

Reynolds EGM, Littlejohn MD. 2019. Genetic markers and methods related thereto. NZ patent (N.Z. patent 768801). New Zealand Intellectual Property Office.

Reynolds EGM, Littlejohn MD. 2019. Genetic markers and methods related thereto. NZ patent (N.Z. patent 768802). New Zealand Intellectual Property Office.

Reynolds EGM, Littlejohn MD. 2019. Genetic markers and methods related thereto. NZ patent (N.Z. patent 768803). New Zealand Intellectual Property Office.

Reynolds EGM, Littlejohn MD. 2019. Genetic markers and methods related thereto. NZ patent (N.Z. patent 768804). New Zealand Intellectual Property Office.

Reynolds EGM, Littlejohn MD. 2019. Genetic markers and methods related thereto. NZ patent (N.Z. patent 768805). New Zealand Intellectual Property Office.

Reynolds EGM, Littlejohn MD. 2019. Genetic markers and methods related thereto. NZ patent (N.Z. patent 768806). New Zealand Intellectual Property Office.

Reynolds EGM, Littlejohn MD. 2021. Genetic markers and methods related thereto. NZ patent (N.Z. patent 777216). New Zealand Intellectual Property Office.

# Chapter 2 Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes

Edwardo G. M. Reynolds[1], Catherine Neeley[2], Thomas J. Lopdell[2], Michael Keehan[1], Keren Dittmer[1], Chad S. Harland[2], Christine Couldrey[2], Thomas J. J. Johnson[2], Kathryn Tiplady[1],[2], Gemma Worth[2], Mark Walker[2], Stephen R. Davis[2], Richard G. Sherlock[2], Katie Carnie[2], Bevin L. Harris[2], Carole Charlier[3], Michel Georges[3], Richard J. Spelman[2], Dorian J. Garrick [1] and Mathew D. Littlejohn[1],[2]

[1]Massey University, Palmerston North, New Zealand

[2]Livestock Improvement Corporation, Hamilton, New Zealand

[3]University of Liège, Liège, Belgium

## 2.1 Abstract

Mammalian species carry ~100 loss-of-function variants per individual (Charlier et al. 2016; Lek et al. 2016), where ~1–5 of these impact essential genes and cause embryonic lethality or severe disease when homozygous (Gao et al. 2015). The functions of the remainder are more difficult to resolve, although the assumption is that these variants impact fitness in less manifest ways. Here we report one of the largest sequence-resolution screens of cattle to date, targeting discovery and validation of non-additive effects in 130,725 animals. We highlight six novel recessive loci with impacts generally exceeding the largest-effect variants identified from additive genome-wide association studies, presenting analogues of human diseases and hitherto-unrecognized disorders. These loci present compelling missense (*PLCD4*, *MTRF1* and *DPF2*), premature stop (*MUS81*) and splice-disrupting (*GALNT2* and *FGD4*) mutations, together explaining substantial proportions of inbreeding depression. These results demonstrate that the frequency distribution of deleterious alleles segregating in selected species can afford sufficient power to directly map novel disorders, presenting selection opportunities to minimize the incidence of genetic disease.

## 2.2 Main

Artificial insemination enables intense selection of males, where in the case of dairy cattle, a single elite bull may be used to inseminate more than one million cows. While permitting dramatic productivity advances, this breeding strategy also promotes rare alleles, where deleterious variants can be driven to problematic frequencies in one or two generations (Daetwyler et al. 2014; Littlejohn, Henty, et al. 2014; Adams et al. 2016). Diminished effective population sizes ($N_e$) also impact the frequency distribution of deleterious alleles. We recently estimated the total number of recessive lethal mutations in Holstein-Friesian cattle (*N*$_e$ ~ 100) as approximately 100-fold fewer than the number of equivalent sites segregating in humans (*N*$_e$ ~ 10,000), although the frequency of any single variant was, on average, approximately 20-fold higher in cattle (0.13% compared with 2.49% minor allele frequency (MAF)) (Charlier et al. 2016). These divergent frequency characteristics should present contrasting power scenarios for the mapping of disease mutations, so we reasoned that recessive syndromes may be able to be mapped in the absence of formal disease classification in selected populations such as cattle – given appropriate surrogate phenotypes. Specifically, phenotypes such as body weight – routinely derived in animal breeding programs and demonstrated to represent a wide array of null alleles in mice (Reed, Lawler, and Tordoff 2008) – might serve as a proxy of such effects, and be detectable through non-additive genome-wide association studies (GWAS). To explore how different model parameters might influence this approach, we tested the sensitivity of recessive, class effect and standard-additive models for detection of a hypothetical recessive mutation. Notably, for a 2.5% MAF mutation explaining 0.1% of the phenotypic variance of a trait with 0.25 heritability, ~50,000 individuals should afford 90% power to detect a recessive effect—as long as tests are conducted using recessive or genotype class definitions (Extended Data Figure 2.1).

To look for non-additive effects in the New Zealand dairy cattle population, we developed a two-step genetic association model that addresses population stratification while estimating dominance and/or additive effects for variants at sequence resolution (Methods). We initially focused our analyses on body weight, stature and body condition score (a visual estimate of body fat content), as growth and developmental traits that might be assumed to capture a range of genetic disorders and impacts on animal fitness (Table 2.1 shows SNP-based heritabilities for these traits). Using an imputed whole-genome sequence dataset comprising 16,129,957 variants, dominance and additive association analyses were then performed using 80,027 mixed-breed animals. Figure 2.1 contrasts genome-wide Manhattan plots from these models. Notably, these analyses highlighted eight non-additive quantitative trait loci (QTL) represented by 4,680 and 1,680 significant variants for body weight and stature, respectively ($P < 5 \times 10^{-8}$). Seven of these QTL were not represented by our standard-additive model (Extended Data Figure 2.2), the exception being a partial dominance effect for the well-described *PLAG1* locus (Karim et al. 2011). A recessive QTL mapping adjacent to the *ARSI* gene similarly presented a locus previously implicated in bovine stature (Cai et al. 2019), although the remaining six QTL appeared to be novel—demonstrating recessive impacts for which five of the signals presented compelling candidate causative mutations (Table 2.1 and Figure. 2.2). In the case of the chr29:44Mb QTL, this locus presented two strong candidates locating to *DPF2* and *MUS81*, with the remainder of loci represented by individual nonsense mutations (*FGD4* and *GALNT2*) or conserved amino acid substitutions (*PLCD4* and *MTRF1*). In all cases, these candidates were the top or near-top ($R^2$ = 0.94–1.0) associated variants for each trait QTL (Figure 2.2 and Table 2.1). Assessing functional annotations of all statistically plausible causative variants for all eight non-additive body weight loci ($R^2 > 0.9$ with the top associated variant; $N$ = 356 variants), the highlighted coding mutations represented a 5-fold enrichment of nonsense and missense variant classes genome-wide ($P$ = 0.001; permutation test; Supplementary Table 1). Significant (yet lower-fold) enrichment of protein-altering variant classes was also observed when considering the 3,926 QTL-linked variants identified through standard-additive GWAS (2-fold enrichment, empirical $P$ = 0.0001; Supplementary Table 1). Not precluding the involvement of regulatory or unidentified structural variants at these loci, these findings suggested that the non-additive candidate mutations were enriched for true positive causative variants. Notably, although no significant non-additive QTL were yielded for body condition score, the top associated genome-wide signal ($P = 7.6 \times 10_{-8}$) presented a novel 78-bp deletion–insertion frameshift of *MYH1* (Extended Data Figure 2.3), a myofilament gene recently implicated in a muscular atrophy disorder in horses (Finno et al. 2018).

Of all signals identified, we were particularly interested in the QTL at the *PLCD4*, *FGD4*, *MTRF1*, *GALNT2*, *DPF2*/*MUS81* and *MYH1* loci, since the functional candidacy of the mutations, and the sign, magnitude and recessive modes of effect highlighted these as potentially representing novel bovine syndromes. Supplementary Table 2 shows detailed molecular genetic descriptions of these candidates. Mutant alleles appeared to be breed specific, and when considered together, cumulative carrier frequencies of these variants were remarkably high, with >40% of purebred Holstein-Friesians and Jerseys carrying at least one mutant (non-ancestral) allele. Extrapolation of these frequencies together with mating records of genotyped sires suggested that ~1% of all New Zealand dairy animals born over the past 10 years would have been homozygous for one or more of the mutations, equating to ~9,700 of the ~940,100 females born annually (Supplementary Table 3 shows mutation-specific estimates).

**Table 2.1 | Association statistics for candidate mutations at recessive loci.**

| Phenotype<br>HF - ($h^2$, $\delta^2$)†<br>J - ($h^2$, $\delta^2$)† | Recessive QTL | Chr2:23Mbp | Chr2:107Mbp | Chr5:78Mbp | Chr7:64Mbp | Chr12:11Mbp | Chr28:1Mbp | Chr29:44Mbp | Chr29:44Mbp | *Chr19:30Mbp* # |
|---|---|---|---|---|---|---|---|---|---|---|
| | Candidate mutation | - | AC_000159.1<br>g.107313998G>A | AC_000162.1<br>g.77632752C>T | AC_000164.1<br>g.63649071C>T | AC_000169.1<br>g.11463981G>A | AC_000185.1<br>g.1312334G>A | AC_000186.1<br>g.44213160A>G | AC_000186.1<br>g.44645469G>T | AC_000176.1<br>g.30114250_30114327<br>delins‡ |
| | Candidate gene | - | PLCD4 | FGD4 | ARSI | MTRF1 | GALNT2 | DPF2 | MUS81 | MYH1 |
| | Predicted consequence | - | Amino-acid sub.<br>NP_001039954.1<br>p.Ala326Thr | Splice donor mut.<br>XM_005206883.3<br>c.1671+1G>A | Amino-acid sub.<br>XP_002689338.1<br>p.Gly301Ser | Amino-acid sub.<br>NP_001030185.1<br>p.Arg341Trp | Splice acceptor mut.<br>NM_001193103.1<br>c.1561-1G>A | Amino-acid sub.<br>NP_001093826.1<br>p.Lys216Arg | Premature stop<br>XP_005227165.2<br>p.Gly70* | Frameshift mut.<br>NP_776542.1<br>p.(Thr1698fs) |
| | N animals physically genotyped for mutation (of total 80,027) | - | 25059 | 25239 | 45 | 25235 | 32442 | 45 | 25231 | 40§ |
| | *AAF (HF, J, ALL) | 0.124,0.002,0.074 | 0.039,0.001,0.024 | 0.039,0.001,0.024 | 0.124,0.002,0.072 | 0.002,0.114,0.045 | 0.055,0.001,0.033 | 0.066,0.001,0.039 | 0.066,0.001,0.04 | 0.006,0.109,0.042 |
| Bodyweight<br>*(0.390, 0.038)*<br>*(0.329,0.059)* | Het Effect +- SE (Kg) | -0.007 ± 0.60 | -0.354 ± 0.879 | 0.834 ± 0.913 | 0.564 ± 0.705 | 1.078 ± 0.697 | -0.837 ± 0.7 | 4.059 ± 0.788 | 4.126 ± 0.816 | *0.798 ± 0.702* |
| | Het P-value | 0.990 | 0.687 | 0.361 | 0.424 | 0.122 | 0.232 | 2.63E-07 | 4.34E-07 | *0.256* |
| | Homo-Alt Effect +- SE (Kg) | -11.8 ± 1.99 | -108.965 ± 8.96 | -48.079 ± 7.641 | -11.981 ± 2.188 | -19.464 ± 3.321 | -112.008 ± 7.272 | -53.092 ± 3.798 | -51.911 ± 4.064 | *-4.643 ± 2.936* |
| | Homo-Alt P-value | 4.0E-09 | 5.03E-34 | 3.13E-10 | 4.37E-08 | 4.59E-09 | 1.55E-53 | 2.13E-44 | 2.34E-37 | *0.114* |
| | Top variant GWAS | 22836659 | 107313998 | 77632752 | 62891843 | 11463981 | 680910 | 44143856 | 44143856 | *27632220* |
| | Mutation R2 with Top Var | - | 1 | 1 | 0.883 | 1 | 0.944 | 0.996 | 0.976 | *0.002* |
| Stature<br>*(0.341,0.049)*<br>*(0.264,0.054)* | Het Effect +- SE (cm) | *0.025 ± 0.045* | -0.025 ± 0.069 | *0.058 ± 0.073* | 0.055 ± 0.052 | *0.025 ± 0.054* | -0.134 ± 0.055 | 0.132 ± 0.06 | 0.136 ± 0.063 | *0.086 ± 0.054* |
| | Het P-value | *0.570* | 0.714 | *0.429* | 0.286 | *0.645* | 0.015 | 0.029 | 0.030 | *0.112* |
| | Homo-Alt Effect +- SE (cm) | *-0.428 ± 0.157* | -5.421 ± 0.668 | *-3.041 ± 0.609* | -1.152 ± 0.168 | *-1.208 ± 0.263* | -8.568 ± 0.621 | -2.565 ± 0.322 | -2.487 ± 0.318 | *0.285 ± 0.241* |
| | Homo-Alt P-value | *6.3E-03* | 4.90E-16 | *6.00E-07* | 6.11E-12 | *4.52E-06* | 2.98E-43 | 1.82E-15 | 5.19E-15 | *0.237* |
| | Top variant GWAS | *21853244* | 107313998 | *77632752* | 63315153 | *11415931* | 1345436 | 44143856 | 44143856 | *30350825* |
| | Mutation R2 with Top Var | - | 1 | *1* | 0.909 | *0.97* | 0.975 | 0.996 | 0.976 | *0.01* |
| Body cond.<br>Score<br>*(0.249,0.018)*<br>*(0.226,0.016)* | Het Effect +- SE (score) | *0.003 ± 0.004* | *0.005 ± 0.006* | *0.012 ± 0.006* | *-0.001 ± 0.005* | *0.009 ± 0.005* | *0.012 ± 0.005* | *0.02 ± 0.006* | *0.02 ± 0.005* | -0.01 ± 0.005 |
| | Het P-value | *0.480* | *0.458* | *0.059* | *0.762* | *0.050* | *0.017* | *5.02E-04* | *2.59E-04* | 0.076 |
| | Homo-Alt Effect +- SE (score) | *-0.044 ± 0.014* | *-0.284 ± 0.063* | *-0.238 ± 0.055* | *-0.028 ± 0.015* | *-0.093 ± 0.025* | *-0.079 ± 0.055* | *-0.05 ± 0.031* | *-0.054 ± 0.03* | -0.133 ± 0.023 |
| | Homo-Alt P-value | *2.1E-03* | *6.03E-06* | *1.51E-05* | *0.060* | *1.85E-04* | *0.156* | *0.115* | *0.073* | 1.22E-08 |
| | Top variant GWAS | *24193486* | *107622778* | *77632752* | *65424545* | *11732788* | *368171* | *46665197* | *46665197* | 29819114 |
| | Mutation R2 with Top Var | - | *0.003* | *1* | *0.0003* | *0.892* | *0.015* | *0.546* | *0.55* | 0.906 |

Non-significant QTL are in italics. N genome-wide significant variants: Bodyweight, 4,680; Stature, 1,860; Body condition score, 0. The top variant GWAS positions were considered to be within same QTL if within 5Mb window centred on candidate causal variant. Effects are shown only for recessive loci, with the *PLAG1* locus (chr14:25Mbp) omitted due to displaying incomplete dominance of an otherwise well described locus. AAF, Alternate allele frequency; HF, Holstein-Friesian; J, Jersey. †SNP-based additive and dominance heritabilities calculated within breed (HF, N=12149; J, N=7502 cows), using a quality-filtered BovineSNP50k genotype dataset. h2, additive heritability; δ2, dominance heritability. #Variant nonsignificant in primary GWAS. ‡Truncated for brevity, full description is AC_000176.1:g.30114250_30114327delinsTGTTATGCTGTTATGTTATGT. §Note imputed genotypes based on left flank of the mutation (since full length indel interpreted by GATK as g.30114250_30114258 del).

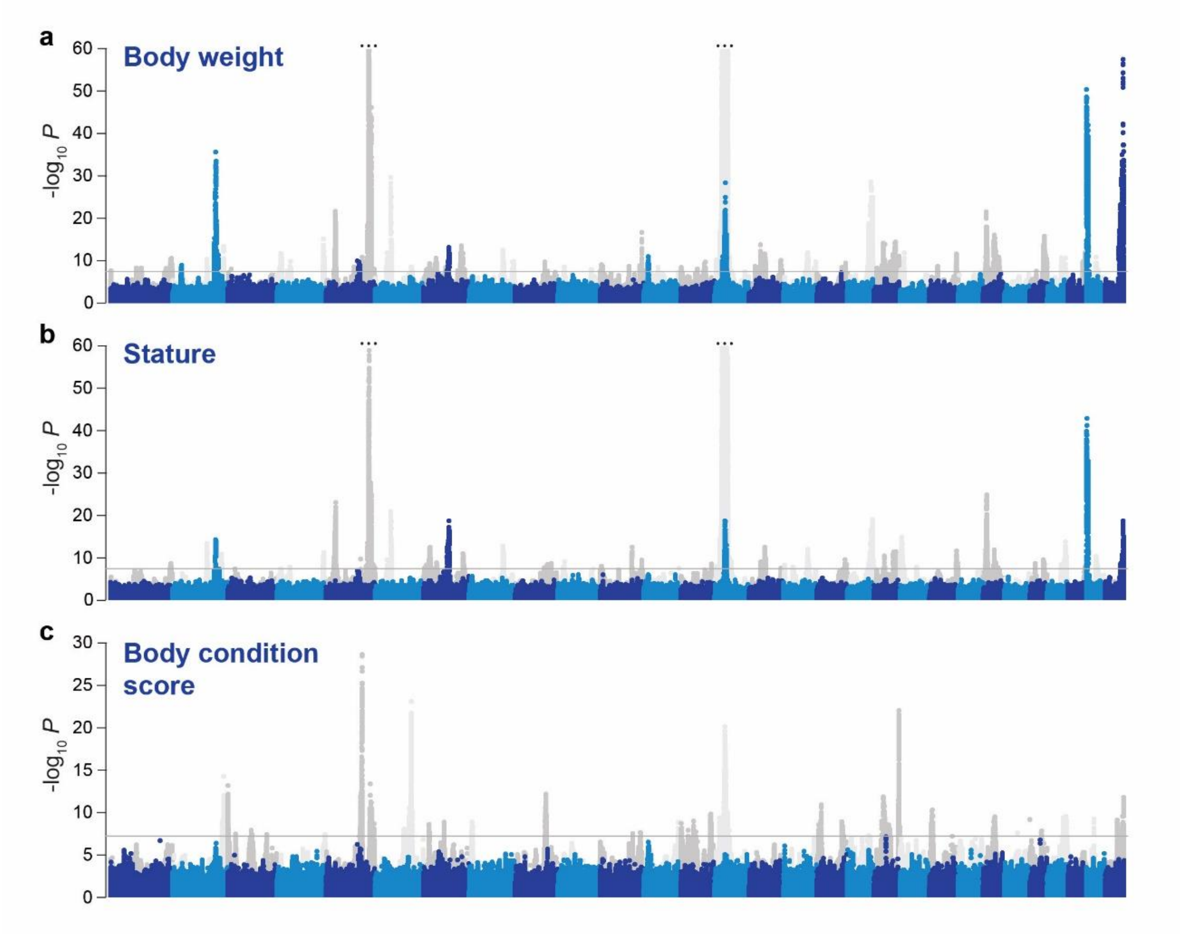


**Figure 2.1 | Dominance and standard additive Manhattan plots for body weight, stature and body condition score.**

a–c, Significance of dominance (blue and light blue) and standard-additive (gray and light gray) estimates for ~16 million imputed sequence variants for body weight (a), stature (b) and body condition score (c) traits; alternating colors are used to demarcate chromosomes. Strength of association is calculated using a Z test with values shown on the y axis; a threshold indicating regions that surpass the multiple testing correction threshold of P = 5 × 10−8 is also indicated (horizontal gray line). Dominance effects were estimated from models that also included additive components, although the additive estimates shown were derived from a separate, standard-additive model (as per the section entitled GWAS in the Methods). For body weight and stature additive estimates, the y axes were truncated for display purposes (indicated by three dots; smallest body weight P = 3.59 × 10−127 and P = 4.97 × 10−442 for chr5 and chr14 respectively, and smallest stature P = 3.23 × 10−97 and 2.0 × 10−306, respectively).

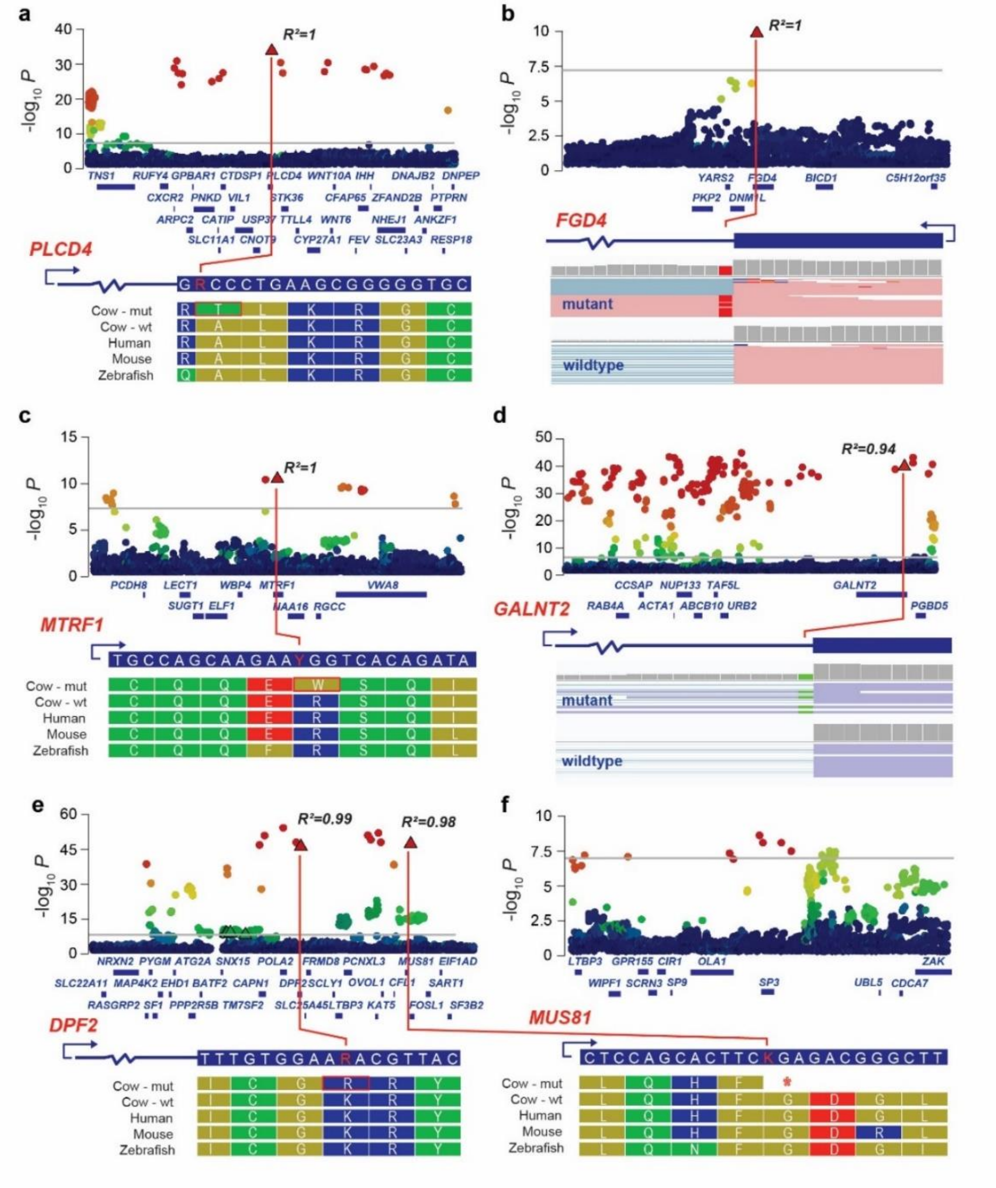


**Figure 2.2 | Regional Manhattan plots for the six novel recessive body weight QTL.**

a–f, 1.5-Mb Manhattan plots showing body weight QTL representing the chr2:107Mb (a), chr5:78Mb (b), chr12:11Mb (c), chr28:1Mb (d), chr29:44Mb (e) and chr2:22Mb (f) loci. P values are shown on the y axis (calculated using a Z test), with the genome-wide significance threshold indicated by the horizontal gray line in each plot. mut, mutant; wt, wild type. Variants are colored by linkage disequilibrium R2 values with the top tag variant per locus; protein-coding variants are shown as triangles. For missense candidates (a,c,e), multispecies protein sequence alignments are shown (residues colored by polarity). For nonsense candidates, mammary RNA sequence alignments show loss of splicing efficiency (b,d), or predicted truncation due to premature termination (*first deleted residue; e).

To explore the diversity of other phenotypes that might be affected, we tested the impacts of the candidate mutations on 16 additional animal performance traits using pedigree-based mixed linear models (Methods). These analyses revealed a variety of other major effects (Supplementary Table 4), with lactation and anatomical impacts generally exceeding those of two of the largest, most well-recognized additive QTL in the cattle literature (*DGAT1* (Grisart et al. 2002) and *PLAG1* (Karim et al. 2011); Figure 2.3 and Extended Data Figure 2.2). While the location of the *FGD4* mutation adjacent to a major milk composition locus (Lopdell et al. 2019) suggested linkage disequilibrium as a possible explanation for some of these associations, representative genotypes of these loci were in weak linkage disequilibrium (maximum $R^2 = 0.018$), and none of the other signals mapped near major known additive loci in our population. Acknowledging genetic correlation between many of these traits, these findings suggested broad pleiotropy for most of the variants, with the size and sign of effects further supporting their status as having pathogenic impacts.

Given their recessive modes of effect, we wondered whether the highlighted loci might underpin inbreeding depression effects on body weight, estimated as −0.77 kg per 1% of inbreeding in the GWAS population. Here, mixed models that included inbreeding coefficients and either did or did not include genotype classes of the eight non-additive QTL were compared (Methods). These collective loci were found to account for 23.5% of the variance otherwise attributable to inbreeding depression, where by contrast, the variants explained 0.7% when fitted as standard-additive effects. Further, most of this effect did not appear to be due to tagging of homozygosity per se, since random, permuted samples of MAF-matched variants assessed as non-additive effects explained an average of 1.5% of inbreeding-attributable variance (range 0–3.4% from 20 samples).

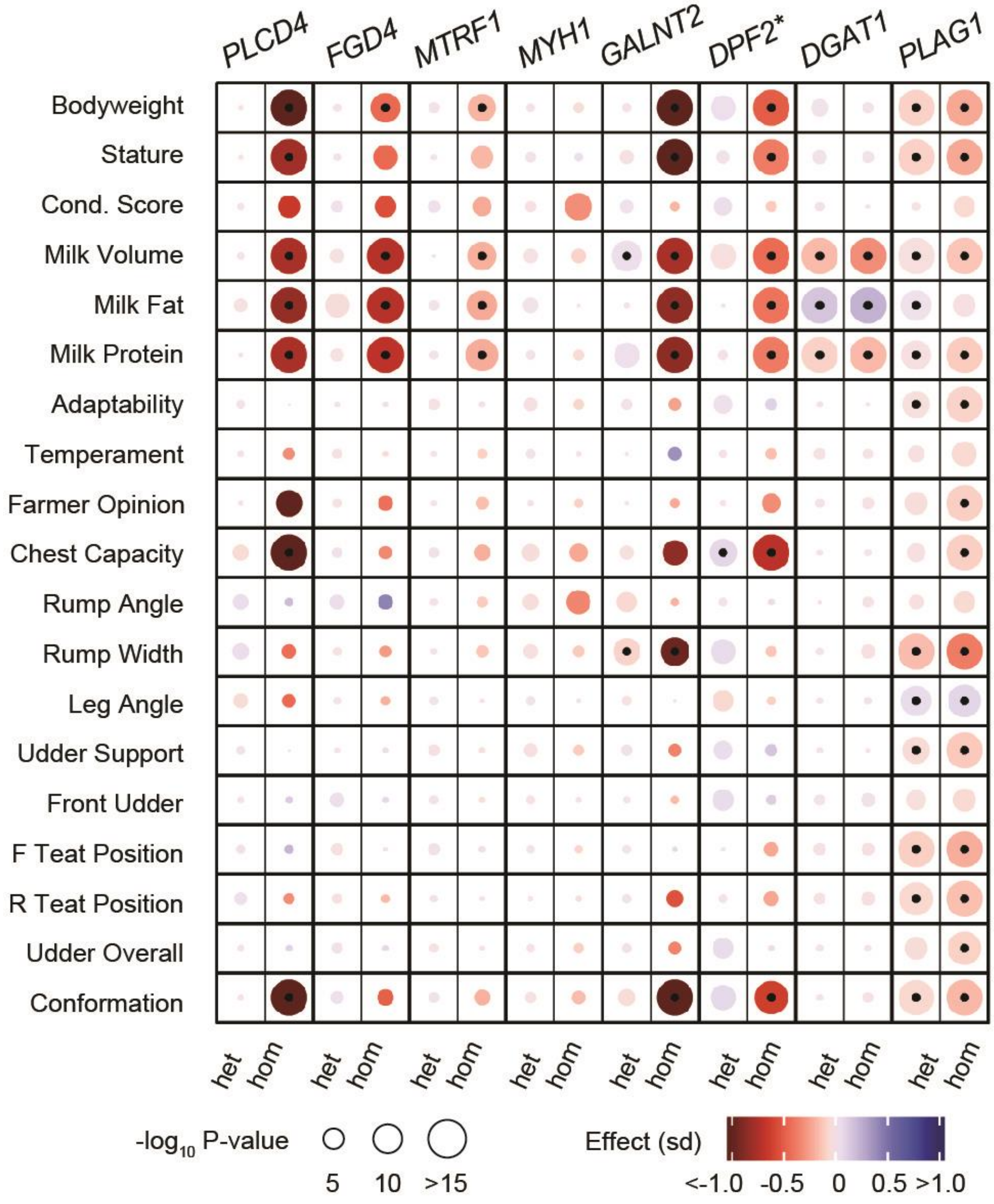


**Figure 2.3 | Heat map showing a diversity of effects for recessive mutations of interest.** Phenotypic effects of the six novel recessive candidate mutations are shown, represented alongside effects for the well-documented, major-effect *DGAT1* and *PLAG1* additive mutations for contrast (right-most columns). For a given gene, columns showing heterozygote and alternative allele homozygote effects are indicated, where circle color denotes effect sign, color intensity denotes effect magnitude, and circle size indicates strength of association (P values are derived from two-sided t-tests; circles containing solid black dots indicate significant association where $P < 5 \times 10^{-8}$). *Note that the chr29:44Mb QTL that presented two tightly linked candidate mutations in *DPF2* and *MUS81* is represented by the *DPF2* mutation alone.

Within-breed representation of the recessive variant genotype classes suggested that homozygote mutant individuals were underrepresented as mature animals for a subset of loci, observations supported by analysis of transmission ratios within carrier crosses (Supplementary Table 5). Since most animals are genotyped in young adulthood, these findings suggested a survival disadvantage for some of the mutations either in utero, or at some other stage of rearing before genotype ascertainment (for example, farmer-directed culling). Two additional populations were then used to validate the effects of the variants. First, a retrospective analysis of milk traits from 31,580 animals was performed using the same single-locus association models applied above. While the utility of this dataset was limited by the number of phenotypes interrogated (that is, the body weight, stature and body condition score traits used for GWAS were not measured in these animals), all loci demonstrated lactation effects of the same sign and similar magnitude to those highlighted in the discovery population (Supplementary Table 6). We next conducted a prospective analysis of juvenile animals—performed to address potential effect size underestimation due to loss of affected animals in early life. Here, pedigree data were used to identify animals <1 year old whose sires and maternal grandsires were both carriers for one or more of the *PLCD4*, *DPF2*/*MUS81* and *FGD4* mutations (prioritized due to effect sizes and frequency). Association testing in this cohort of 568 calves identified significant effects ($P < 0.05$) for the *PLCD4* and *FGD4* variants, with the *DPF2*/*MUS81* effect nonsignificant in juvenile animals (Supplementary Table 7).

To understand the largest-effect syndromes in detail, we conducted more detailed molecular and physiological analyses of the *PLCD4*, *FGD4*, *GALNT2* and *DPF2*/*MUS81* mutations. First, we assessed the expression consequences of the nonsense variants using a large, pre-existing (Littlejohn et al. 2016) mammary RNA-sequencing (RNA-seq) dataset.

Transcripts bearing the *GALNT2*, *FGD4* and *MUS81* mutations did not appear to be subject to nonsense-mediated RNA decay, where expression QTL (eQTL) analysis suggested increased expression of *FGD4* and *MUS81* in carrier animals (Extended Data Figure 2.4). Splicing-based eQTL analysis confirmed loss of splice fidelity for the *GALNT2* and *FGD4* variants, however, with mutant transcripts showing intron retention and activation of cryptic splice sites (Extended Data Figures 2.4 and 2.5). More detailed analyses were then performed on animals homozygous for these mutations, conducted on a research farm utilizing control animals broadly matched for age, sire and breed. Supplementary Tables 8 and 9 detail anatomical and blood metabolic associations between mutants and controls. In humans, analogous mutations in *DPF2*, *FGD4* and *GALNT2* cause Coffin–Siris syndrome (Vasileiou et al. 2018), Charcot Marie Tooth syndrome (Delague et al. 2007; Stendel et al. 2007) and a congenital disorder of glycosylation (GALNT2-CDG), respectively. Peculiarities in hoof conformation were potentially suggestive of Coffin–Siris syndrome in some *DPF2* homozygotes (Extended Data Figure 2.6), although quantitative analysis did not show significant differences (Supplementary Table 8). In *FGD4* homozygotes, apparent loss of coordination, and histology showing lesions of axonal degeneration, Schwann cell hyperplasia and demyelination consistent with Charcot Marie Tooth syndrome in peripheral nerves (Extended Data Figure 2.7), confirmed a bovine form of that disorder. Likewise, significant reductions in circulating triglycerides, and the markedly reduced body weight and stature of *GALNT2* homozygotes (Extended Data Figure 2.8), are features common to most patients with GALNT2-CDG—as well as mouse and rat *Galnt2*-knockout models (Khetarpal et al. 2016; Zilmer et al. 2020). Liver transcriptomic analysis similarly suggested perturbed lipid metabolism in *GALNT2*-homozygous calves, showing a >10-fold increase in expression of transcripts encoding the gluco- and lipo-regulatory hormone FGF21 (Ge et al. 2012) ($P$ = 3.9 × $10^{-7}$; Supplementary Table 10). Further discussion and presentation of results pertaining to the *DPF2*/*MUS81*, *FGD4* and *GALNT2* QTL in these human disease contexts is given in the Supplementary Note.

Although no clear pathogenic effects have been ascribed to mutations in the *PLCD4* gene in humans, it is noteworthy that *PLCD4* homozygotes displayed the most striking anatomical impacts, showing ~100-kg reductions in body weight, and abnormal body conformation (Extended Data Figure 2.8 and Supplementary Tables 4 and 8). Associations for stature and body composition traits have been highlighted near human (Buniello et al. 2019) and ovine (Bolormaa et al. 2016) *PLCD4*, although mice homozygous for *Plcd4*-null mutations appear anatomically normal (Fukami et al. 2001)— suggesting that additional work will be required to definitively establish the causality of the p.Ala326Thr alteration at this locus. Although we did not investigate *MTRF1*-mutant calves, we note that this gene similarly has no human disease implication—nor knockout mouse model. These findings suggest that p.Arg341Trp represents one of the first deleterious MTRF1 mutations reported for mammalian species, acknowledging the comparatively modest effects attributed to this mutation.

We note that several recent non-additive GWAS published in humans (Zhu et al. 2015) and cattle (Bolormaa et al. 2015; Jiang et al. 2017; 2019) have presented significant loci, although contrary to the findings reported here, major recessive effects were not highlighted in these studies. In a human context, this contrast can be reconciled by a comparatively high $N_e$ and consequent low average allele frequency of deleterious mutation (Charlier et al. 2016; Lek et al. 2014), where compared to cattle, fewer rare allele homozygotes will be presented for most mutations. While the largest previous non-additive cattle GWAS used a population size far exceeding that reported here (*N* = 294,000 cows (Jiang et al. 2019)), this analysis was conducted using medium-density SNP-chip genotypes (~57,000 variants filtered at MAF > 5%). This suggests low-MAF recessive mutations may have been incompletely captured in this study, and indeed, when visualizing our body weight GWAS results at BovineSNP50k resolution (~46,000 variants), only one of the major-effect recessive loci remains represented (*DPF2*/*MUS81* locus; Extended Data Figure 2.9). These observations demonstrate the importance of MAF to sensitivity of discovery, so to investigate these influences more broadly, we further simulated a population of 80,000 animals harbouring recessive mutations of variable effect size, segregating at 1–5% MAF (Methods). Association analysis of these data shows even the largest-effect mutations are no longer discernible at ~1% MAF (Extended Data Figure 2.9), highlighting the lower bound of detection for population sizes similar to that investigated here. These findings also represent the best-case scenario, since no account of genotype or imputation error is made. We (and others (Pausch et al. 2017)) observe that imputation accuracy declines markedly at frequencies approaching 1% in cattle (Extended Data Figure 2.10), and although the mutations highlighted here segregate at 2–5% MAF and appear to have been accurately imputed (imputation allelic $R^2$ (AR2) metrics of >0.95), even higher MAF causal variants that do not phase or impute well are likely to be obscured from detection. Expanded imputation reference datasets and physical genotyping of functionally prioritized candidates are thus likely to play key roles in future discovery, in addition to the increased sample sizes required to represent homozygotes for these alleles.

In summary, we highlight the existence of non-additive deleterious alleles segregating at marked frequencies in cattle, where we detail six novel putative causative mutations with effects ranging from mild (3.5% reduction in body weight) to severe (>25% reduction in body weight and increased early-life mortality). These discoveries demonstrate the use of proxy phenotypes to directly map deleterious effects in the absence of prior disease identification, an approach that holds promise for the identification of similar effects in other selected species. Importantly, these results create new opportunities to improve the health and welfare of animals, where genetic screening and a broader awareness of syndromes will allow breeding and management strategies to minimize genetic disease.

## 2.3 Methods

### 2.3.1 Animal populations.

Supplementary Table 11 summarizes the different animal populations, their demographic details and respective analyses performed in this study. The 'Animal populations' description here outlines five of these populations in particular, for which the most detailed studies were performed (*N* = 130,725 animals total, though excluding the RNA sequence population that is described further in that section). Animals within these five key datasets mostly consisted of commercially farmed, outbred cows that had been genotyped and phenotyped as part of commercial livestock improvement activities. These five datasets were as follows: (1) the 'discovery population' used for initial GWAS, (2) a 'validation population' used for point-wise association tests of putative causative mutations, (3) a 'homozygous depletion' population used for analysis of genotype representation statistics and carrier cross outcomes, (4) a group of 'prospectively genotyped calves' and (5) a group largely overlapping with (4) that was subjected to detailed phenotypic analysis at a research farm (the 'research farm study').

The discovery population comprised a total of 80,027 cows, representing the union of samples for which body weight (*N* = 79,945), stature (*N* = 75,041) or body condition score (*N* = 75,617) phenotypes and imputed sequence genotypes were available. The validation population comprised 31,580 animals that had also been genotyped and imputed to genome-sequence-resolution data, and for which milk fat, protein and volume yield data were available. The homozygous depletion population comprised a total of 15,379 bulls and 35,790 cows. Two analyses were performed to investigate the relative representation of candidate causative mutation classes in these animals, with homozygous depletion statistics calculated using 49,115 purebreds (defined as 16/16$^{th}$s breed proportion based on pedigree), and carrier cross trio analysis conducted using an overlapping set of 4,276 animals. These animals overlapped with the discovery and validation populations, with the main difference being that this population also included males, and some SNP-chip-genotyped animals for which phenotypic records were not available.

The prospectively genotyped calves study represented 568 calves that were identified as candidate homozygous affected animals for one of the *PLCD4*, *FGD4* or *DPF2*/*MUS81* recessive effects. These candidates were selected through analysis of pedigree records to identify animals less than 1 year old whose sires and maternal grandsires were both carriers for one of the recessive alleles. Assuming a 20% error in recorded parentage assignment, these calves were expected to present a 1 in ~15 probability of being homozygous for one of the alleles of interest. The research farm study comprised a subset of 34 of these prospectively genotyped calves, and an additional 12 young animals assessed as part of a 2013/2014 investigation of the *GALNT2* mutation. These 46 animals were subjected to detailed molecular and physiological analyses (see the section below entitled Phenotypic analysis, and the Supplementary Note). For simplicity of communication of results and methods, these two deeply phenotyped groups of animals are considered and presented together in the manuscript.

### 2.3.2 Phenotypic analysis.

The phenotypes used for the GWAS in the discovery population consisted of measured body weight (also referred to as live weight), measured stature or subjectively assessed body condition score. Lactation traits, and a range of other, additional phenotypes, were also investigated in the discovery population using single-locus models. Definitions of these phenotypes are provided in the Supplementary Methods, and further in the 'Evaluation system for traits other than production' booklet (Advisory Committee on Traits Other than Production 2020). Mixed linear models were used to fit class variables and covariates for all phenotypes before genetic analysis. These models were applied to the entire cattle population for national genetic evaluation, and differed slightly by trait, based on the context and properties of each phenotype. Broadly, they included fixed effects for cohort or contemporary group, age (in days) at calving and pairwise heterosis between breed designations (Holstein, Friesian, Jersey, Ayrshire and other). Further trait-specific model parameters are detailed in the Supplementary Methods.

The research farm studies were investigated as part of two, more detailed, prospective phenotypic analyses. The group of 34 calves consisted of 9 control animals, and affected (homozygous) animals representing either the *PLCD4* (*N* = 8), *FGD4* (*N* = 9) or *DPF2*/*MUS81* (*N* = 8) mutation classes. Control animals were part of the same study and common to all three of the mutation classes, being broadly matched for age, sire and breed. The 12 animals separately assessed as part of the *GALNT2* study consisted of 6 affected individuals and 6 controls. Both of these studies were performed on an AgResearch Ruakura research farm in Hamilton, New Zealand, with control and affected animals grazed and managed together. Supplementary Tables 5 and 6 detail the range of phenotypes collectively measured on these animals, and a description of the more involved analyses (including blood biochemistry, anatomical measures and histological procedures) is provided in the Supplementary Note.

### 2.3.3 Sequencing, mapping and sequence informatics.

Whole-genome sequencing of the 565 animals that formed the reference population for sequence imputation was performed as previously described (Littlejohn, Henty, et al. 2014; Littlejohn et al. 2016). Briefly, animals of HF, J or HF × J breeds were sequenced on an Illumina HiSeq 2000 instrument targeting 100-bp paired-end reads. Genome sequence data were aligned to the UMD3.1 genome assembly using BWA MEM 0.7.8 (Li 2013), yielding mean and median mapped read depths of 15× and 8×, respectively. Variant calling was conducted using the GATK HaplotypeCaller (version 3.2) (DePristo et al. 2011) with base quality score recalibration applied.

To quantitatively assess the splicing and gene expression consequences of the *FGD4*, *GALNT2* and *MUS81* nonsense variants, analyses were performed using a large, pre-existing RNA-seq dataset based on mammary biopsy of 389 lactating cows. Details regarding the animals and sequencing methods underpinning this dataset have been described previously(Littlejohn et al. 2016). Additional data processing relevant to the current manuscript is described in the Supplementary Note. It should be noted that unlike all DNA-based analyses that referenced the UMD3.1 genome, RNA-seq-based data were mapped to the ARS-UCD1.2 assembly. Further clarification on how these two assemblies were utilized is given in the Supplementary Note.

To generate gene expression phenotypes for eQTL analysis, reads mapping to each gene were counted using the featureCounts module of Subread (version 1.5.3) (Liao, Smyth, and Shi 2014), and then normalized and transformed using the variance-stabilizing transformation function implemented in DESeq2 (Love, Huber, and Anders 2014). Splice efficiency phenotype generation was similar to that recently described (Fink et al. 2020), calculated as the number of reads that spliced or did not splice at a given intron–exon boundary, expressed as a proportion and transformed using the logit function. These criteria yielded per-junction estimates of splicing efficiency, with RefSeq annotation release 106 reference annotations XM_024992557.1 and NM_001193103.2 used to define gene structures for *FGD4* and *GALNT2*, respectively. Gene expression and splice phenotypes were then used to perform association analysis, as further described below, in the section entitled Single-locus models. For RNA sequencing of liver tissue, biopsy was performed using an established protocol (Lucy et al. 2009) targeting the 12 young animals investigated as part of analysis of the *GALNT2* mutation (6 affected, 6 controls). Further details regarding biopsy, RNA extraction and subsequent sequencing, read mapping and expression phenotype derivation are provided in the Supplementary Note.

### 2.3.4 DNA extraction, genotyping and imputation.

The majority of the study animals (*N* = 130,145) were genotyped using SNP chips, with DNA extraction performed using either ear-punch tissue samples or blood. Samples were processed by GeneMark using Qiagen BioSprint kits or by GeneSeek using a MagMAX system (Life Technologies). Genotyping was performed using one or more of a variety of platforms, including the Geneseek GGPv1, GGPv2, GGPv3, GGP50k, Illumina BovineSNP50k or BovineHD 777k SNP chips. For the 32,455 samples typed on the GGPv3 platform, 3,779 protein-altering variants were directly interrogated in these animals, where sequence-derived missense and nonsense variants had been included as custom content on that platform as part of a previous study (Charlier et al. 2016). A subset of the same custom content (*N* = 349 variants) was also typed on the GGP50k platform (*N* = 10,224 of 130,145 samples). Table 2.1 outlines the number of animals from the discovery population that were physically genotyped for the loci of primary interest, based on genotyping with these two panels. A description of assays used to target single mutations of interest is given in the Supplementary Note.

Imputation of genome sequence data in the study animals was performed using Beagle v4 (Browning and Browning 2009) and has recently been described in detail (Jivanji et al. 2019), although we also included steps specific to the current study. Specifically, in addition to the 19,320,540 sequence variants imputed using standard parameters, we reintroduced the physically genotyped, protein-altering GGPv3 variant sites (see above) that were otherwise lost due to an AR2 > 0.95 phasing quality threshold applied to the genome sequence reference. Before GWAS, an additional frequency filter was also applied to remove variants that did not have at least five individuals per genotype class (roughly equivalent to MAF < 0.01). Together, application of these steps yielded the final dataset of 16,129,957 variants used for association analysis. Extended Data Figure 2.10 shows AR2 distributions by MAF class for all variants used for GWAS, as well as the relationship between AR2 and other statistical parameters post-GWAS. The Supplementary Note elaborates on the baseline sequence imputation strategy and methods used to impute and filter the BovineSNP50k panel dataset that was used for population stratification adjustment.

### 2.3.5 Association analysis.

*GWAS.*

We applied a GWAS method to test the effects of all imputed sequence variants, one at a time, while simultaneously adjusting for genomic effects that lay outside the given genomic segment of interest. For heritability estimations performed before GWAS, see Supplementary Note.

*Overview.*

Conceptually this method involved fitting the model:

$$\mathbf{y} = \mathrm{T}\mathbf{b} + \mathrm{M}\_\boldsymbol{\alpha}\_ + \mathbf{e} \qquad (1)$$

where **y** is a vector of phenotypic deviations for one trait, **b** is a vector of genotype class effects for the sequence variant of interest, *T* is a design matrix relating records to genotype class for the sequence variant, $\boldsymbol{\alpha}\_$ is a vector of random SNP-chip additive marker effects spanning the whole genome except the segment of interest such that $\boldsymbol{\alpha}\_ \sim N(0, I\sigma_{\alpha}^{2})$, where *I* is an identity matrix of order equal to the number of marker effects and $\sigma_{\alpha}^{2}$ represents the marker effect variance, $M\_$ is a matrix obtained from *M* (a matrix relating records to SNP markers (encoded [0, 1, 2])), by deleting the columns corresponding to the region flanking the sequence variant, and **e** is an error vector with $\mathbf{e} \sim N(0, D)$, where for traits with single observations, $D = I\sigma_{e}^{2}$, *I* is an identity matrix of order equal to the number of phenotypic records and $\sigma_{e}^{2}$ represents the residual error variance. For traits represented by the mean of a variable number of repeated observations, the diagonal elements of *D* varied according to the number of observations in the mean.

This model was applied in a two-step method. First, we adjusted phenotypic deviations for population structure using marker effects ($M\_\boldsymbol{\alpha}\_$) (1), sampled using a BayesC0 algorithm based on the Markov chain Monte Carlo method (Fernando and Garrick 2013). We applied a leave-one-segment-out (LOSO) approach by adjusting phenotypic deviations for $M\_$ and $\boldsymbol{\alpha}\_$ instead of *M* and $\boldsymbol{\alpha}$, respectively. Second, each sequence variant was separately tested for association with its respective LOSO-adjusted phenotypic deviations using two association models, a standard-additive model and a dominance model. Further details on how these methods were applied are provided in the Supplementary Note.

*Standard-additive model.*

In the standard-additive model, **b** is a vector of an intercept and an additive effect ($\beta$), and $T$ is a design matrix relating records to genotype for the sequence variant. We obtained a vector of plausible intercept and standard-additive genotype effects for the adjusted phenotypic deviations, and made inference on the standard-additive effect of each sequence variant.

*Dominance model.*

In the dominance model, we fit genotypic additive ($a$, covariate encoded [0, 1, 2]) and dominance ($d$, covariate encoded [0, 1, 0]) effects. We chose a model formulation to ease the interpretability of results and enable recessive, over-dominant and partial dominance effects to be readily differentiated, an approach that differs from other parameterizations (Zhu et al. 2015; Jiang et al. 2019) (see Supplementary Note for further considerations on model contrasts with prior studies). For each sequence variant, a vector of plausible samples of the additive ($\tilde{a}$) and dominance ($\tilde{d}$) effects was computed from the samples of genotype class effects ($\tilde{b}$) via the second and third elements of the vector $T^{-1}\tilde{b}$, where

$$T = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 0 \end{bmatrix}.$$

*Summarizing results.*

The respective plausible distributions of genotype effects for each imputed sequence variant, $x$, were then summarized by their posterior means, $\beta_x$, posterior standard deviations, $\sigma_x$, and $z$ statistics, $z_x$, following a standard normal distribution as in Bernal Rubio et al. (2015).

$$z_x = \frac{\beta_x}{\sigma_x}, where\ z_x \sim N(0, 1) \qquad (2)$$

For each variant, the statistical significance of additive and dominance genetic effects was evaluated using a $Z$ test. For GWAS analyses, we considered variants to be statistically significant using a multiple testing threshold of significance at $P = 5 \times 10^{-8}$. To identify individual QTL from our standard-additive model and permit effect size comparison with variants highlighted from dominance GWAS, we used GCTA-COJO (Yang et al. 2012), similarly selecting variants at $P < 5 \times 10^{-8}$.

### 2.3.6 Single-locus models.

We used a pedigree-based mixed linear model for estimation of single-locus effects on additional traits including phenotypic deviations, gene expression levels and body weights from prospectively genotyped calves and animals in the research farm study. This model incorporated repeated measures, where applicable, and consisted of the following equation:

$$\mathbf{y} = \mathrm{X}\mathbf{b} + \mathrm{Z}\mathbf{u} + \mathrm{W}\mathbf{p} + \mathbf{e} \qquad (3)$$

where **y** is a vector of phenotypic deviations, gene expression levels or body weights, **b** is a vector of fixed effects for the variant of interest, $X$ is a design matrix relating records to respective fixed effects, **u** is a vector of random breeding value effects such that $\mathbf{u} \sim N(0, A\sigma_u^2)$, where $\sigma_u^2$ represents the additive genetic variance and $A$ is the additive relationship matrix conditional on the pedigree, $Z$ is a design matrix relating records to breeding values, **p** is a vector of random permanent environment effects such that $\mathbf{p} \sim N(0, I\sigma_p^2)$, where $\sigma_p{}^2$ represents the permanent environment variance and $I$ is an identity matrix of order equal to the number of phenotypic records, $W$ is a design matrix relating records to permanent environment effects and **e** is a random error term, where **e** ~ $N(0, I\sigma_e{}^2)$, where $\sigma_e{}^2$ represents the residual error variance. The random variables **u**, **p** and **e** are assumed to be uncorrelated. For further details regarding specific model characteristics and implementations for each phenotype, see Supplementary Note.

### 2.3.7 Enrichment of functional categories.

We aimed to test for the enrichment of annotation categories for sets of statistically plausible causative variants. For non-additive QTL, a set of 356 variants was tested based on a linkage disequilibrium threshold that selected all variants with $R^2 > 0.9$ with the top associated variant at each body weight locus. SnpEff 4.3 (Cingolani et al. 2012) was used to determine the predicted functional effect of the sequence variants based on *Bos taurus* annotations from Ensembl release 86 for the UMD 3.1 assembly. We then randomly selected 10,000 samples of MAF-matched variants to perform a permutation test and estimate whether the observed proportion of plausible causative variants with a predicted nonsense or missense annotation was greater than expected by chance. This same process was repeated for analysis of 3,926 additive QTL tag variants; additional details regarding variant selection, permutation and other annotation-related analyses conducted in the paper are given in the Supplementary Note.

### 2.3.8 Analysis of homozygote frequencies and depletion.

To estimate the number of historical animals that would have been affected by one or more of the key recessive mutations of interest, we used pedigree records in conjunction with sire genotypes to calculate per-individual homozygote probabilities. These estimates leveraged genotype information from 5,550 sires to highlight 2,799,022 historical animals for analysis, with probabilities calculated using the equation presented in the Supplementary Note. Homozygous depletion was assessed for each candidate causative mutation within the purebred population of its respective breed of origin, applying a standard likelihood ratio test as previously reported (Charlier et al. 2016) and further described in the Supplementary Note. Carrier cross genotype proportions were also assessed for each candidate causative mutation, using a goodness-of-fit test that aimed to detect loss of individuals from the homozygous alternative genotype class (Supplementary Table 5).

### 2.3.9 Inbreeding analysis.

To assess whether the non-additive QTL might explain some of the variance in inbreeding depression of body weight, we first calculated inbreeding coefficients using all animals in the GWAS discovery population (Supplementary Note). Following removal of animals with inbreeding coefficients of zero (to avoid inclusion of animals with inadequate pedigree information— *N* = 68,578 cows remaining), we then formulated and compared the following two models:

$$\mathbf{y} = 1\mu + \mathbf{F}b_1^R + \mathrm{M}\boldsymbol{\alpha} + \mathbf{e} \qquad (4)$$

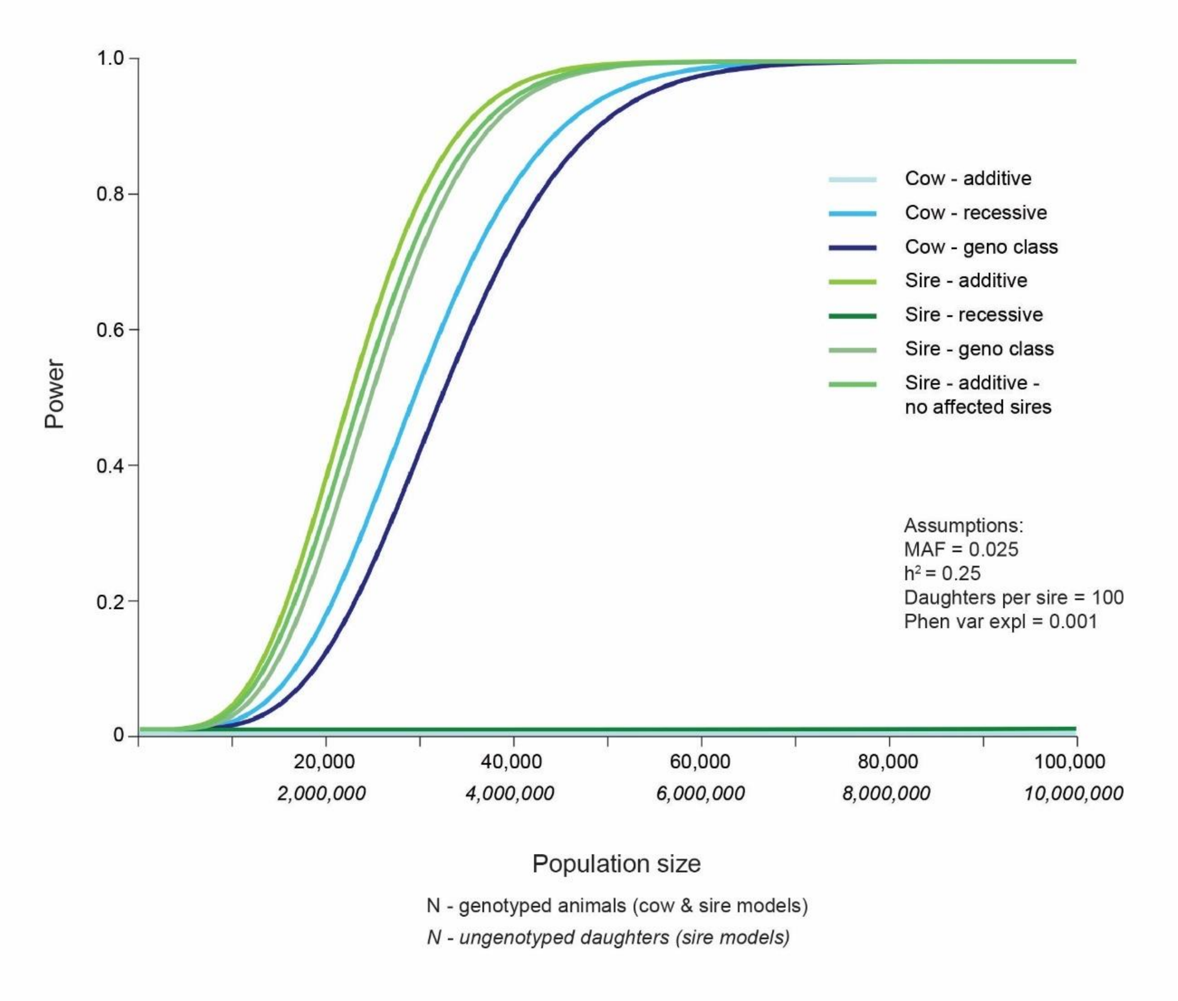
$$\mathbf{y} = 1\mu + \mathbf{F}b_1^F + \mathrm{X}\boldsymbol{b_2} + \mathrm{M}\boldsymbol{\alpha} + \mathbf{e} \qquad (5)$$

Terms are as described in the GWAS section above (1) with the following additions: **y** is a vector of phenotypic deviations for body weight (pre-adjusted for effects described in the section entitled Phenotypic analysis including pairwise heterosis), $\mu$ is the fixed effect representing the overall mean, $b_1^R$ and $b_1^F$ are the regression coefficients of body weight on inbreeding representing the reduced and full models respectively, **F** is a vector of pedigree-derived inbreeding coefficients, $\mathbf{b_2}$ is a vector of genotype class effects for the recessive QTL and $X$ is a design matrix relating records to additive and dominance effects of non-additive QTL. $\mathbf{F}b_1^R$ and $\mathbf{F}b_1^F$ are the respective vectors of the inbreeding depression of body weight for each individual. $\mathbf{F}b_1^R$, $\mathbf{F}b_1^F$ and $\mathbf{b_2}$ are fixed effects, where $\boldsymbol{\alpha}$ are random. These models were fitted using a BayesC0 algorithm as implemented in GenSel using standard priors and convergence was assessed using the Geweke diagnostic. To assess the impact of the recessive QTL on the effect of inbreeding, we compared $b_1^R$ and $b_1^F$ ; further details regarding comparison of these models, and complementary analyses performed to estimate the contribution of random selections of variants to inbreeding depression, are given in the Supplementary Note.

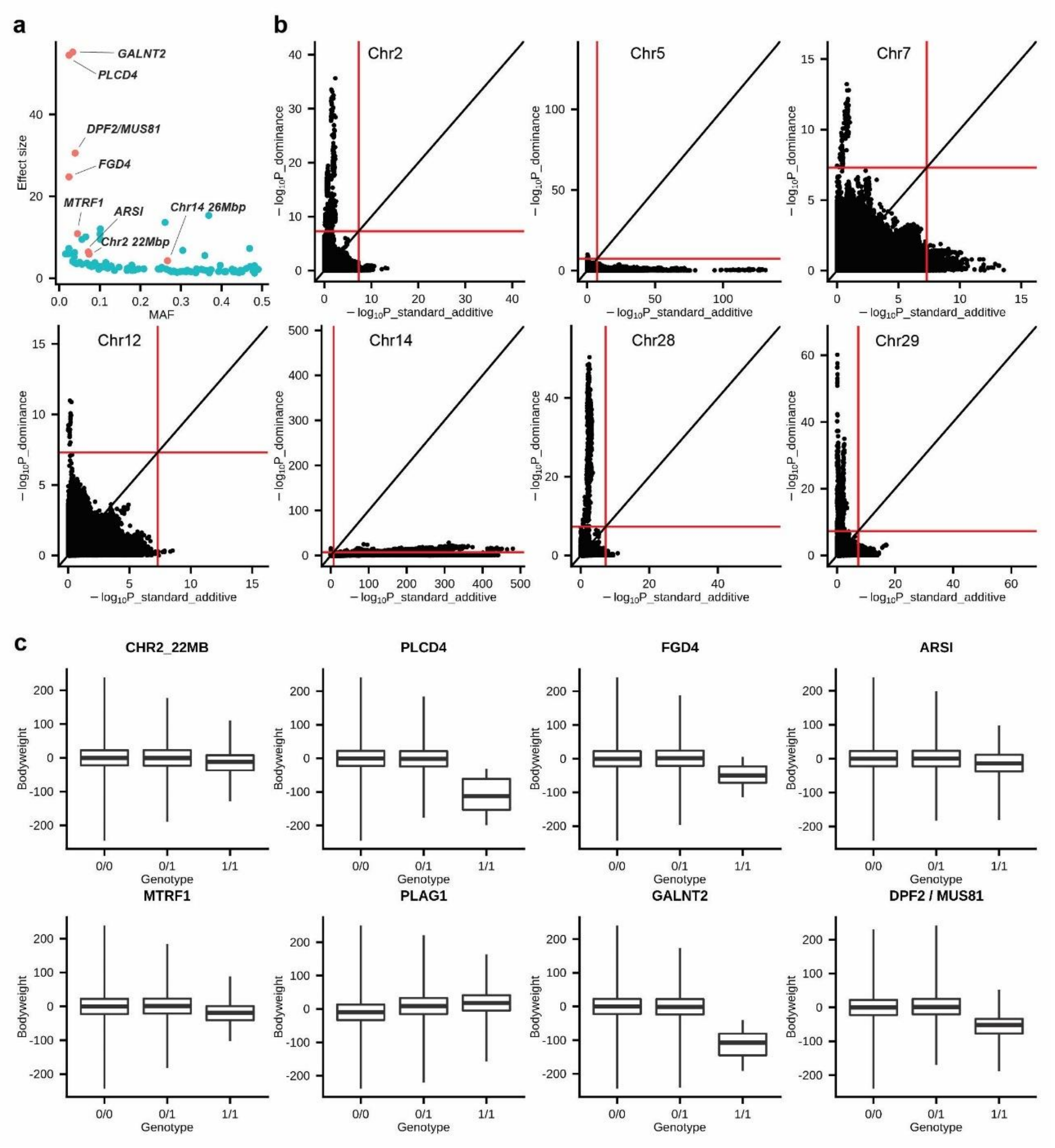
### 2.3.10 Simulation.

We performed simulation analyses to investigate factors influencing sensitivity of detection and confirm that our model could differentiate recessive QTL from standard-additive effects more broadly. Here, QMSim (Sargolzaei and Schenkel 2009) software was used to simulate a population with 800,000 SNP markers and 750 standard-additive QTL distributed across the 29 *B. taurus* autosomes, using parameters broadly aligned with those described in Brito et al. (2011). The exact parameters and protocol are described in the Supplementary Note. To generate non-additive QTL in these data, we randomly selected 30 markers from our simulated dataset with allele frequencies ranging from 0.01 to 0.05 to designate as recessive mutations. A complete recessive effect equivalent to one-half or one full standard deviation of the phenotype was then assigned to these markers (comparable to the magnitude of significant QTL described in this study). Our GWAS methods were then applied to this dataset to determine the sensitivity of QTL detection.

## 2.4 Extended Data Figures



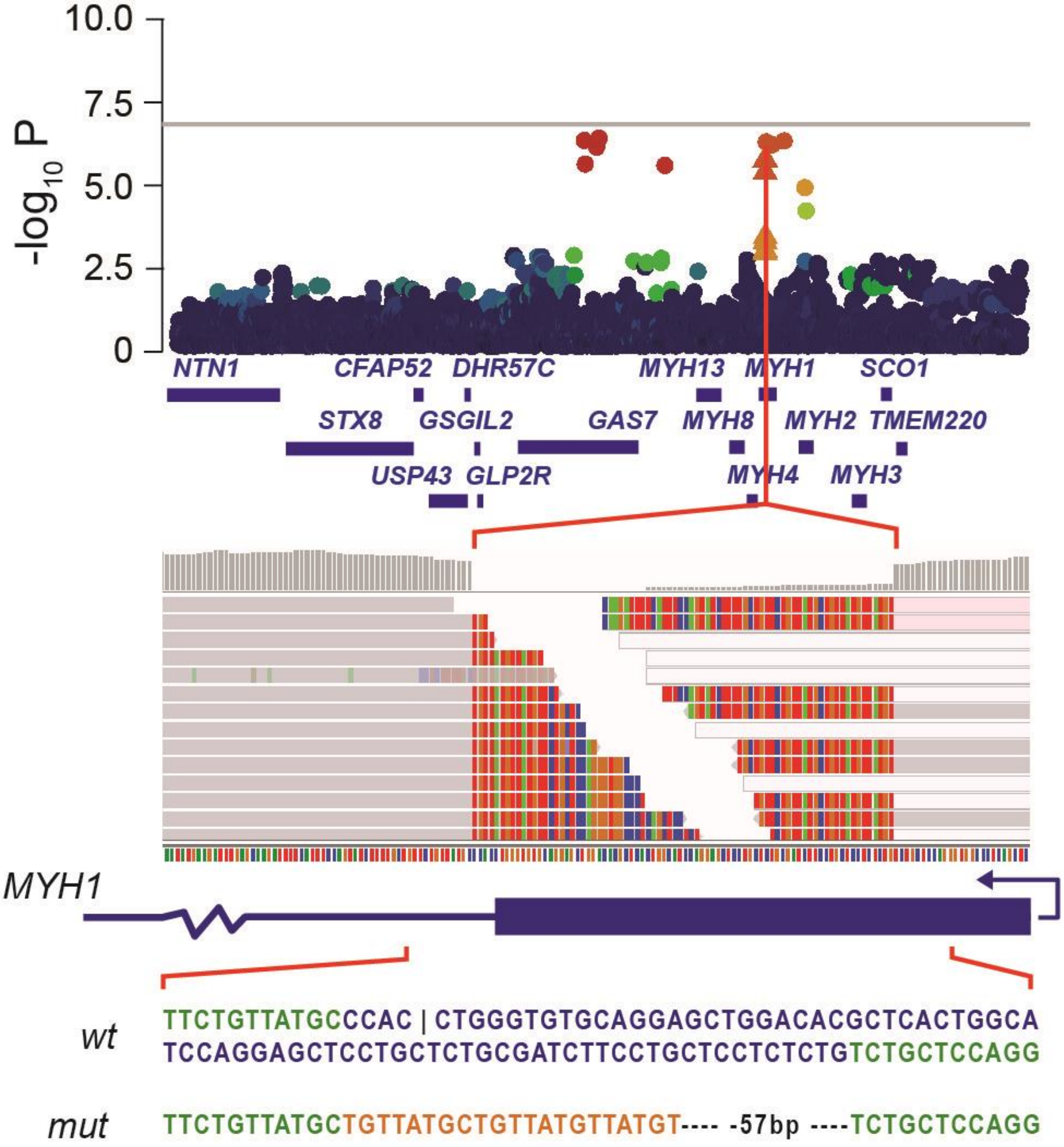
**Extended Data Figure 2.1 | Power calculations for different association models and sample sizes.** Plot contrasting the power of detection for different phenotype sources (i.e. cows, sires), different sample sizes, and different models (additive, recessive, class effect, and additive – without affected sires) given a locus explaining 0.1% of the phenotypic variance with a minor allele frequency of 2.5%, and heritability of 0.25. For sire models, each genotyped sire is assumed to have 100 un-genotyped daughters. Note that although sire models generally present higher power than cow models for a given number of genotyped animals, these models are not directly comparable since analyses based on breeding values typically leverage far fewer sires than studies using cows directly. Also note power for these sire models is provided for context only, since breeding values were not leveraged in analyses reported in this manuscript.



**Extended Data Figure 2.2 | Dominance and additive QTL contrasts of allele frequencies, effect sizes, genotype group means, and p-values.**

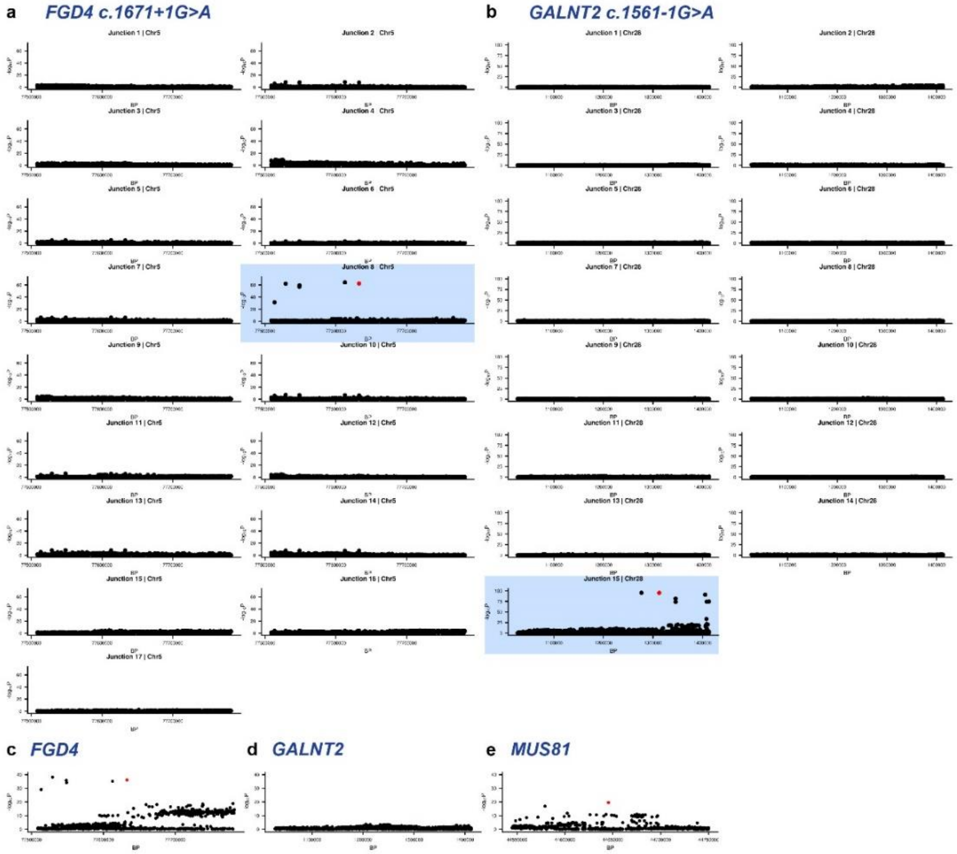
**a**, Plot contrasting minor allele frequency (MAF) and absolute effect size (Effect size, kg) of QTL identified in the standard-additive model (blue), and the dominance model (red) for body weight. Note that for equitable effect size comparison both additive and dominance estimates are represented as allele substitutions (i.e. effect of the heterozygote compared to the reference homozygote), so dominance effects only represent half the effect observed in homozygous individuals. **b**, Chromosome-wide scatterplots contrasting P-values of the eight recessive QTL for the standard-additive model (p_standard_additive) and dominance model (p_additive/p_dominance). P-values were computed using

Z-tests, the genome-wide significance thresholds of $p = 5 \times 10^{-8}$ are drawn in red, an x = y line is drawn in black. Note that only seven chromosomes are presented since two effects were identified on chromosome 2 (and thus are not readily differentiated). **c**, Box plots showing adjusted-bodyweight genotype means for the 8 non-additive loci in the discovery population (N=79,945 cows). Genotypes used for display represent putative causative mutations or lead associated variants where no obvious candidate was identified. Note the largely additive effect presented by the *PLAG1* locus, highlighted in GWAS due to a partial dominance effect. Box plots show median (centre line), interquartile range (box limits), and upper and lower whiskers (maxima and minima data points).



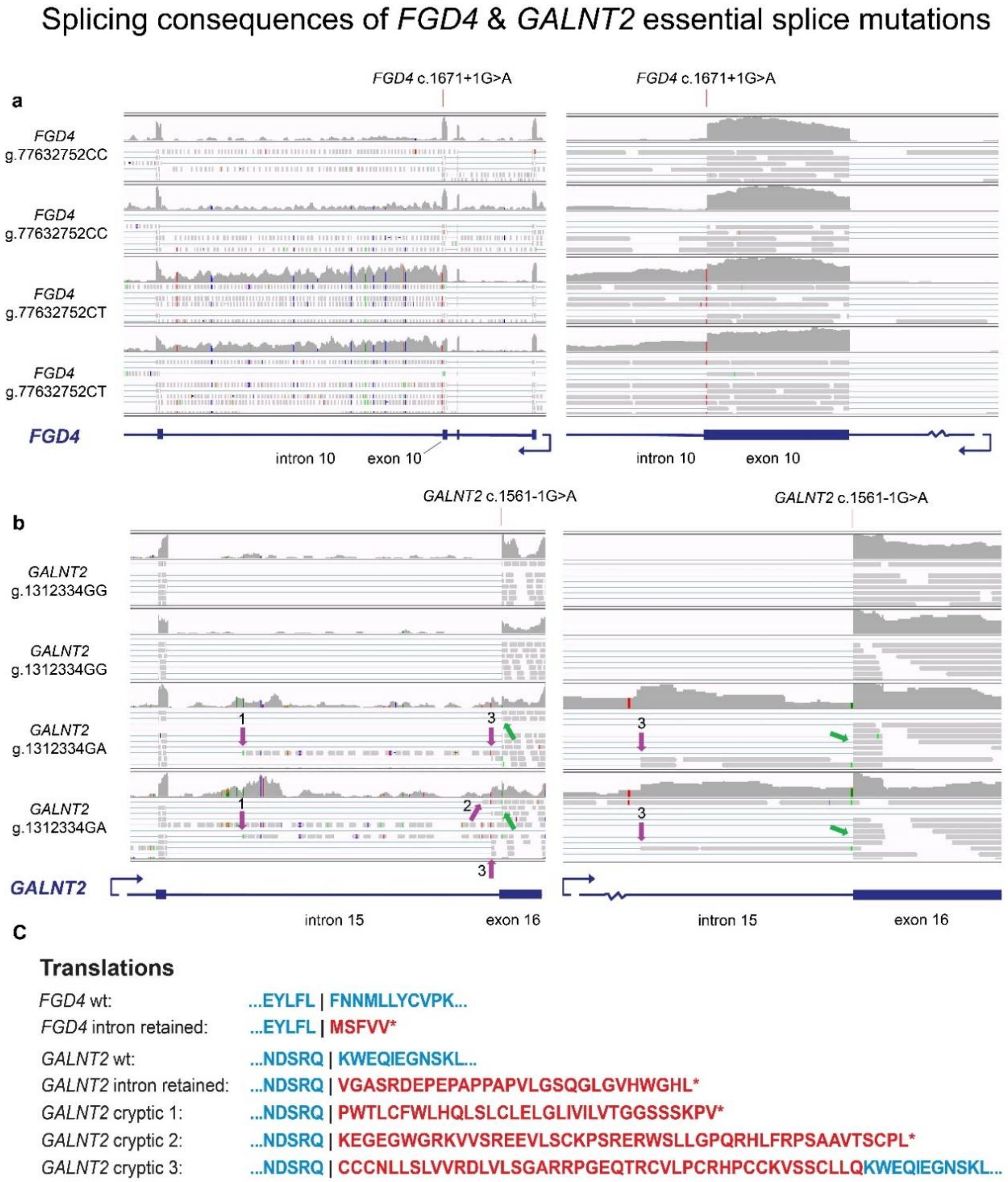
**Extended Data Figure 2.3 | Manhattan plot of Chr19 recessive QTL showing *MYH1* frameshifting indel.**

1.5Mbp sequence interval showing the top genome-wide non-additive association signal from analysis of body condition score in 75,617 cows; P-values were calculated using Z-tests. The genome-wide significance threshold of P < 5 × $10^{-8}$ is indicated by the horizontal grey line, note no variants at this locus surpassed this threshold (smallest P=7.6 × $10^{-8}$). Lead variants of the signal tag a 78 bp compound insertion deletion variant evident from inspection of whole genome sequence alignments. Genome sequence alignment of homozygous animal shown, resulting in predicted knockout of MYH1 due to simultaneous loss of 19 amino acids and introduction of a premature stop codon at exon 34.



**Extended Data Figure 2.4 | Per-junction and per-gene Manhattan plots for splicing efficiency and gene expression QTL.**

Manhattan plots showing splicing efficiency and whole-transcript expression QTL effects (P-values calculated using 2-sided t-tests). Splicing efficiency analysis was performed for the *FGD4* (**a**) and *GALNT2* (**b**) genes, with associations highlighting junctions for which splicing appears to be genetically modulated in cis. The proportion of spliced to un-spliced reads at each junction has been treated as an individual phenotype, with association analysis performed using intervals of imputed sequence data spanning the annotated gene structures, and ± 100kbp 5' and 3' of the gene boundaries. The splicing junction predicted to be impacted by the splice donor (*FGD4*) and acceptor (*GALNT2*) mutations is indicated by the blue highlighted panels, with the candidate causative mutation indicated in red. Whole transcript eQTL analysis was performed to assess possible gene expression impacts as a consequence of nonsense mediated RNA decay (NMD) for the *FGD4* (**c**), *GALNT2* (**d**), and *MUS81* (**e**) genes that harbour nonsense mutations. Note that for the two genes that show significant eQTL (*FGD4* and *MUS81*), the mutant allele is overexpressed and thus no NMD is apparent. In the case of the *MUS81* and *GALNT2* genes, lack of apparent NMD can be anticipated given the position of the *GALNT2* c.1561-1G>A mutation in the final exon, and the presence of an in-frame start codon (p.Met76) following the p.Gly70 mutation for *MUS81*.



**Translations**

*FGD4* wt: ...EYLFL | FNNMLLYCVPK...

*FGD4* intron retained: ...EYLFL | MSFVV*

*GALNT2* wt: ...NDSRQ | KWEQIEGNSKL...

*GALNT2* intron retained: ...NDSRQ | VGASRDEPEPAPPAPVLGSQGLGVHWGHL*

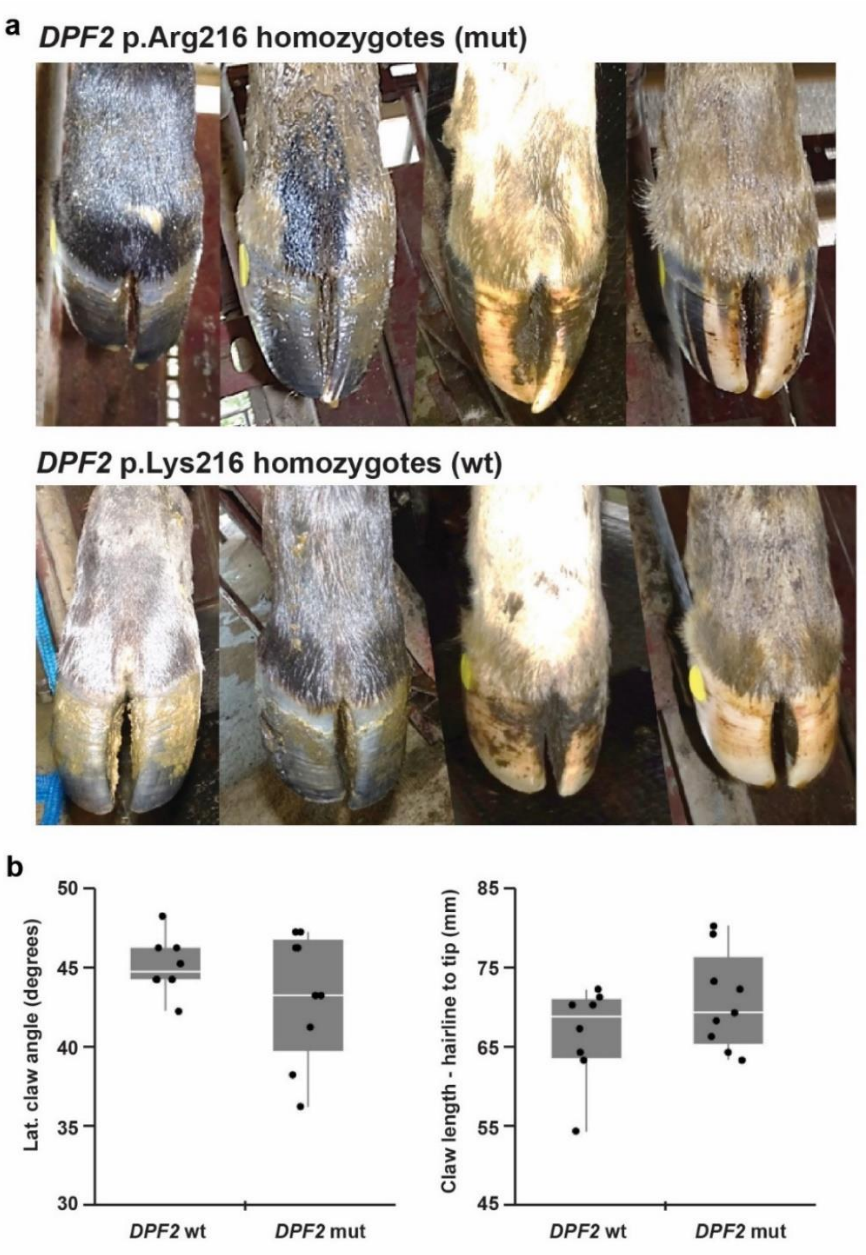
*GALNT2* cryptic 1: ...NDSRQ | PWTLCFWLHQLSLCLELGLIVILVTGGSSSKPV*

*GALNT2* cryptic 2: ...NDSRQ | KEGEGWGRKVVSREEVLSCKPSRERWSLLGPQRHLFRPSAAVTSCPL*

*GALNT2* cryptic 3: ...NDSRQ | CCCNLLSLVVRDLVLSGARRPGEQTRCVLPCRHPCCKVSSCLLQKWEQIEGNSKL...

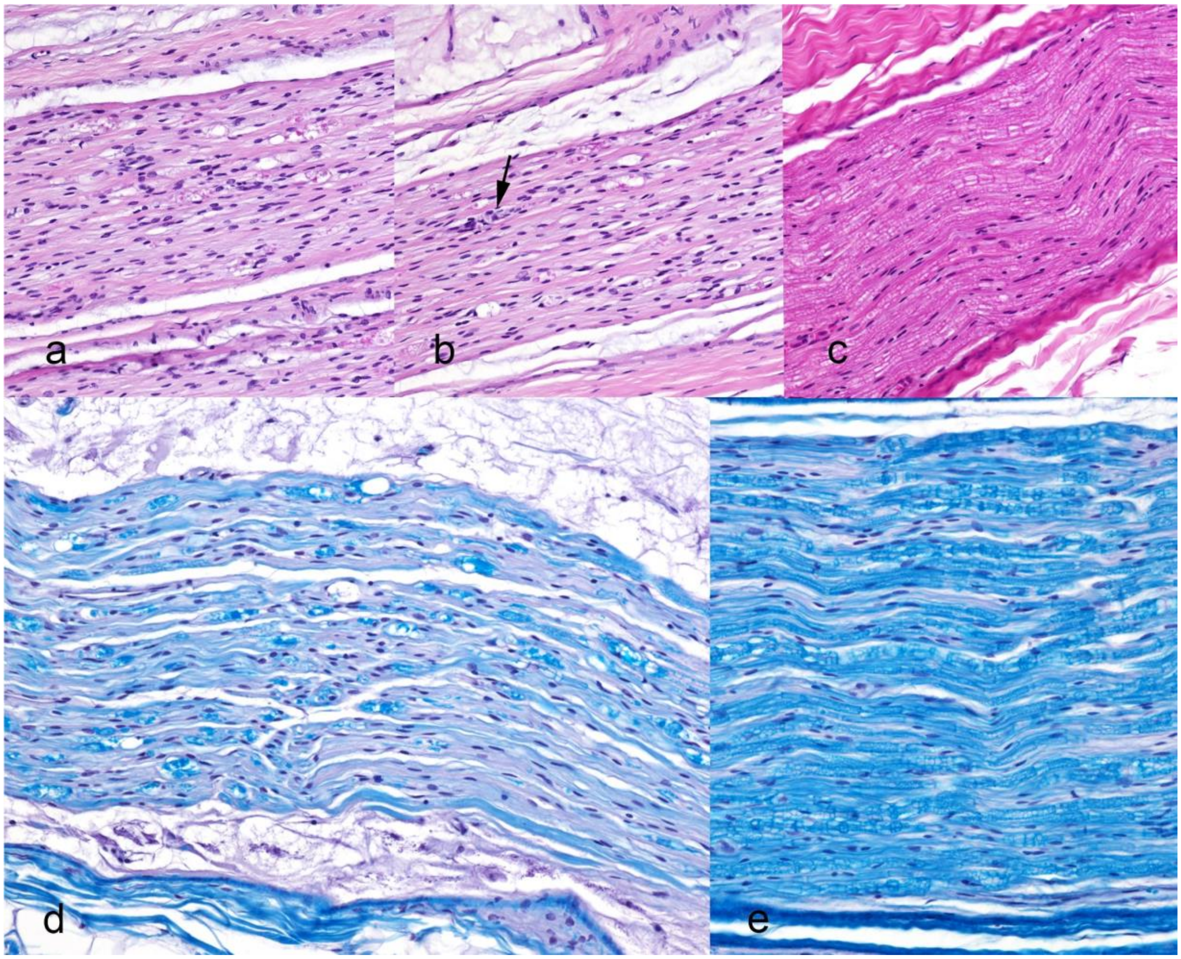
**Extended Data Figure 2.5 | Splicing consequences of *FGD4* and *GALNT2* essential splice mutations.**

Mammary RNA-seq alignments for the *FGD4* (**a**) and *GALNT2* (**b**) genes, showing wildtype and carrier animals for the *FGD4* c.1671+1G>A and *GALNT2* c.1561-1G>A splice mutations (two animals representing each genotype class per gene). Intron and exon numbers reference the ENSBTAT00000007175.5 and ENSBTAT00000006404.5 transcript annotations for the *FGD4* and *GALNT2* genes respectively. Right-most panels show intron-exon boundaries of the mutation-implicated splice junction, left-most panels show kilobase-level views of the whole intron and adjoining exon junction. Coverage tracks demonstrate clear intron retention for *FGD4* heterozygous mutants, without obvious cryptic splicing. Animals heterozygous for *GALNT2* mutant transcripts show less uniform intron retention, though at least three recurrent cryptic splice sites indicated by the purple arrows (green arrows show annotated junctions). **c** Putative translations for these alternatively spliced transcripts are indicated (light blue=reference splice, red=mis-splice), where the first base of the new acceptor exon boundaries are: cryptic 1 chr28 g.1309085; cryptic 2 chr28 g.1312087; cryptic 3 chr28 g.1312203. Note that all intron retention and cryptic splices are predicted to cause premature termination, with the exception of the '*GALNT2* cryptic 3' isoform that encodes a 44aa 5' frame-extension of exon 16.

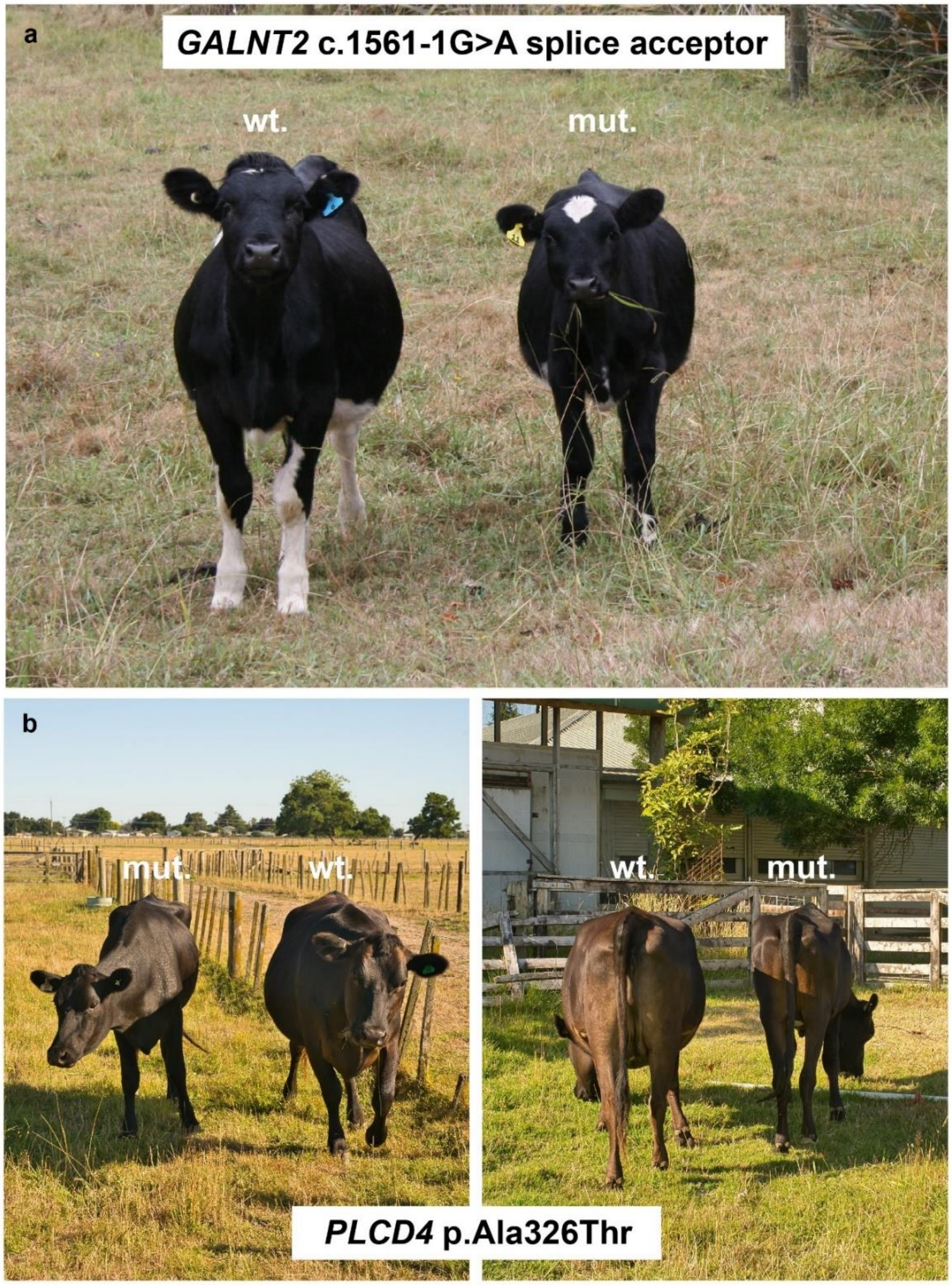

a
DPF2 p.Arg216 homozygotes (mut)
DPF2 p.Lys216 homozygotes (wt)
b
Lat. claw angle (degrees)
50
45
40
35
30
DPF2 wt
DPF2 mut
Claw length - hairline to tip (mm)
85
75
65
55
45
DPF2 wt
DPF2 mut

**Extended Data Figure 2.6 | Hoof anatomical observations for DPF2 mutant and control individuals.**

Figure showing hoof characteristics of DPF2 mutant and control cattle. Photographs (**a**), show the right rear hooves of representative mutants and controls, with the hooves of some mutants showing subjective differences including overlapping claw-tips and longer claws overall compared to controls (see centre two animals in mutant group). However, quantitative comparisons based on hoof measurements (**b**) did not reveal significant differences between groups (N=8 mutant and 9 control animals respectively: Supplementary Table 8). Box plots show median (centre line), interquartile range (box limits), and upper and lower whiskers (maxima and minima data points).

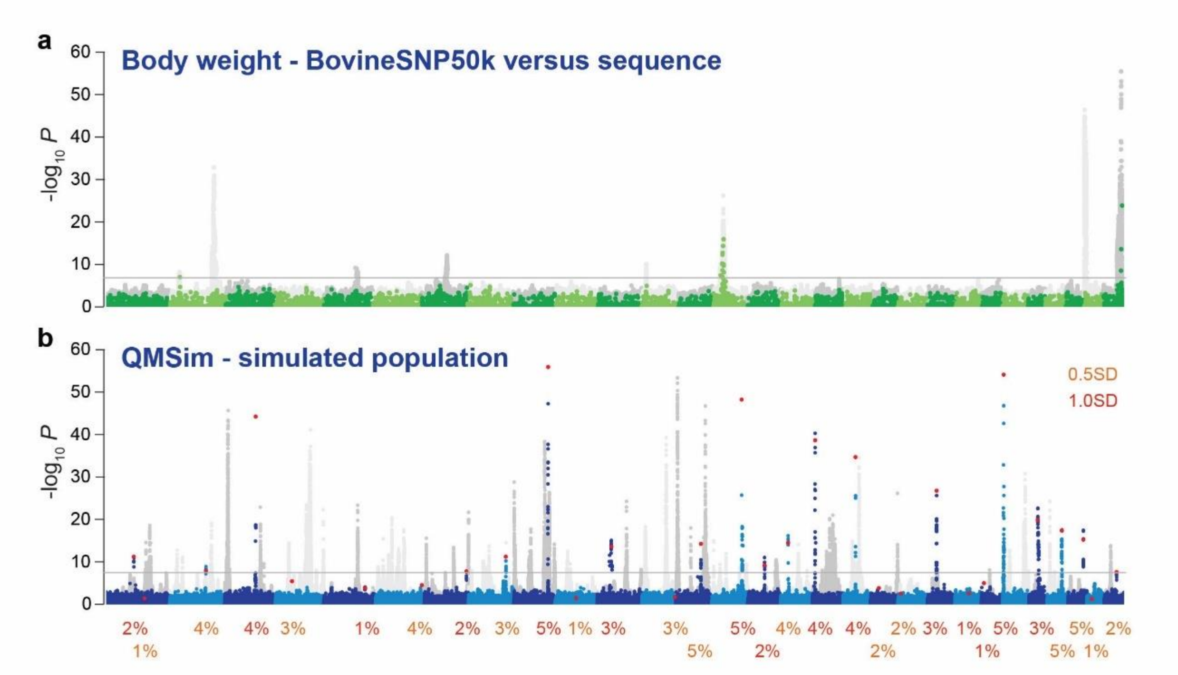


**Extended Data Figure 2.7 | Nerve histology of FGD4 mutant and control individuals.** Common digital nerve of forelimb from two different FGD4 homozygotes (**a** and **b**) showing hypercellularity, Schwann cell hyperplasia, axonal swelling and degeneration (black arrow) (2000X, HE). **c**, Common digital nerve of forelimb from control animal (2000X, HE). **d**, Saphenous nerve from an FGD4 homozygote showing lack of myelin staining consistent with demyelination (2000X, Luxol fast blue). **e**, Saphenous nerve from a control animal (2000X, Luxol fast blue). Micrographs are representative of the lesions found in 7 different nerves examined from 2 FGD4 homozygotes and 2 control animals. Each nerve was examined in 3 locations, with both transverse and longitudinal sections. Bar = 50 µm.

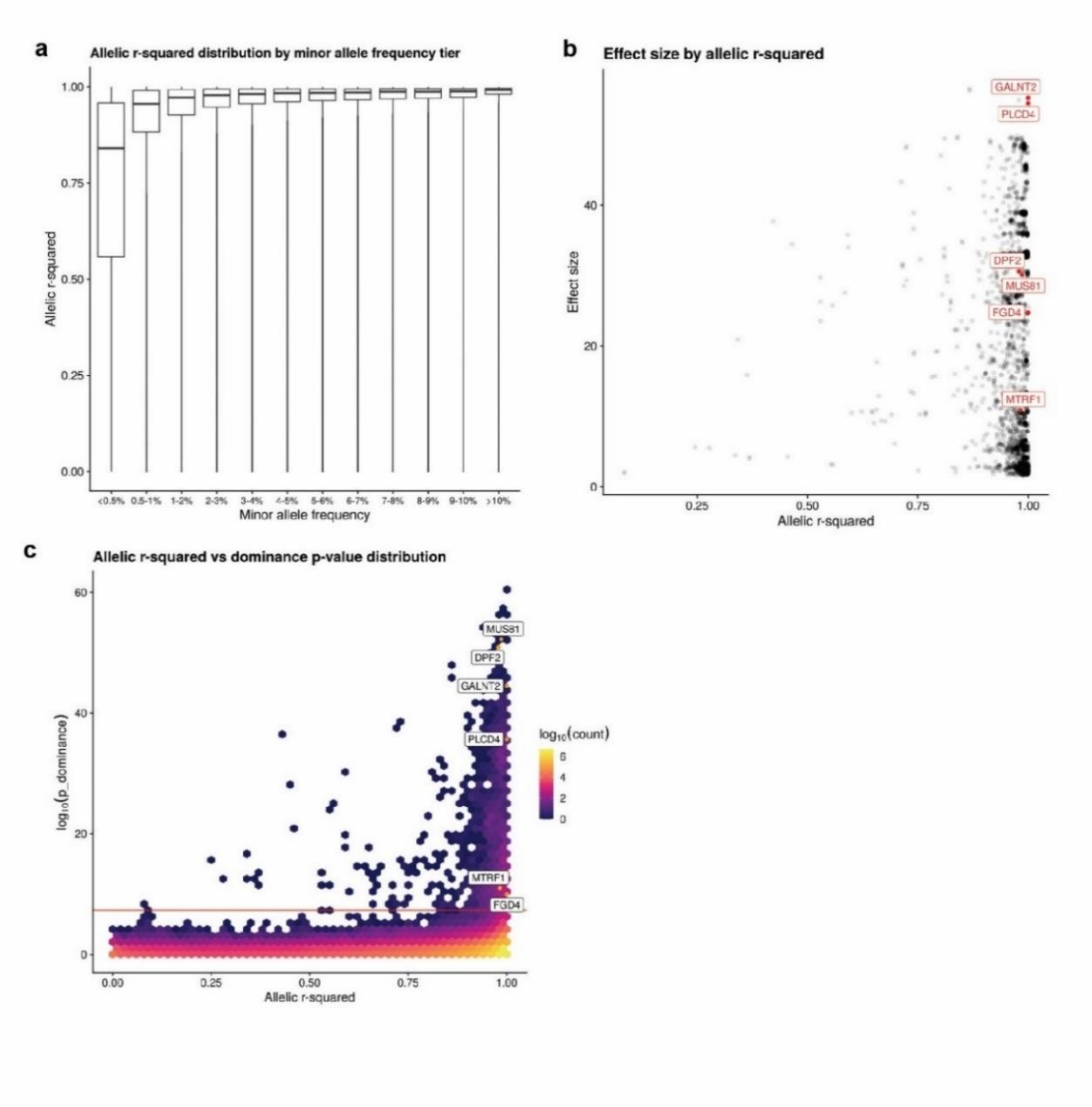


**Extended Data Figure 2.8 | Photographs of *GALNT2* mutant and control individuals; *PLCD4* mutant and control individuals.**

Photographs contrasting homozygous mutant and homozygous reference animals for the *GALNT2* c.1561-1G>A splice acceptor mutation (**a**), and *PLCD4* p.Ala326Thr mutation (**b**). For the *PLCD4* variant, front and rear images contrast the same two animals. Animals represent individuals from the research farm studies that were neither the smallest nor largest animals within each of their genotype classes. Photos are unstandardised and provided for qualitative purposes.



**Extended Data Figure 2.9 | GWAS Manhattan plots exploring sensitivity of QTL detection for reduced resolution genotype data, and simulated loci varying in effect size and frequency.** Manhattan plot showing impact of marker density on discovery of non-additive bodyweight GWAS signals (**a**; P-values computed using Z-tests, horizontal grey line indicates the genome-wide significance threshold of $P < 5 \times 10^{-8}$). Here, dominance estimates from sequence-based bodyweight GWAS (grey dots) are plotted alongside a subsetted version of these same data filtered to represent the content of BovineSNP50k SNP-chip platform (green dots). While two of the modest effect, comparatively higher MAF QTL retain significance (i.e. Chr2:22Mbp and *PLAG1* locus), only the *DPF2*/*MUS81* QTL is represented among the major-effect, recessive signals. (**b**) Manhattan plot showing the influence of MAF and effect size on sensitivity of detection in a simulated dataset. Dominance estimates (blue dots) are contrasted with standard-additive estimates (grey dots), showing sensitivity of detection for 30 recessive causative mutations (red dots). Recessive effects were generated by randomly selecting variants from 1-5% MAF bins from the pool of simulated genotypes (frequencies indicated at bottom), with effect sizes assigned as 0.5 standard deviations (SD; light orange) or 1.0 SD (dark orange) per mutation. Mutations were selected to represent all chromosomes (two on chromosome 1).



**Extended Data Figure 2.10 | Visualisation of sequence imputation allelic R-squared statistics by minor allele frequency, dominance effect sizes, and dominance p-values in the GWAS dataset.** Plots showing imputation allelic $R^2$ (AR2) values of genotypes from the discovery population, where AR2 is taken to reflect accuracy of imputation, representing the squared correlation between the allele dosage with the highest posterior probability and the true allele dosage (Browning and Browning 2009). **a**, Box plots showing distributions of AR2 within different MAF classes for the 16,128,757 sequence variants used for GWAS. Box plots show median (centre line), interquartile range (box limits), and upper and lower whiskers (maxima and minima data points). **b**, Plot showing absolute dominance effect size (Effect size, kg) for genome-wide significant variants ($P < 5 \times 10^{-8}$) from the bodyweight GWAS, visualised by AR2. Also indicated are the candidate causative mutations of interest; effects are expressed as allele substitutions and thus represent half the effect observed in homozygous mutant individuals **c**, Scatter/density plot showing relationship between P-value and AR2 for the sequence variants tested in the bodyweight GWAS (dominance model), with mutations of interest also indicated.

## 2.5 Supplementary Note

Three of the loci highlighted in this study implicate genes for which analogous variants have been demonstrated to cause human syndromes. Given the potential interest of these findings, we provide here an elaborated description of observations and experiments conducted to investigate these mutations. This document also contains elaborated discussion and considerations on non-additive model formulation, highlighting differences (and similarities) with previously implemented non-additive GWAS models.

### 2.5.1 Cattle mutations with human disease analogues

*FGD4 – a gene implicated in human Charcot Marie Tooth Disease*

Using non-additive association analysis, we detected a significant locus for bodyweight on bovine chromosome 5 at 77.6 Mbp. This signal was represented by a single, significant variant at the locus, representing a nonsense mutation in the FYVE, RhoGEF and PH domain containing 4 gene (*FGD4;* Figure 2.2), a gene for which nonsense variants in humans have been proposed to underlie Charcot Marie Tooth disease (CMT). This disease is the most common inherited neurological disorder in humans (affecting ~1 in 2500 people)(Szigeti and Lupski 2009), so it is noteworthy that ~1 in 590 cows are expected to be affected by the *FGD4* mutation in the NZ dairy population (see Supplementary Table 3 and main manuscript for frequency calculations).

Charcot Marie Tooth disease describes a group of peripheral neuropathies characterised by nerve degeneration and muscular atrophy, primarily affecting the feet, hands, and legs (Bird 1993). These changes may lead to disability including hand weakness and loss of sensation, difficulties standing and walking, and increased risk of injuries due to a propensity to trip and fall. The disease displays variable penetrance and shows autosomal dominant, recessive, and X-linked forms of inheritance (Szigeti and Lupski 2009). CMT is categorised on these modes of inheritance and the electrophysiological and histopathological parameters of the disease. Mutations in human *FGD4* are classified as CMT4H, a subgroup of the CMT4 autosomal recessive demyelinating forms of disease (Delague et al. 2007). Frabin, the protein encoded by *FGD4*, is an actin-filament binding protein that has GDP/GTP exchange activity specific for Cdc42, and is involved in the regulation of the actin cytoskeleton and cell morphology (Ono et al. 2000). The precise role of Frabin in CMT is not well described, though appears to relate to its interaction with Cdc42, given that knockout of either *Fgd4* or *Cdc42* in mice results in both impaired nerve development, and loss of myelin in adult nerve fibres (Horn et al. 2012).

In the context of the current study, we note that the recessive mode of inheritance of *FGD4* mutants in CMT4H is consistent with the recessive presentation of the *FGD4 c.1671+1G>A* splice donor mutation in cattle reported here. We are unaware of other cattle or large animal models of CMT, so the relative severity of the cattle mutation is unknown, though it is noteworthy that recessive, human CMT4 forms of the disease tend to comprise earlier onset, more severe cases that lead to pronounced disability (de Sandre-Giovannoli et al. 2005). Upon identification of the *FGD4 c.1671+1G>A* mutation as a candidate for the bodyweight QTL, we identified local farms with homozygous animals highlighted from our GWAS. Upon visiting one of these farms and identifying the animal in question, the farmer remarked that this cow had "nerve issues". This was an unsolicited observation, where no prior mention of the purpose of inspection had been made to the farmer. Further anecdotal accounts of this animal, and another (recently culled) animal of the same family lines suggested both were prone to stumble, where in the case of the culled animal, it had on one occasion fallen and not been able to get back up unassisted.

During the inspection, the animal identified from GWAS was subsequently walked around a concrete yarded area where it was seen to demonstrate a stumbling phenotype. Here, while walking a straight line, its rear leg momentarily collapsed after which it immediately recovered and continued walking as normal again.

Following these observations, we conducted a genotypic screen of 568 animals to identify calves homozygous for the *FGD4* (and other) mutations. This analysis identified 22 *FGD4* homozygotes, which at ~9 months of age were ~9.8kg lighter than controls (P=0.018*;* Supplemental Table 7). Nine of these homozygous mutant animals were then recruited for the farm trial described in greater detail in our accompanying manuscript. These animals were grazed together with age-matched control animals (and animals representing homozygotes of the other mutations of key interest). Growth rates of *FGD4* homozygotes were significantly reduced (P=2.65x10$^{-24}$; Supplemental Table 8), with the bodyweight difference between these animals and controls widening to 49.7kg by 24 months of age (P=9x10$^{-3}$; Supplemental Table 8). At the end of the farm trial (~27 months of age), *FGD4* mutant animals subjectively demonstrated behavioural differences and instances of loss of motor control. In these cases, routine animal handling procedures such as confinement in a cattle crush and head bail appeared to lead to increased restlessness and agitation in FGD4 mutant animals, with some animals collapsing to a 'kneeling' position – a behaviour not observed in controls or mutant animals representing the other genotype classes. These observations were made in the months prior to post mortem dissection and histological examinations that demonstrated neuronal abnormalities in *FGD4* mutant animals (Figure 2.10).

Given that muscle atrophy is a hallmark of human CMT, we assume that the significant bodyweight association for the *FGD4* mutation is due to wastage of muscle (Table 2.1). The finding of lower serum creatinine ($P<0.05$) in *FGD4* mutants also supports a muscle wastage hypothesis (Schutte et al. 1981), and is of further note given the importance of creatine metabolism to the nervous system generally (Andres et al. 2008). Other significant associations in the discovery population included negative impacts on lactation traits, namely reduced milk protein, fat, and volume yield (Table 2.1). Although we are unaware of lactation impacts attributed to CMT in human studies, we speculate these effects reflect compound stresses from disease and the additional metabolic demands of lactation, and we note that worsening of symptoms may be experienced in pregnant individuals with CMT1 (Rudnik-Schöneborn et al. 1993).

### *DPF2 – a gene implicated in human Coffin-Siris Syndrome*

As part of the same genome scan of bodyweight referenced above and reported in detail in the accompanying manuscript, we detected a significant locus for bodyweight on bovine chromosome 29, with peak association at 44.1 Mbp. This locus presented two highly associated candidate causative mutations (both $R^2>0.98$ with the lead variant; Figure 2), comprising a p.Lys216Arg amino acid substitution in *DPF2*, and a p.Gly70* premature stop mutation in *MUS81*. Details of these candidates are discussed below, with an emphasis on the *DPF2* missense variant given the gene's implication in human Coffin-Siris Syndrome.

Coffin-Siris Syndrome (CSS) is a rare genetic disorder caused by mutations in BRG1-associated factor (BAF) chromatin-remodelling complex-subunit genes that include double PHD fingers 2 (*DPF2)* (Knapp et al. 2019; Zarate et al. 2016). This syndrome is characterised by intellectual disability and developmental delay, coarse facial features, fingernail and/or toenail abnormalities, and a variety of other clinical features (Kosho, Miyake, and Carey 2014). The functions of the genes so far implicated in CSS suggest the syndrome results from alterations in chromatin remodelling, manifesting through dominant modes of inheritance via dominant-negative, gain of function, or loss of function mechanisms (Vasileiou et al. 2018).

Two papers have recently described *de novo* mutations in *DPF2* as responsible for CSS, highlighting missense and nonsense variants with variable clinical presentation (Knapp et al. 2019; Vasileiou et al. 2018). All six missense variants highlighted in these studies impact the PHD domains of *DPF2*, whereas the cattle p.Lys216Arg substitution reported here maps to the C2H2-type zinc finger domain. The recessive presentation of effects for the chr29 QTL similarly contrasts with the dominant negative effects observed for human *DPF2* mutations. Although the p.Lys216Arg substitution represents a conservative change, it is noteworthy that this residue (and C2H2 domain overall) is highly evolutionarily conserved (W. Zhang et al. 2011) (Figure 2), and given the bodyweight and stature effects that are also a feature of human *DPF2* mutations (Vasileiou et al. 2018), we thus considered p.Lys216Arg a plausible candidate mutation for the QTL.

To examine this possibility further, we investigated hoof conformation as a potential bovine-equivalent to the fingernail and toenail abnormalities seen in human CSS. Here, we pursued both quantitative and qualitative measurements of hooves in *DPF2* mutants and controls as part of the same farm trial investigating *FGD4* mutant phenotypes. Supplementary Table 8 and Figure 2.9 shows claw length and toe angle measurements of these animals, with photographs of hooves from animals of different genotype also represented. Although no significant differences were observed for hoof dimension data, the mutant group appeared to present more variable hooves (Figure 2.9). Hooves were also assessed by a veterinarian specialising in lameness, and a professional hoof trimmer/inspector – considered as part of a blinded inspection to characterise the animals either as 'positive' or 'normal' for affection status. Three mutants and one control animal were characterised as 'positive' in this analysis (the latter qualified as being 'mild'), though neither the mutant nor control animals were judged to be overtly abnormal in these inspections. Although these analyses did not demonstrate resolute differences in the hooves of *DPF2* homozygotes, these findings do not necessarily preclude the causality of the p.Lys216Arg variant, given the highly variable (and in some cases mild) phenotypic presentation of nail abnormalities in human CSS (Knapp et al. 2019; Vasileiou et al. 2018).

As mentioned above, the chr29 44Mbp bodyweight and stature locus presented two candidate coding variants as potentially responsible for these effects. Given that hoof data did not conclusively implicate the involvement of the *DPF2* p.Lys216Arg variant, it is therefore of interest to consider the potential role of the *MUS81* p.Gly70* premature stop mutation in these QTL. The *MUS81* gene encodes the catalytic subunit of an endonuclease responsible for cleavage of branched DNA substrates in eukaryotic chromosomes, proposed to help maintain chromosomal stability through processing of stalled DNA replication forks (Hanada et al. 2007). Mice bearing homozygous *Mus81* null mutations are born and develop normally, though have decreased survival following DNA damage when exposed to DNA crosslinking agents (Dendouga et al. 2005). Cells from these mice are similarly more sensitive to DNA damage, show increased chromosomal aberrations, and show abnormalities in cell cycle progression (Dendouga et al. 2005); human *MUS81* null cell lines have also been shown to replicate DNA more slowly (Fu et al. 2015). These observations suggest potential DNA replication and/or chromosomal abnormalities as an alternative explanation for the phenotypic changes seen in *MUS81/DPF2* homozygotes.

Given the two competing candidate mutations, it is also interesting to consider whether they could be differentiated through large-scale population-based genotyping, and at what scale of data that this could be achieved. As part of animal breeding activities and with a view to implementing the discoveries reported in this manuscript, we have recently included the *DPF2*, *MUS81*, and other highlighted recessive variants as custom content on a newly-developed low-density Illumina SNP chip platform. At time of writing, we had physically genotyped the two candidate SNP in the first 98,002 animals typed on this platform, where the $R^2$ between SNP was 0.985 in this largely mixed breed population of calves. We identified a total of 106 animals with recombinant haplotypes in this dataset, however as expected, most animals presented heterozygote, major allele homozygote diplotypes. However five animals were homozygous for one of the mutations of interest, and manual inspection of SNP-chip signal intensity and genotype clusters confirmed the likely validity of these recombinants. Although these individuals did not have phenotypes and thus are unable to be analysed in the immediate term, these findings suggest differentiation of the mutations should be possible as largescale population data continues to accumulate.

### *GALNT2 – a gene underlying a human O-linked glycosylation disorder*

The most significant bodyweight and stature effects identified from GWAS highlighted loci on chromosome 28 at 0.7Mbp and 1.3Mbp respectively. The top-associated SNP in these analyses presented a highly correlated ($R^2$>0.9) c.1561-1G>A splice acceptor mutation in

*GALNT2* as potentially responsible for these effects (Figure 2). Of the seven recessive mutations of key focus in our accompanying manuscript, this *GALNT2* loss-of-function variant was the only such mutation for which some prior phenotypic implication had been made, where we had provisionally mapped the variant as a candidate stature mutation (unpublished), and subsequently as a variant for which homozygotes were depleted in our population (Charlier et al. 2016). We further discuss this candidate within the context of recent papers highlighting *GALNT2* loss-of-function variants as underpinning lipid metabolic phenotypes and a novel human glycosylation disorder, below.

The *GALNT2* gene encodes the polypeptide N-acetylgalactosaminyltransferase 2 enzyme responsible for mucin-type *O* glycosylation of proteins. Although this enzyme can be assumed to have many substrates, its most prominent role is as a modulator of circulating high density lipoprotein cholesterol (HDL) and triglyceride levels – as first demonstrated in human GWAS of blood lipid concentrations (Willer et al. 2008; Kathiresan et al. 2008). The role of the gene in lipid metabolism was 84urtherr elaborated following a study implicating non-coding, missense, and nonsense mutations causing similar phenotypes, namely reduced HDL and triglycerides (Khetarpal et al. 2016). This study also highlighted candidate proteins as the likely mediators of these effects, showing differential glycosylation of human and/or rodent ANGPTL3, ApoC-III, and PLTP (Khetarpal et al. 2016). Although this study did not highlight a pathogenic consequence to the nonsense variant, a subsequent analysis based on patients homozygous for this variant, and additional nonsense and missense variants recently implicated *GALNT2* loss-of-function mutations as underlying a new congenital disorder of glycosylation – termed GALNT2-CDG (Zilmer et al. 2020).

The description of GALNT2-CDG is based on a single, recent paper, reporting the same lowered triglyceride and HDL levels highlighted in other analyses of human *GALNT2* polymorphisms, in addition to many other intellectual, behavioural and anatomical clinical features (Zilmer et al. 2020). These phenotypes include developmental delay and autistic behaviours, epilepsy, white matter lesions as evidenced by brain MRI, dysmorphic facial features, short stature, lower bodyweight, and microcephaly (two of seven patients investigated only). This paper also investigated two rodent models of *Galnt2* knockout in detail. Male mice, and both male and female rats homozygous for *Galnt2* null mutations were significantly smaller than heterozygous or homozygous wild-type animals. Startle testing, social responses, and other mouse behavioural assessments also showed differences between homozygous knockouts and controls (Zilmer et al. 2020).

Experiments to investigate the cattle *GALNT2* mutation pre-date description of GALNT2-CDG, though it is interesting to compare our analyses with those findings. Foremost, the reduced growth parameters of homozygous individuals appears to be shared for human, mouse, rat, and cattle mutations (Table 2.1). We also observed a significant reduction in circulating triglycerides, though HDL was not significantly different in cattle (Supplementary Table 9). Creatinine was significantly lower in mutant cattle (Supplementary Table 9), consistent with the reductions observed in all patients with GALNT2-CDG. The substantial embryonic lethality observed in the rodent knockout models is also consistent with our findings. In heterozygous crosses, 14% of mice, and 12% of rats were homozygous for the null mutations (c.f. 25% expected) (Zilmer et al. 2020), where we identified 16% *GALNT2* c.1561-1G>A homozygotes from heterozygous matings in our cattle population (Supplementary Table 5). Neurological and behavioural traits were not assessed in the current study, though given the range of such features presented in patients with GALNT2-CDG, and in knockout rodent models, similar complementarity might also be anticipated for the cattle mutation described here.

Given the key role of the liver in lipid metabolism, we obtained liver biopsy samples representing *GALNT2* c.1561-1G>A homozygotes and controls, with the aim of conducting RNA-seq to investigate transcriptional differences between genotypes. Differential expression analysis (see Supplementary Methods below) revealed a total of 14 genes significantly over-expressed, and 15 genes significantly under-expressed in *GALNT2* c.1561-1G>A mutants (adjusted P-value<0.05; Supplementary Table 10). Given observations of altered lipid metabolism demonstrated for *GALNT2* null mutations across species, perhaps the most striking change was a ~14-fold increase in expression of *FGF21* in the liver of *GALNT2* mutants - a hormone responsible for modulating a range of glucose and lipid metabolic functions (Ge et al. 2012). Pharmacologic administration of FGF21 hormone has been shown to lower circulating triglycerides in humans and mice (Schlein et al. 2016), though a causal role for FGF21 here is at odds with the differential glycosylation analyses that have proposed ANGPTL3, ApoC-III, and/or PLTP as responsible for these effects in *GALNT2* null individuals (Khetarpal et al. 2016). Alternatively, increases in *FGF21* expression might mediate or be a consequence of the marked body weight differences observed between genotypes, given that overexpression causes weight loss in multiple species (Véniant et al. 2012), and that expression of the hormone may be induced as a consequence of fasting (Markan et al. 2014). If reduced caloric intake is indeed the cause of differential *FGF21* expression in *GALNT2* c.1561-1G>A mutants, this observation might fit with the observation of significantly lower serum albumin in homozygotes (Supplementary Table 9). Several collagen and extracellular matrix proteins (*COL12A1, COL4A5, ECM1, ECM2, SERPINF1*) were also differentially expressed between mutants and controls (Supplementary Table 10). Although the relevance of these findings is unknown, the broad roles of these genes in tissue morphogenesis and more specific implication of *COL12A1 and SERPINF1* in bone and muscle development (Becker et al. 2011; Zou et al. 2014), may have relevance to the reduced stature and bodyweight phenotypes seen in *GALNT2* mutants. Work to investigate these potential relationships – particularly studies able to highlight the differential glycosylation targets of wildtype and *GALNT2* mutant cattle, would be of key future interest.

### 2.5.2 Additive and dominance model considerations

We implemented an association model aimed foremost at detecting deleterious recessive alleles. The formulation of this model differs from that applied in other recent publications aiming to detect nonadditive effects in human (Zhu et al. 2015), and cattle (Jiang et al. 2019) GWAS, therefore, we outline these contrasts and similarities below. Further considerations regarding these model parameterisations and effect estimations are discussed in Falconer (1960), and more recent summaries and comparisons of variance components can be found in Vitezica et al. (2013) and Sun et al. (2014).

In the context of most 'traditional' GWAS, an additive-only model is implemented, seeking to detect what we term in this article as a 'standard-additive' effect (β). The three genotype classes [G11, G12, G22] are encoded as a covariate representing the number of alternate alleles [0, 1, 2]. An extension to this approach may fit an additional covariate to represent a dominance deviation effect (*δ)* (Falconer 1960). The dominance covariate encodes genotype classes as [0, 2p, 4p-2], where *p* represents the frequency of the alternate allele and the standard-additive estimate $\beta$ is preserved in the two-effect model

($\beta$, $\delta$). The non-additive GWAS studies by Zhu et al. (2015) and Jiang et al. (2019) apply models incorporating that approach, represented by a design matrix, $\mathbf{X_1}$ (1).

An alternative model designation to identify non-additive effects may fit an additive effect ($a$, covariate encoded [0, 1, 2]) and a dominance effect ($d$, covariate encoded [0, 1, 0]), as applied in the current study. This approach can be represented by a slightly different design matrix, $\mathbf{X_2}$ (2). A third model may fit the three genotypes as class effects ($G_{11}$, $G_{12}$, $G_{22}$ encoded [1, 0, 0], [0, 1, 0], and [0, 0, 1], respectively) and can be represented by the design matrix, $\mathbf{X_3}$ (3).

These three formulations all have rank 3 and are equivalent such that any one of the formulations can be fitted and the estimated effects transformed to generate the estimated effects that would have been obtained from any of the other models (Henderson 1985). Considering (3) as the 'base model' because of its simplicity, we can use knowledge of equivalence (4) to derive effect estimates from formulations (1) and (2) via (5).

Where $\mathbf{T}$ represents a design matrix such as $\mathbf{X_1}$**,** or $\mathbf{X_2}$, $\boldsymbol{b}$ is a vector of effect estimates for the relevant model formulation, $\mathbf{X_3}$ is as defined in (3), and $\boldsymbol{g}$ is the vector of genotype class means in the 3rd model. $\mathbf{T^{-1}}$ is a matrix of contrasts used to transform effect estimates from the genotype class effects model (3) to either alternate model. The matrices of contrasts used to calculate the effect estimates from the first two models are:

$$X_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 1-p & 1-2p & p \\ -0.5 & 1 & -0.5 \end{bmatrix} \qquad X_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 0 & 0.5 \\ -0.5 & 1 & -0.5 \end{bmatrix}$$

Here, we observe the dominance contrast ($d$) and the dominance-deviation contrast (δ) are identical [-0.5, 1, -0.5], indicating the dominance effects estimated between the models implemented in other human(Zhu et al. 2015) and cattle (Jiang et al. 2019)non-additive GWAS are the same as that implemented here. However, the standard-additive ($\beta$) and additive ($a$) effects are different between the two models, where $\beta$ is the same as that derived from an additive-only model, while $a$ represents the difference between the homozygote genotype classes ([1-p, 1-2p, p] vs [-0.5, 0, 0.5], respectively). It can be observed that $\beta$ fluctuates with allele frequency, while $a$ does not, and the effects are related as $\beta = a + (1 - 2p)d$.

Vitezica et al. (2013) termed these two alternative implementations as either the 'classical' parameterisation (1), or the 'genotypic' model (2), noting that the former is more relatable to pedigree based concepts and allows direct estimates of breeding values, while the latter has more intrinsic biological meaning (i.e. more readily represents the biological effects of mutations). One obvious advantage to the classical model is the preservation of β, meaning that both standard-additive and dominance estimates are generated from the same model. The 'genotypic' model on the other hand allows straightforward calculation of a dominance coefficient K, which informs on the likely biological mechanism of the presented mutation (where $K = d / |a|$). If $K \approx 1$, a completely dominant or recessive mechanism is suggested, and if $0 < |K| << 1$, partial dominance is suggested (as observed for the PLAG1 locus effect presented in this paper). Model selection should therefore incorporate experimental aims and interpretation considerations, though both models are likely fit for purpose if dominance estimates are of primary interest.

## 2.6 Supplementary Methods

Content presented here details expanded descriptions of methods and approaches described in the main manuscript.

### 2.6.1 Power calculations

We compared additive, recessive, and genotype class effect models at different sample sizes for cows and sires for a hypothetical recessive mutation. Here, we stipulated the heritability of the trait, the number of un-genotyped daughters within each sire family (for sire models), minor allele frequency of the causative mutation, and the proportion of phenotypic variance attributable to the mutation. The power of detection was then contrasted by assessing the variance explained by each model, based on a non-central F-distribution using a critical value equivalent to a p-value of $5x10^{-8}$.

### 2.6.2 Additional animal populations

In addition to the five major-use populations described in the main manuscript (See Methods and Supplementary Table 11), two additional populations were used to perform other, more limited analyses. These included a dataset of 98,002 animals genotyped on an Illumina XT low-density parentage testing SNP-chip (683 variants), used to investigate LD between the *DPF2* and *MUS81* mutations physically genotyped on that chip. These animals are referenced only in the Supplementary Note (above) and were relatively poorly characterised - comprising mostly 2020-born calves without phenotypes. Estimation of the total number of animals homozygous for the key mutations of interest was conducted using a population of 2,799,022 animals that similarly sits outside the five major populations described in the main manuscript. The majority of these animals lacked genotypic information, though were represented by sufficient pedigree and sire genotype information to support those analyses. Further demographic information on these two populations is given in Supplementary Table 11.

### 2.6.3 Phenotypic analysis

In addition to phenotypes used for GWAS, a variety of other traits were assessed in the discovery and validation populations. These included first lactation traits such as milk (L/lactation; a lactation refers to a standardised 268 day lactation), fat (kg/lactation), protein (kg/lactation), as well as 13 traits other than production. These non-production traits included adaptability to milking (how quickly an animal settled into the milking routine), shed temperament (measure of placidity), overall opinion (farmer's subjective opinion of an animal), capacity (strength and depth of chest), rump angle, rump width, legs (leg straightness), udder support (strength of suspensory ligament), front udder (strength of attachment to the body), front and rear teat placement (position of front and rear teats on the udder), udder overall (an amalgamation of all udder traits), and dairy conformation (overall conformation). More detail on non-production traits can be found in the 'Evaluation system for traits other than production' booklet (Advisory Committee on Traits Other than Production 2020). Beyond the phenotypic model considerations detailed in the main manuscript, other trait-specific adjustments were made to data prior to genetic analysis. For lactation data, these included effects for stage of lactation, record type (lactation traits may be recorded at am milkings, pm milkings or both), and effect of induced calving. Body weight data were standardised using a multiplicative factor to adjust body weights to a mature equivalent which allowed for our admixed breed population. Since individuals had multiple body weight records across lactations, that information was utilised as previously described (Garrick, Taylor, and Fernando 2009), namely by calculating the mean of an individual's yield deviations and weighting this record through the manner in which the diagonals of the inverse residual variance-covariance matrix were constructed. Phenotypic deviations for traits without multiple records were obtained by adjusting measured phenotypes for estimates of the fixed effects derived in the national genetic evaluation models above. Body weight measurements for the 568 pedigree-identified, prospectively genotyped calves were obtained using electronic scales under contract by AsureQuality (Auckland, New Zealand).

For phenotyping conducted as a part of the research farm studies, a number of other, more detailed measures were made. Growth rate data were derived from measurements made using electronic cattle scales, with weighing performed on a monthly basis for three months for the *GALNT2* study, and on a fortnightly basis for 12 months for the *PLCD4*, *FGD4*, *DPF2*/*MUS81* study. Additional anatomical phenotypes were gathered on the latter group, including spine length (distance from the intervertebral gap between T1 and T2, to the intervertebral gap between the hips), tibia length (superior tip of the tibia to tip of the tuber calcaneus), stature (height measured just anterior of hip bones), and chest circumference (measured immediately posterior to the front legs). Since a candidate mutation for the chr29:44Mbp QTL included a missense variant in *DPF2*, a gene presenting finger and toenail phenotypes when mutated in humans (Vasileiou et al. 2018), measurements of hoof anatomical characteristics were also conducted for the relevant subset of animals. These measurements included the length of the claw (distance from the hairline to toe-tip for the outer claw), and the lateral toe angle (dorsal slope of the toe). A qualitative assessment of hoof confirmation in *DPF2* homozygotes and control animals was also performed by a veterinarian specialising in lameness, and a qualified hoof trimmer/inspector. For investigation of nerve pathology anticipated as a consequence of the *FGD4* splice site mutation, peripheral nerves (sciatic, femoral, saphenous, ulna, median and common digital nerve of the fore and hind limb) from two *FGD4* homozygotes, and two wildtype animals were dissected immediately post mortem and placed in 10% neutral buffered formalin until processed for histology. The nerves were further dissected and sections placed in a tissue cassette, after which they were dehydrated in graded alcohol and embedded in paraffin wax. Sections (3 µm) were cut and stained with hematoxylin and eosin (HE) and luxol fast blue (LFB).

Blood biochemical and metabolite measurements were derived on all research farm calves. To help assess the reproducibility of these measures, tests were conducted at multiple points in time, with blood samples for these analyses taken on three separate occasions for the *GALNT2* study (Dec 13th and 16th 2013, Feb 25th 2014), and two separate occasions for the *PLCD4*, *DPF2*/*MUS81*, *FGD4* study (18th Feb and 24th June 2019). Most tests were performed by Gribbles Veterinary (Hamilton, New Zealand), comprising commercial veterinary diagnostic assays to measure serum albumin, beta hydroxybutyrate, total calcium, creatinine, phosphate, urea, total T4 (thyroxine), triglycerides, nonesterified fatty acid concentrations, plasma glucose concentrations, and serum gamma glutamyl transferase, glutamate dehydrogenase activities. Further testing was also conducted under contract by the Liggins Institute (The University of Auckland, New Zealand) to measure the concentration of insulin-like growth factor 1 (human IGF-1 kit assessed on a Cobas e411 Immunology Analyser), insulin (Mercodia Bovine Insulin ELISA), and high and low density lipoprotein assays (using Roche Diagnostics homogeneous enzymatic colorimetric assays analysed on a Cobas c311 analyser).

### 2.6.4 RNA sequencing, sequence informatics and genome assembly considerations

Genome sequence data processing is described briefly in the main manuscript, and more comprehensively, in prior publications (Littlejohn et al. 2016; Littlejohn, Henty, et al. 2014). For analysis of RNA sequence data, reads were first processed using Trimmomatic (version 0.39)(Bolger, Lohse, and Usadel 2014) to remove leading and trailing low-quality bases, and then mapped to the ARS-UCD1.2 reference genome (GCF_002263795.1). To minimise mapping biases for downstream gene expression analyses, 14.54 million sites in this reference had been masked to replace known variant bases with nucleotides that matched neither the reference nor alternative allele. Masked bases represented sites from whole genome sequence data imputed in a subset of 99 of the 389 RNA-sequenced animals (representing cows investigated as part of a separate, as yet unpublished, study), filtered to remove sites with MAF<1%, and a DR2 <0.9. Mapping was performed using STAR (version 2.7.0)(Dobin et al. 2013), and conducted in two stages. First, reads were mapped using the RefSeq annotations to identify novel splice junctions, and subsequently, a second round of mapping was performed to incorporate these novel junctions along with those defined in the reference annotations.

To obtain liver samples for differential gene expression analysis as part of investigation of the *GALNT2* mutation, biopsies were obtained using an established protocol (Lucy et al. 2009), comprising incision and needle penetration of an area of the right rib cage at the 11th intercostal space, having first sterilised the skin surface and applied local anaesthetic. The resultant ~200 mg samples were snap frozen in liquid nitrogen with total RNA subsequently extracted using a Qiagen RNeasy kit (Qiagen). Libraries were sequenced on an Illumina Hiseq 2000 instrument using 100bp paired end reads to an average yield of 40.7M read pairs per sample. Reads were also mapped using STAR, based on the same masked version of the ARS-UCD1.2 reference genome referenced above. To generate transcript read counts for differential expression analysis, a custom script was used to count reads splicing over junctions from gene structures captured in RefSeq annotation release 106. These data were then used to perform case/control analysis using DESeq2 (Love, Huber, and Anders 2014), with genes considered significant at $\alpha=0.05$ after accounting for multiple hypothesis testing using the multiple testing procedure implemented in that software.

### 2.6.5 Utilisation of multiple genome assemblies

Most analyses presented in the main manuscript reference the UMD3.1 bovine genome assembly, though several components of data, results, and positional information are also presented based on the newer ARS-UCD1.2 assembly. Presentation of information representing both references reflects the availability of data resources at the time the analyses were performed, and the desire to use the latest assembly where possible. For the avoidance of ambiguity, however, we clarify the usage of these reference assemblies as following: Except where specified, genome positions reference the UMD3.1 assembly. Genome sequence alignments, and population-wide genotypes that were imputed from these data were also based on the UMD3.1 assembly. Candidate mutation identification and enrichment analyses performed using these genotypes were based on the corresponding UMD3.1 transcript annotations. ARS-UCD1.2 positions were located for these ~16m UMD3.1-derived variants, using the NCBI Genome Remapping Service (https://www.ncbi.nlm.nih.gov/genome/tools/remap; ~99% successfully repositioned and reported as supplementary data). Alongside UMD3.1 transcript annotations, ARS-UCD1.2 annotations are also reported for the seven candidates of primary interest (Supplementary Table 2). RNA-seq alignments were performed by mapping against the

ARS-UCD1.2 assembly (see previous section for detail). Here, splicing and gene expression phenotypes were derived based on ARS-UCD1.2 annotations, though association analyses were performed using UMD3.1 imputed genotypes in line with all other association analyses presented in the paper.

### 2.6.6 Genotypes and imputation

Most genetic analyses in the manuscript utilised animals genotyped using SNP-chips, though for prospectively genotyped calves and the cohorts used for detailed phenotypic analysis, specific mutations of interest were targeted using single variant assays. This strategy included genotyping of the *PLCD4* g.107313998G>A and *FGD4* g.77632752C>T mutations using AgriSeq assays (ThermoFisher), a method that uses sequencing of multiplexed amplicons on the Ion Torrent platform (Thermofisher). Custom Taqman assays (Thermofisher) were used to interrogate the *PLCD4* g.107313998G>A, *FGD4* g.77632752C>T, *DPF2* g.44213160A>G, *MUS81* g.44645469G>T, and *GALNT2* g.1312334G>A variants in the research farm studies, with DNA extraction and genotyping performed by GeneMark.

Imputation of sequence-resolution data in SNP-chip genotyped animals has recently been described in detail (Jivanji et al. 2019). Briefly, imputation consisted of a stepwise procedure to first unify genotype content across SNP chip platforms, yielding an ‘all animals imputed to the content of all panels’ dataset. Prior to genome sequence imputation, the sequence reference was phased using Beagle 4 software (Browning and Browning 2009). In this step, the allelic $R^2$ (AR2) metric was used to identify and remove variants with substandard phasing metrics from the reference (<0.95 AR2), where AR2 indicates the squared correlation between the predicted allele dosage and the true allele dosage (Browning and Browning 2009). This filter yielded 19,320,540 whole genome sequence variants, with this reference then used to impute the target animals using Beagle 4 (Jivanji et al. 2019; Browning and Browning 2009). A step that recovered physically genotyped, custom content representing protein-altering variants of key interest was also applied as an exception to the AR2>0.95 filter. This step, and additional study-specific genotype filtering criteria are detailed in the main manuscript.

A subset of 34,738 markers from the Illumina BovineSNP50k panel were used to perform population stratification adjustment in the GWAS discovery cohort (see 'Accounting for population stratification' section, below). These variants represented an 'all animals imputed to the BovineSNP50k panel' dataset, comprising both imputed samples and individuals physically-typed on the BovineSNP50k platform (Jivanji et al. 2019). These data represented content that had been quality-filtered according to Mendelian concordance criteria, a minimum MAF of 0.02, deviation from Hardy-Weinberg equilibrium (excluding variants with $p < 0.15$, calculated within breed), and LD pruning criteria (variants with R2 > 0.9 removed)(Jivanji et al. 2019).

### 2.6.7 SNP-based heritability estimation

We estimated the additive and dominance heritabilities for animals within the discovery population for bodyweight, stature, and body condition score. To avoid complications in estimation across breeds, heritabilities were calculated for Holstein-Friesian and Jersey purebred animals separately (N=12,149 and 7,502 animals respectively). Genetic relationship matrices (GRM) were constructed using the same quality-filtered 34,738 SNP from the BovineSNP50k dataset referenced above, with heritabilities subsequently estimated using the restricted maximum likelihood (REML) approach implemented in GCTA (Yang et al. 2011).

### 2.6.8 Association analysis

*GWAS*

In addition to the GWAS methods described in the main manuscript (Methods), further details describing the practical considerations, population stratification adjustment procedures, and leave one segment out approach are expanded upon here.

*Practical Considerations*

We applied Markov chain Monte Carlo (MCMC) based Gibbs sampling methods to draw plausible samples for genotype and marker effects described in (1, Methods). In this way, a Gibbs sampler would generate a sample $\boldsymbol{b}$ (i.e. $\tilde{b}$) given a sample for $\boldsymbol{\alpha}$ (i.e. $\tilde{\alpha}\_$) based on the model equation:

$$[y - M\_\tilde{\alpha}\_] = Tb + e$$

where $\tilde{\alpha}\_$ is a Gibbs sample from the model equation:

$$[y - T\tilde{b}\ ] = M\_\alpha\_ + e$$

However, the computational effort required to make inference from the MCMC samples obtained from Gibbs sampling of each of these 16 million models is substantial. To reduce that effort, Gibbs samples of the genotype class effects were obtained conditional on samples of the marker effects from the rest of the genome, but the samples of the marker effects from the rest of the genome were obtained conditional on a joint sample of all the marker effects from the segment of interest rather than on the sequence variant from the segment of interest.

*Accounting for population stratification*

Samples for the marker effects from the whole genome were Gibbs samples based on the BayesC0 algorithm implemented in GenSel using standard priors (Fernando and Garrick 2013). The sampler generates a Markov chain of plausible SNP-chip marker effects and variance components conditional on all else. This step can be represented as:

$$y = 1\mu + M\alpha + e$$

where equation terms are as previously described in (1, Methods), and $\boldsymbol{\mu}$ is the fixed effect representing the overall mean, and ***1*** is a vector of ones. The set of 34,738 SNP-chip markers were used to sample marker effects that represent the genomic relationships between individuals in the sample and can be used to adjust for population stratification (Kang et al. 2008). For each growth and developmental trait, GenSel was run with a chain length of 30,000, a burn in of 5,000, and thinning of 50. Thus 500 sets of MCMC samples were used for testing the conditional effect of the sequence variant class given the marker effects from the rest of the genome. Convergence of the Markov chain was established using the Geweke diagnostic such that when >95% of variables appeared stationary, the multi-variate chain was considered converged (Geweke 1991).

*Leave one segment out (LOSO)*

To avoid fitting SNP-chip markers in high linkage disequilibrium concurrently with the sequence variant being tested, we adopted a LOSO method (Yang et al. 2014). Our LOSO method involved dividing each chromosome into a series of overlapping segments beginning every 5Mbp and spanning 10Mbp for a total of 503 segments genome-wide. The 10Mbp left-out segment was selected as that best centred around the sequence variant being tested, and the columns of M (1, Methods) corresponding to BovineSNP50k

SNP-chip markers in this segment were correspondingly deleted to obtain M_. The overlapping segments allowed for a minimum distance of 2.5Mbp between any SNP-chip marker used in the BayesC0 step and any sequence variant being tested.

### 2.6.9 Single locus models

Linear mixed models (see 3, Methods) were used to analyse several datasets including phenotypic deviations, gene expression levels, and bodyweights from prospectively genotyped calves and animals in the research farm study. In this approach, we used a single site Gibbs sampler to apply a Markov chain Monte Carlo method for covariate effects, breeding value effects, and permanent environment effects as implemented in the Julia package, JWAS (Cheng, Fernando, and Garrick 2018). For each marker-phenotype combination, we used similar criteria for convergence to that applied for GWAS.

Here, we elaborate on specific model characteristics for each dataset. For phenotypic deviations, we fit genotype class effects as fixed effects and tested differences between genotype classes by directly calculating differences from the samples in their respective Markov chains. For analysis of splicing efficiency substitution effects pertaining to mutations predicted to disrupt splicing, we made inference for each marker on the expression of each intron for the *FGD4* and *GALNT2* transcripts. For the bodyweight analyses, we tested a case-control class effect of the homozygote alternate genotype. For the prospectively genotyped calves' bodyweights, we included contemporary group, age at weighing, and breed as covariates. For the research farm study bodyweights, we included date of weighing as a class effect, age x date and age x variant as interaction terms, and omitted the Z**u** random term from the model (3, Methods).

### 2.6.10 Functional enrichment and annotation-based analyses

Following identification of 356 statistically plausible causative variants for all non-additive QTL ($R^2$ >0.9 with the top-associated variant per body-weight locus), we aimed to test whether these variants were enriched for protein-altering mutations. To this end, permutation analysis was performed by randomly sampling 356 variants from all markers tested during GWAS, stratifying the selection of variants within 5% MAF bins representing the 356 -linked variants (0-1%, 1-5%, 5-10%, 10-15% etc.). We permuted this process 10,000 times to obtain a null distribution of the number of nonsense and missense variants, and then calculated an empirical p-value (North, Curtis, and Sham 2002) for our observed nonsense and missense variant count. This same process was also repeated for the 139 COJO-selected variants from the standard-additive GWAS, highlighting 3,926 linked variants for permutation analysis (Supplementary Table 1). Other bioinformatic and annotation-related analyses were also performed to predict functional consequences for mutations of interest. These included alignment of protein sequences representing other vertebrate species to visualise evolutionary conservation, identification of mutation-implicated protein domains, and retrieval of SIFT (Ng and Henikoff 2003) and Functional-And-Evolutionary Trait Heritability (FAETH) (Xiang et al. 2019) scores (variously represented in Figure 2 and Supplementary Table 2).

### 2.6.11 LD calculations at the FGD4 locus

Since the recessive QTL represented by the *FGD4* splice mutation mapped moderately close to a major known milk composition locus in our population (Lopdell et al. 2019), we determined the extent of LD between variants representing these signals in the discovery population. Here, pairwise LD was calculated between the *FGD4* candidate mutation (chr5 g.77632752C>T), and two variants respectively representing the previously reported milk yield (chr5 g.75685770A>C; rs208473130), and milk protein percentage (chr5 g.75651326G>A; rs208375076) QTL (Lopdell et al. 2019) (both yielding values of R2=0.018).

### 2.6.12 Population-based estimates of homozygote frequencies

We used pedigree records in conjunction with genotype information from 5,550 sires to estimate the historical number of animals born each year that would have been homozygous for the recessive mutations of key interest (i.e. *PLCD4*, *FGD4*, *MTRF1*, *GALNT2*, *DPF2*/*MUS81*, or *MYH1* variants). These estimates were based on data from 2010-2019, where pedigree information was first used to identify all individuals eligible for analysis, filtering to exclude animals whose sire and maternal grandsire did not have imputed genotypes, and/or whose dam and maternal grand-dam did not have recorded breed details (N=2,799,022 animals for all years retained). Using the allele frequencies reported in the discovery population (Table 2.1), we then calculated the probability of being homozygous for the recessive mutations of key interest for all eligible animals per year. This probability $P_i$ was calculated using the equation below, where $a_{S,i}$ is the sire genotype (coded 0, 1, 2 for the number of mutant alleles), $a_{MGS,i}$is the maternal grand-sire genotype, and $p_{D,i}$ and $p_{MGD,i}$ are the breed-specific population allele frequencies of the dam and maternal grand-dam, respectively.

$$P_i = \frac{a_{S,i}}{2} \times \frac{a_{MGS,i} + p_{D,i}(1 + \ p_{MGD,i})}{4}$$

$$N = \sum\nolimits_i P_i$$

These probabilities were then summed to calculate the expected number of homozygous individuals ($N$) among all eligible animals. We then used breed composition information of the national herd to extrapolate these estimates, thereby estimating the number of affected animals born per year in the entire New Zealand dairy population. These estimates were averaged across the ten years and are reported in Supplementary Table 3.

### 2.6.13 Analysis of homozygous depletion

Homozygous depletion was assessed for each candidate causative mutation within the purebred population of its respective breed of origin, as in Charlier et al. (2016). We used 29,771 purebred Holstein-Friesian, or 19,344 purebred Jersey cattle for this analysis, using a standard likelihood ratio test:

$$LRT = 2\ln\left(\frac{L|H_1}{L|H_0}\right)$$

Where

$$L|H_1 = \left(\frac{n_{mm}}{n_{mm} + n_{mw} + n_{ww}}\right)^{n_{mm}} \times \left(\frac{n_{mw} +\ n_{ww}}{n_{mm} + n_{mw} + n_{ww}}\right)^{n_{mw}+n_{ww}}$$

and

$$L|H_0 = \left(\frac{2n_{mm} +\ n_{mw}}{2(n_{mm} + n_{mw} + n_{ww})}\right)^{2n_{mm}} \times \left(1 - \left(\frac{2n_{mm} +\ n_{mw}}{2(n_{mm} + n_{mw} + n_{ww})}\right)^2\right)^{n_{mw}+n_{ww}}$$

and $n_{xx}$ is the number of individuals within each genotype class (m: mutant allele, w: wild-type allele). LRT was assumed to have a $\chi^2$ distribution with one degree of freedom.

### 2.6.14 Inbreeding analysis

Inbreeding depression describes the negative impact of inbreeding on an animal's phenotype, and can be derived through the use of an animals' inbreeding coefficient.

These coefficients are computed through analysis of pedigree information to characterise the probability that the alleles at any random genomic locus are identical by descent. To estimate the potential contribution of the non-additive QTL to inbreeding depression of bodyweight, we calculated inbreeding coefficients and formulated mixed models that did or did not include genotype classes of the eight non-additive QTL, as represented by equation 4 & 5 in the main manuscript. For ease of reference these models and terms are repeated here:

$$\mathbf{y} = 1\mu + \mathbf{F}b_1^R + \mathrm{M}\boldsymbol{\alpha} + \mathbf{e}$$

$$\mathbf{y} = 1\mu + \mathbf{F}b_1^F + \mathrm{X}\boldsymbol{b_2} + \mathrm{M}\boldsymbol{\alpha} + \mathbf{e}$$

Terms are as described in the GWAS section (1) with the following additions; **y** is a vector of phenotypic deviations for body weight (pre-adjusted for effects described in 'Phenotypic Analysis' including pairwise heterosis), $b_1^R$ and $b_1^F$ are the regression coefficients of bodyweight on inbreeding representing the reduced and full models respectively, **F** is a vector of pedigree-derived inbreeding coefficients omitting animals where F = 0, $\boldsymbol{b_2}$ is a vector of genotype class effects for the recessive QTL, X is a design matrix relating records to additive and dominance effects of non-additive QTL. $\mathbf{F}b_1^R$ and $\mathbf{F}b_1^F$ are the respective vectors of the inbreeding depression of body weight for each individual. $b_1^R$, $b_1^F$, and $\mathbf{b_2}$ are fixed effects, where $\boldsymbol{\alpha}$ are random. To assess the impact of the recessive QTL on the effect of inbreeding, we compared $b_1^R$ and $b_1^F$, where the proportion of variance explained was calculated as:

$$Variance\ Explained = \frac{var(Fb_1^R) - var(Fb_1^F)}{var(Fb_1^R)}$$

Where $b_1^R$ and $b_1^F$ represent the posterior means of each parameter, and $var(Fb_1^R)$ and $var(Fb_1^F)$ are the variance of the vector computed from $Fb_1^R$ and $Fb_1^F$, respectively.

To assess the impact of the dominance term in these models, we repeated this analysis fitting the same 8 QTL tag variants solely as additive effects. We also sought to determine how the variance explained by the 8 non-additive QTL differed from that of a randomly selected set of variants. Here, we randomly sampled sets of 8 variants from the GWAS dataset, stratifying selections within MAF bins in an approach similar to the method described in the 'Functional enrichment and annotation-based analyses' section, above. The variance attributable to these selections was then assessed applying the equations above, with the analysis permuted 20 times.

### 2.6.15 Simulation

Simulation analyses were performed to investigate the impact of MAF and effect size on sensitivity of QTL detection, and judge the utility of our association model more broadly. To these ends, we simulated a population using QMSim software (Sargolzaei and Schenkel 2009), specifying a total of 800k SNP and 750 standard-additive QTL distributed across all 29 autosomes. The various population parameters and simulation protocol was guided by that described in Brito et al. (2011), with the exact procedure described hereafter: First, a historical population of 1000 animals was simulated for 500 generations, and subsequently reduced to 200 individuals over a further 100 generations to simulate a bottleneck. This population was then increased to 20,000 over 60 generations to simulate breed expansion. In the next simulation step, 20 recent generations were simulated by selecting 200 males and 10,000 females from the historical population. Selection was performed using BLUP EBVs on a single simulated trait with a heritability of 0.3 and phenotypic variance of 1.0. Sires and dams were randomly mated, with each dam producing one progeny per year, with a replacement rate of 20% for dams and 60% for sires. To approximate the genetic composition of the bovine genome, we simulated 29 chromosomes totalling 2333 cM. A total of 800,000 SNP markers were generated to be evenly distributed across the chromosomes, while 750 QTL were randomly distributed with effects drawn from a gamma distribution with shape parameter equal to 0.4. A mutation rate of $10^{-5}$ was used for markers and QTL. The parameter and seed files used for simulation are available as supplementary data.

We then used this population to extract genotypes and phenotypes for individuals born in the final 8 generations totalling 80,000 animals and 800,000 markers. We randomly selected 30 markers from this dataset (MAF=0.01-0.05), assigning a complete recessive effect to each - thereby defining as recessive causative mutations. These markers were assigned either a 0.5 or 1.0 standard deviation effect on phenotype, where we subsequently applied our GWAS methods to attempt to detect these (and other additive) simulated QTL. Note that the marker density used was sparse compared to that derived on our real population (i.e. 800k variants versus 16m). While this might be anticipated to influence detection of additive loci, LD effects were not of concern for detection of recessive mutations since these were selected directly from the pool of 800k genotypes (i.e. the causative mutations were captured directly).

# Chapter 3 GWAS model development and application

Foreword to Chapter 3

This Chapter is a supplement to the text presented in Chapter 2. It describes and explains additional work involved in the development and implementation of the GWAS model prior to its applications in Chapters 2 and 5. The text begins by exploration of which model would be best for detecting non-additive effects, especially completely recessive effects, and we discuss the implementation of this model using the Julia programming language.

## 3.1 Abstract

Genetic analyses such as genome wide association studies have predominantly focussed on explaining the additive genetic architecture of complex traits.  While this approach has been instrumental in improving our understanding of complex disease and providing for selection on economically important quantitative traits, these approaches have ignored the impacts of non-additive genetic mechanisms. One reason for the absence of non-additive investigation is a lack of algorithms and software designed to detect these effects. Here, we design and implement an intra-locus non-additive GWAS algorithm that can be applied across millions of sequence resolution variants in over 100,000 individuals. This tool has been used to detect non-additive QTL and identify causal mutations elucidating the genetic architecture of complex traits, as described in other chapters in this thesis.

## 3.2 Introduction

A cornerstone of genomic research is understanding how genetic variants influence phenotypes. A single biallelic causal locus may act via additive or dominance mechanisms to influence a phenotype. Additive mechanisms occur when each allele at a locus acts independently such that the substitution of a wild type allele for an alternate allele has some constant effect on the phenotype. Additivity is a commonly modelled mechanism as it represents the transmissible effect inherited from one generation to the next. Therefore, the contribution of additive loci to heritable genetic variance has been the prize of livestock breeders and genetic evaluators throughout the 20th century (Lush 1940; Falconer 1960; Lynch and Walsh 1998). Economically important additive genetic loci have been described, for example, a mutation in the *DGAT1* gene causes large additive effects on milk production traits in cattle (Grisart et al. 2002).  Other studies have discovered several hundred independent additive genetic loci contributing to human height (Wood et al. 2014; Marouli et al. 2017).

Dominance mechanisms occur when there is an interaction between alleles at a locus such that the heterozygote differs from the mean of the homozygotes. There is a spectrum of possible mechanisms which can be defined into three types of dominance: partial dominance, complete dominance, or overdominance. Partial dominance occurs when the heterozygote differs from the mean but does not equal or exceed either homozygote class. Commonly described additive loci often present partial dominance effects, for example, a mutation in the *ABCG2* gene presents a partial dominance mechanism on milk-fat percentage and milk protein percentage in cattle (Cohen-Zinder et al. 2005). Complete dominance and complete recessive mechanisms occur when the heterozygote effect equals either of the homozygote genotype classes. These are often represented by qualitative phenotypes such as the polled (hornless) phenotype in cattle where a single mutant allele results in a polled animal (Georges et al. 1993), and embryonic lethal mutants where the presence of two mutant alleles at a locus results in the embryo failing to develop normally (Charlier et al. 2016).  Overdominance occurs when the heterozygote performance exceeds that of either homozygote genotypes. Single locus examples of overdominance in mammals are rare, a relatively famous exception is the mutation that causes the callipyge muscle hypertrophy phenotype in sheep (Cockett et al. 1996; Freking et al. 2002). The muscle hypertrophy phenotype only occurs when an animal is heterozygous for the mutant allele due to the animal having inherited its mutant allele from its sire, this is termed polar overdominance (Cockett et al. 1996). Together, the accumulation of these additive and dominance mechanisms as well as inter-locus interactions comprise the genetic architecture of a trait.

Heritability, the ratio of genetic variance to phenotypic variance, is a standard metric used to describe genetic architecture. Heritability is typically described in terms of narrow or broad sense heritabilities, defined as the proportion of phenotypic variance attributable to additive genetic variance or total genetic variance, respectively. However, the numerator of broad sense heritability (total genetic variance) can also be broken down into its additive, dominant, epistatic, and gene by environment genetic components, leading to respective heritability ratios (Lush 1940; Visscher, Hill, and Wray 2008). Investigating additive heritability ($h^2$) and dominance heritability ($\delta^2$) provides insight to how quantitative trait loci (QTL) might manifest in the trait of interest.

When we wish to model the biological effect of a biallelic variant, we can consider the three genotype classes at a single locus [$A_1A_1$, $A_1A_2$, $A_2A_2$] and assign each a genotypic value [-*a*, *d*, *a*] (Figure 3.1, (Falconer 1960)).

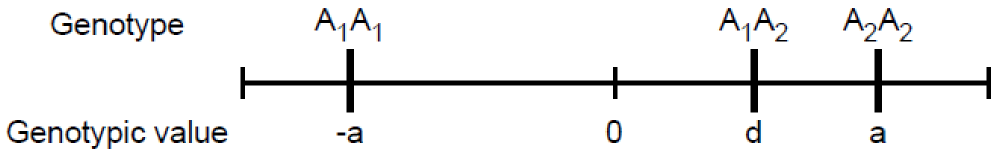


**Figure 3.1 | Assignment of genotypic values.**
Number line illustrating the assignment of genotypic values [-a, d, a] to genotype classes. This example shows a partial dominance effect (Falconer 1960).

To characterise the different intra-locus effect mechanisms, Falconer (1960) defined a dominance coefficient, k where $k = d / |a|$, that indicates the degree of dominance. If $k = 0$ the locus effect is solely additive, if $k < 1$ the locus presents a partial dominance mechanism, if k = 1 the locus presents a completely recessive or completely dominant mechanism, and if k > 1 the locus presents overdominance (Figure 3.2; (Falconer 1960)).

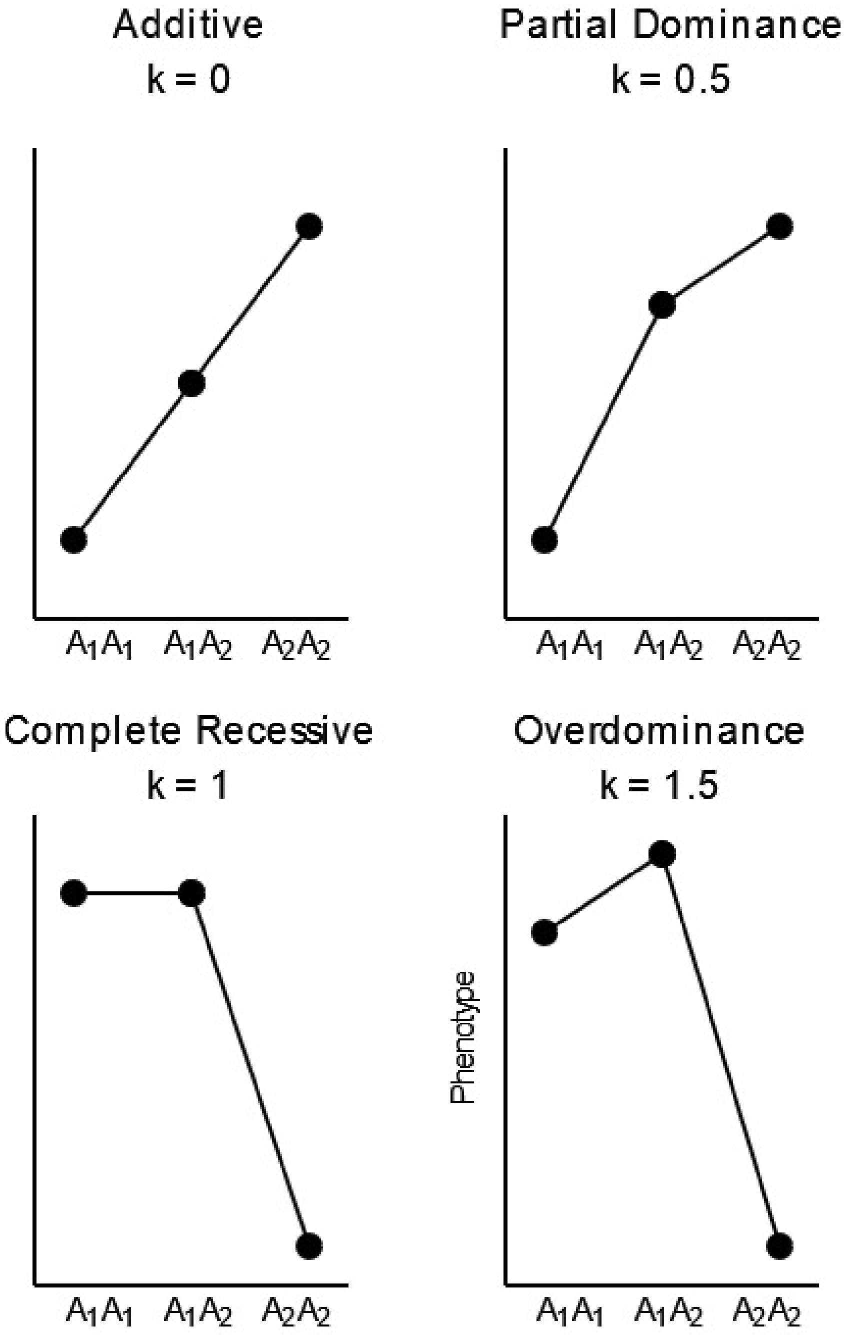


**Figure 3.2 | Diagrams of dominance genetic mechanisms.**

Diagrams comparing mean phenotypes across genotype classes (A1A1, A1A2, and A2A2) given differing genetic mechanisms. The parameter k indicates the dominance coefficient of each mechanism.

Heterosis and inbreeding depression are complementary biological phenomenon that explain how offspring can be superior or inferior, compared to the average of their parents, respectively. Non-additive mechanisms including the dominance and recessive mechanisms described here have been implicated as forming the basis of both these phenomena (Falconer 1960). The majority of the admixed dairy cattle population of New Zealand comprises Holstein-Friesian, and Holstein-Friesian-Jersey crossbreds (Livestock Improvement Corporation 2020). Both inbreeding depression and pairwise-heterosis parameters (estimated via pedigree records) are fitted in national genetic evaluation models, however the genetic basis of these parameters may present new opportunities for selection.

Genome wide association studies (GWAS) are based on a statistical approach that can be used to investigate the genetic architecture of complex traits (Risch and Merikangas 1996). GWAS can be used to model additive and dominance effects across millions of genetic variants and provide insights on the genetic mechanisms through which these variants might act. GWAS has led to a number of discoveries across species and promises to aid in the prevention and treatment of disease in humans (Visscher et al. 2017) and the improvement of selection in livestock (Georges, Charlier, and Hayes 2018).

A key aspect of robust GWAS studies is to account for potential confounding effects like population structure. Population structure encompasses the known and unknown genetic relationships between individuals in a sample and can lead to spurious associations which reflect ancestry rather than one or more causal variants (Lander and Schork 1994). Genetic relationship matrices (GRMs) fitted in a linear mixed model have been a common way to account for population structure. Such models have been implemented in software like GCTA (Yang et al. 2011), and GEMMA (Zhou and Stephens 2012) and those techniques have been shown to remove spurious associations (Eu-ahsunthornwattana et al. 2014). Other methods such as fitting random marker effects can also be used to account for population structure (Toosi, Fernando, and Dekkers 2018) and avoid the computational complexity deriving from relationship matrix construction and inversion.

Despite recognition of the existence of dominance mechanisms and their influence on quantitative traits, their contributions have often been considered to be negligible relative to additivity, and thus ignored (Hill, Goddard, and Visscher 2008; Crow 2010). It is harder to exploit dominance effects in selection schemes and genetic prediction models, so dominance effects have not been investigated to the same extent as additive effects even when non-negligible dominance genetic variance exists. Recently, there have been attempts to detect dominance QTL in human and cattle populations (Zhu et al. 2015; Jiang et al. 2019). Zhu et al. (2015) tested for the presence of dominance effects across 79 quantitative traits in humans and identified only a single partially dominant QTL at the *ABO* locus for two blood related phenotypes. In what may be the largest GWAS to investigate dominance to date, Jiang et al. (2019) tested for dominance effects based on a dataset of over 290,000 dairy cattle, and identified several dominance QTL for five milk yield and milk composition traits. While significant variants were identified, researchers were not able to identify causal mutations underlying these QTL.

There are many biological and experimental factors which contribute to our power to detect causal mutations via GWAS. As presented in Chapter 1, experimental factors that may increase power include increased sample size, higher marker density, and improved imputation quality. The manner in which the GWAS model is parameterised is a fourth experimental factor to consider when attempting to improve a models' power. Designing a fit-for-purpose model to best capture the genetic variation contributed by differing mechanisms may thus provide the power required to better elucidate the genetic architecture of complex traits.

Here, we present the development of a non-additive GWAS model designed to detect either dominant or recessive QTL. We investigate the power of differing model parameterisations at detecting a range of dominance mechanisms. An algorithm to fit this model was implemented in the Julia programming language and used in Chapters 2 and 5 to successfully detect dominance QTL affecting economically important traits in dairy cattle.  We further investigated how to account for population structure between genotyping panels and between reference genomes and explored the effect of dominance population structure on spurious associations.

## 3.3 Methods

### 3.3.1 Animal populations

Our study consisted of two interrelated datasets used to represent the same New Zealand dairy cattle population, referred to here as the 'Heritability Dataset', and the 'Application Dataset'. The Heritability Dataset was used for estimation of genetic variance and heritability, and the Application Dataset was used for reference assembly and dominance population structure comparisons.

The Heritability Dataset consisted of 12,149 cows that were reportedly 16/16ths Holstein Friesian (HF) and another 7,502 cows that were 16/16ths Jersey (J). These purebred animals were a subset of those in the admixed 'discovery population' previously described (Reynolds et al. 2021).  The Application Dataset consisted of 124,356 dairy cows from a mixed breed population described in Chapter 5 (Reynolds et al. 2022), where

20,888 are HF, 13,182 are J, 67,519 are crosses of varying proportions of those two breeds (HFXJ), and 22,767 are HF or J crossbreeds with minor proportions of other breeds.  An individual's breed composition has historically been coded in 16ths, in which case a 16/16ths animal implies it is purebred; however, this does not preclude the possibility that a distant ancestor may be crossbred.

### 3.3.2 Phenotypes

The Heritability Dataset investigated growth and developmental traits in purebred female cows, namely, live weight (kg; HF = 12,149, J = 7,502), stature (cm; HF = 10,753, J = 7,088), and body condition score (score; HF = 10,858, J = 7,148). The Application Dataset investigated the first-lactation milk-fat yield phenotype (kg/Lactation; N = 124,356). Prior to analysis, each phenotype was adjusted for nuisance effects derived from the national genetic evaluation of the entire cattle population (~30 million animals). That evaluation fitted a linear mixed model with fixed effects such as contemporary group, age at calving, and stage of lactation, pairwise heterosis, and breed group. Growth and developmental traits were adjusted for contemporary group, age at calving, and pairwise heterosis, while the milk-fat phenotype was adjusted for contemporary group, age at calving, and stage of lactation as well as other lactation specific covariates. Animals with multiple records had these aggregated to their mean and a weighting was applied reflecting the amount of information in the aggregated phenotypic deviation (Garrick, Taylor, and Fernando 2009).

### 3.3.3 Genotypes

Study animals in the Heritability dataset were genotyped on a variety of medium and/or high-density SNP-chip platforms. All missing loci were imputed first to the BovineSNP50 panel and subsequently to the BovineHD panel. For further details see Chapter 2 –

Methods (Reynolds et al. 2021). Imputed genotypes from both the Bovine SNP50k and Bovine HD panels were filtered using similar quality controls. We removed individual variants if they presented a high missing genotype rate (>0.01), a low minor allele frequency (<0.02), a high deviation from expected Hardy-Weinberg equilibrium (>0.15, calculated within breed), or a low imputation quality based on allelic $R^2$ ($AR^2 < 0.95$). Final filtering consisted of LD pruning to remove variants in high linkage disequilibrium with another on the panel ($R^2 > 0.9$). These filters resulted in a set of 34,738 markers on the Bovine SNP50k panel, and 280,570 markers on the Bovine HD panel.

Animals in the Application dataset had sequence-imputed genotypes generated as described in Chapter 5 (Reynolds et al. 2022). That dataset consisted of 16,640,294 variants for GWAS.

### 3.3.4 GWAS Model

Genome wide association studies are frequently used to find which genotypes influence phenotypes (Visscher et al. 2017). Most implementations only investigate additive effects and ignore non-additive effects like dominance. We aimed to design a linear mixed model to detect intra-locus non-additive (dominance and recessive) effects in complex quantitative traits in cattle. We intended to investigate over 16 million imputed variants, described in Chapter 2 and Chapter 5, and chose a single locus approach to test each variant one-at-a-time. While doing so we also aimed to account for the complex relationship structures of dairy cattle in the genotyped dataset to avoid spurious associations. This linear mixed model can be represented by this equation.

$$y = Tb + \mathrm{M}\alpha + \mathrm{e} \tag{1}$$

Where $\boldsymbol{y}$ represents a vector of adjusted phenotypes, $\boldsymbol{b}$ is a vector of fixed effects representing the genotype of interest, and $\boldsymbol{T}$ is the design matrix relating records to the genotype of interest. The vector $\boldsymbol{\alpha}$ comprises random effects representing population structure and the design matrix $\boldsymbol{M}$ relates records to these random effects. The vector $\boldsymbol{e}$ represents random residuals for each record. Throughout this Chapter we specify further details of these model terms.

### 3.3.5 Model Specification

While recessive causal variants were of particular interest in developing this model, we also wanted to capture other intra-locus effects such as partial dominance and over-dominance. We designed simulations to select the best encoding of $\boldsymbol{T}$ (1) to detect these causal mechanisms. We compared the computational aspects and tested the power of different encoding specifications of single locus models to this end, comparing additive, recessive, and genotype class effect specifications to select an appropriate model for detecting non-additive variation. We considered four types of mechanisms: additive, partial dominance, complete recessive, and over-dominance shown in Figure 3.2. We simulated either sire phenotypes or cow phenotypes across different sample sizes for these different mechanisms caused by a mutation. We assessed the power of each model in each simulation by calculating the variance explained using analysis of variance (ANOVA) techniques.

Three different model specifications were used to calculate the variation explained by the locus of interest. First, the additive model which fits an intercept and a covariate indicating the number of alternate alleles in the genotype (2). Second, the recessive model fits an intercept and a covariate indicating whether the genotype is homozygous-alternate or not (3). Third, the genotype class means model fits three genotype class effects indicating the individual's genotype at the locus (4).

$$\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} \quad (2)$$

$$\begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} \quad (3)$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (4)$$

### 3.3.6 Simulated datasets

We simulated mutations with the 4 different effect mechanisms presented in Figure 3.2, these are additive (k = 0), partial dominance (k = 0.5), complete recessive (k = 1), and overdominance (k = 1.5). For each mechanism, we simulated a low frequency causal variant with a minor allele frequency of 0.025. Minor allele frequency is a key parameter given that lower frequency variants are harder to detect; we chose this frequency as it is similar to that reported for a previously described *GALNT2* mutation (Charlier et al. 2016; Reynolds et al. 2021). That causal mutation explained 0.1% of the phenotypic variation.

Many economically important phenotypes are routinely measured on cows via herd-tests. A key aspect of artificial selection in New Zealand dairy cattle is the use of a small number of sires across the national herd of several million cows (Tacon 2002; Georges, Charlier, and Hayes 2018). This means understanding the phenotypes of bulls is crucial to genetic gain, however one cannot measure important cow-specific traits like milk production on bulls. To overcome this challenge, the phenotypes of a sire's daughters can be aggregated and assigned as the phenotype of the sire, and this value can then be used in selection and association models.  Here, we have modelled one such cow-specific phenotype where each cow has their own phenotype record, and each sire phenotype was calculated as the aggregate of 100 of their non-genotyped daughters. The simulation investigated sample sizes ranging from 5,000 to 100,000 cattle (or 500,000 to 10,000,000 non-genotyped daughters to simulate sire phenotypes) to attempt to determine how many animals are required to have sufficient power in detecting mutations under each mechanistic presentation. This phenotype had an additive heritability of 0.25 which is broadly similar to previously reported heritabilities for liveweight and milk volume phenotypes of 0.39 and 0.28, respectively (Sun et al. 2014; Reynolds et al. 2021).

### 3.3.7 Estimating the power of the model

The power of a statistical test is the probability of rejecting the null hypothesis when it is false.  We assessed the power of each model specification using analysis of variance (ANOVA). We tested the hypothesis that treatment group means are all equal by comparing between treatment group variation and within treatment group variation. The quotient of these variance terms generates an F-statistic which we used to calculate the power of the test using a non-central F-distribution and a critical value equivalent to a p-value of $5x10^{-8}$. We chose this significance threshold as it is a standard multiple testing threshold used in GWAS studies (Visscher et al. 2017). This analysis was implemented in R [https://github.com/egmreynolds/PowerCalculations.git] using the pwr library (Champely et al. 2020).

### 3.3.8 Fitting marker effects

Accounting for population structure is an important aspect of a GWAS model (Yu et al. 2006). This can be done using a pedigree relationship matrix, a genomic relationship matrix (GRM) or marker effects as outlined in Chapter 1 (Henderson 1976; Vanraden

2008; Toosi, Fernando, and Dekkers 2018). Relationship matrices for *n* individuals are *n x n* symmetric matrices describing the genetic relationships between each of pair of individuals. As sample size increases so does the computational complexity of inverting that relationship matrix. The effort can be alleviated by fitting marker effects for *n* individuals and *m* markers. Marker effects are represented by a *n x m* matrix describing allele distributions present in the sample, and the variance -covariance matrix of those effects is diagonal making its inversion trivial.  That was the method implemented for model design.

### 3.3.9 Heritability estimation

We first compared the genetic variance explained by two different SNP-chip platforms with different marker densities, aiming to select which marker set to fit as marker effects. Genetic variance components and heritabilities were estimated in a similar way to that described in Chapter 2.  We constructed additive and dominance GRMs using 34,738 and 280,570 markers from the Bovine SNP50k and Bovine HD panels, respectively, for each breed.  We used GCTA to estimate additive and dominance variance components and their respective heritabilities using the restricted maximum likelihood (REML) approach (Yang et al. 2011), where additive heritability ($h^2$) is the ratio of additive genetic variance to phenotypic variance, and dominance heritability ($\delta^2$) is the ratio of dominance genetic variance to phenotypic variance. To compare possible differences in dominance heritability estimation, we also used GCTA to estimate dominance genetic variance while omitting the additive GRM from the model.

### 3.3.10 Heritability comparison

We were interested in comparing heritability estimates between SNP chip panels, between breeds, and between REML models. REML estimators are asymptotically normal but their exact distribution is unknown (Searle, Casella, and McCulloch 2009; Cressie and Lahiri 1993). Their sample distribution is approximated by a mean and standard error, but these cannot be compared between estimates using Z-tests, or confidence intervals. Instead, a conservative method was used, appealing to Chebyshev's inequality (Chebyshev 1867). Chebyshev's inequality is a generalised probability theorem that produces an upper limit for confidence intervals that can be applied to any probability distribution. This generality may reduce its power but increases its utility. The inequality can be used to calculate the maximum probability that there is no difference between two probability distributions with known mean and variance.  This probability can be calculated using

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Where $X$ and $\mu$ are REML estimates from different distributions such that $X$ - $\mu$ indicates the difference between the REML estimates of each distribution, $k$ indicates the magnitude of the difference between $X$ and $\mu$, and $\sigma^2$ represents the variance of the difference of the REML estimators while assuming the covariance between the two estimators is zero. We considered a difference significant at $p \leq 0.1$.

### 3.3.11 Marker effect estimation

We chose to fit marker effects from the BovineSNP50 panel as random effects in our GWAS model and we aimed to estimate all random marker effects simultaneously. The Bayesian Alphabet is used to describe a series of Bayesian models for estimating the posterior distributions of marker effects (Fernando and Garrick 2013). These Bayesian alphabet models can be implemented using Gibbs sampling techniques where inferences on marker effects and variance components are made on plausible samples drawn from their posterior distributions (Fernando and Garrick 2013). In BayesC0, each plausible marker effect ($\alpha_i$) is drawn from a univariate normal distribution with variance $\sigma_{\alpha}^{2}$ (Habier et al. 2011; Fernando and Garrick 2013). These posterior marker effect distributions can be used to account for the population structure while testing another variant of interest (Toosi, Fernando, and Dekkers 2018). We chose this BayesC0 approach to estimate marker effects in our GWAS model.

### 3.3.12 Proximal contamination

When fitting population structure parameters derived from genomic data like GRMs or marker effects, double fitting of the variant of interest can occur and this causes what is known as proximal contamination. This is because the variant of interest (or another variant in moderate to high linkage with the variant of interest) already contributes to the GRM or marker effects, which can shrink the estimate of the effect and lead to it being falsely rejected. Leave one chromosome out (LOCO) and leave one segment out (LOSO) are two approaches to address this problem. LOCO involves ignoring variants on the same chromosome as the variant of interest. Similarly, LOSO involves ignoring variants in the same chromosomal segment as the variant of interest, where the size of the segment can vary depending on the effective population size which is determined by the population structure. We implemented the LOSO approach with 10Mbp segments as we aimed to account for as much population structure as possible while avoiding proximal contamination.

### 3.3.13 Model Implementation

We implemented the GWAS model developed in two steps to reduce computational effort. Step 1 involved estimating the effects of population structure and adjusting the phenotype using a LOSO approach.  Step 2 uses these LOSO-adjusted phenotypes to test each of 16 million genome-wide variants using the specified encoding to detect dominance and recessive effects. We also needed to implement the majority of the code ourselves such that it fit the model developed in Chapter 2 (Methods).  We chose the relatively new programming language, Julia (Bezanson et al. 2017), which specialises in fast computation and is positioned for data science and statistical and functional programming.

*Step 1*

Step 1 involved fitting a BayesC0 algorithm across a set of markers and subsequently using these markers' effects to account for population stratification.  GenSel is a software which can be used to run multiple Bayesian models including BayesC0 in an efficient manner (Fernando and Garrick 2013). GenSel also allows control of sampling approaches such as Markov chain length and burn-in amount, as well as a range of definitions of priors for variance components.  We used GenSel to sample plausible marker effects for SNP chip markers genome wide.

An important aspect of our model is that it allows phenotypes to be the aggregates of different numbers of records. For example, one individual may have been measured three times for the same trait, whereas other individuals may have only been measured once. This introduces differences in the error variance of each measurement where the record measured three-times has a smaller variance than the record measured once (Garrick, Taylor, and Fernando 2009). To account for this difference, we fit **D**, an n x n diagonal matrix where the diagonal elements vary according to the number of observations in the aggregated record (Chapter 2 - Methods).

The remainder of the processing and analysis for the two-step GWAS model was implemented in the Julia programming language. To determine whether inferences obtained from the Markov chains had converged we need to extract and interpret the marker effects from GenSel result files. Assessing the convergence of a multivariate chain is very difficult, to overcome this we tested convergence for each variable via the univariate Geweke diagnostic (Geweke 1991). The Geweke diagnostic works by comparing the distributions from the first 20% of the chain with the last 10% of the chain using a t-test. If greater than 95% of variables appeared stationary, the multi-variate chain was considered to have converged. This process allowed us to determine convergence of the marker effect Markov chains and use these data to assess population structure.

The leave one segment out (LOSO) approach required adjusting phenotypes for marker effects from all SNP chip markers except those within the relevant 10Mbp window. A key challenge to this analysis is the matrix multiplication $\boldsymbol{M\alpha}$ (Chapter 2 – (1)), where $\boldsymbol{M}$ is an $n$ x $m$ matrix, and α is an $m$ x $c$ matrix where $n$ indicates the number of individuals, $m$ the number of SNP chip markers, and $c$ represents the number of plausible marker effect samples. As $n$, $m$, and $c$ increase, this calculation becomes more computationally challenging. The variables $n$ and $m$ are based on the sample set available while $c$ can be modified. The number of plausible marker effect samples needs to be sufficiently large to represent the effect distribution of each variant. We chose c = 500 to fit this purpose as it results in a sufficiently large number of samples to summarise the distribution, while being small enough such that the computation of LOSO adjusted phenotypes remains feasible. The resulting *n x c* matrices are subtracted from the phenotypic deviations resulting in *c* LOSO-adjusted phenotypic deviation vectors of order *n* to be used in the subsequent GWAS analysis.

*Step 2*

In Step 2, we aimed to test each genome-wide sequence variant for association with the LOSO-adjusted phenotypes. Using the genotype means parameterisation (4), we estimated genotype class effects for each vector of plausible phenotypes adjusted for population structure. These genotype class means can then be converted to other substitution, additive, and dominance effects by transforming the effect estimates to those of an equivalent model (See Chapter 2, Supplementary Note; (Henderson 1985)). Matrix multiplication is a computationally demanding task and is required to calculate fixed effect estimates.  As described in Chapter 2 – Methods, if ***T*** is a design matrix relating individuals to genotype classes for a sequence variant of interest, instead of calculating ***T'DT*** we can calculate the dot product of ***T*** and ***D*** and sum the columns to create the output diagonal of the cell counts (***T'DT***).  This adjustment may save on computational resources, especially as samples sizes increase and if the model becomes increasingly complex.

Through this computation, vectors of plausible effect estimates are drawn from the posterior distributions of each sequence variant, *x*. These can be summarised by their posterior means, $\beta_x$, posterior standard deviations, $\sigma_x$, and z-statistics, $z_x$, following a standard Normal distribution as in Bernal Rubio et al 2015. The statistical significance of the genetic effects was then evaluated using a Z-test (5).

$$z_x = \frac{\beta_x}{\sigma_x}, \qquad \text{as if } z_x \sim N(0,1) \tag{5}$$

Many computers come with multiple CPUs, but most basic code only uses one of these CPUs at a time. In some cases, software can be written to utilise multiple CPUs in parallel, so the program evaluates its code faster, this is termed multi-threading or parallel computing. Here, we are testing millions of genome wide markers one at a time where markers are tested independently of each other. This provides an opportunity to use multiple CPUs to process the GWAS in less time than performing each test sequentially. On the other hand, starting up a CPU can also take time especially if it needs access to a larger dataset as it does in this case. We struck a balance between sequential and parallel computing, where up to 10,000 markers are tested on each CPU available. This allows many CPUs to run at the same time, so wall clock time is divided by the number of CPU available.

### 3.3.14 Transitioning to a new genome assembly

We used the model and algorithm described above in Chapter 2 where we investigated dominance and recessive effects on growth and developmental phenotypes measured on over 80,000 cows (Reynolds et al. 2021). During that research, the world of bovine genomics made a major transition between reference genome assemblies from UMD 3.1 (UMD, Zimin et al. (2009)) to ARS-UCD1.2 (ARS, Rosen et al. (2020)) and, in turn, we wanted to make use of this new resource. Much of the process of transitioning our dataset was undertaken by scientists in the research and development department at LIC. To use our GWAS model on the new genome assembly, we needed to create a new set of genomic positions for the marker effects on ARS, we used the ARS-based BovineSNP50 genotyping panel as the basis of this and the filtering applied to this dataset is described below. We used the milk-fat yield phenotype in the Application Dataset to assess how well the ARS marker set accounted for population structure compared to the UMD marker set (described in Chapter 2) by comparing association results of imputed to sequence genotypes on chromosome 1.

### 3.3.15 Filtering the ARS BovineSNP50 panel

We aimed to extract a new set of marker positions from the ARS-based BovineSNP50 panel. We filtered this panel in a similar way to the manner in which the UMD-based panel was filtered in Chapter 2, also presented in Chapter 5. Briefly, the 50K panel had 54,708 autosomal SNPs, we filtered these SNPs to remove markers with high missing genotype rates (> 0.01), low minor allele frequency (< 0.02), or high deviations from expected Hardy-Weinberg equilibrium (> 0.15, calculated within breed). We then removed markers that appeared to impute poorly (dosage $R^2 < 0.9$), and markers in high LD with another marker on the panel (pairwise $R^2 > 0.9$, within 1 Mbp). These conditions resulted in a set of 31,451 SNP chip marker positions.

### 3.3.16 Dominance population structure

We were interested in how dominance population structure may cause spurious associations in GWAS. In Chapter 2, we adjusted phenotypic deviations for pairwise heterosis effects derived from the national genetic evaluation models. The heterosis covariates were estimated using pedigree and breed composition records which can be inaccurate and ignore un-recorded ancestral relationships. Since we were interested in studying the dominance population structure of our sample, we chose not to adjust for pairwise heterosis in this Chapter and in Chapter 5.  We contrasted GWAS results when fitting only additive marker effects (additive-only) versus the results obtained when fitting both additive and dominance marker effects (additive and dominance). We quantified inflation by calculating the genomic control inflation factor, calculated as $\lambda_{GC} = \frac{median(Z_i^2)}{median(\chi_1^2)}$, where $Z_i^2$ represents a vector of observed squared Z-statistics from GWAS and $\chi_1^2$ is the expected chi-squared distribution with a median value of 0.455 (Devlin and Roeder 1999). We used the milk fat yield phenotype and chromosome 1 to make this comparison.

## 3.4 Results

### 3.4.1 Model specification simulation

We assessed the power to detect causal variants of differing genetic mechanisms under additive, recessive, and genotype class models. Figure 3.3 presents the power of these models at varying sample sizes for cow and sire datasets. These power calculations indicate the genotype class means model has the highest or near-highest power to capture variation across genetic mechanisms across cow simulations. We observe the power of the additive model and recessive model to fluctuate depending on the genetic mechanism of the locus. These results indicate the additive model will be much less effective at detecting complete recessive QTL and the recessive model similarly will not effectively detect complete additive QTL, but the genotype class means model will have adequate power in both such cases (given sufficient sample sizes). In sire simulations, power is typically higher than the equivalent cow models similar to results previously observed (Weller, Kashi, and Soller 1990) and calculations indicate the additive models and genotype class models have the strongest power, while the recessive model has the lowest across genetic mechanisms.

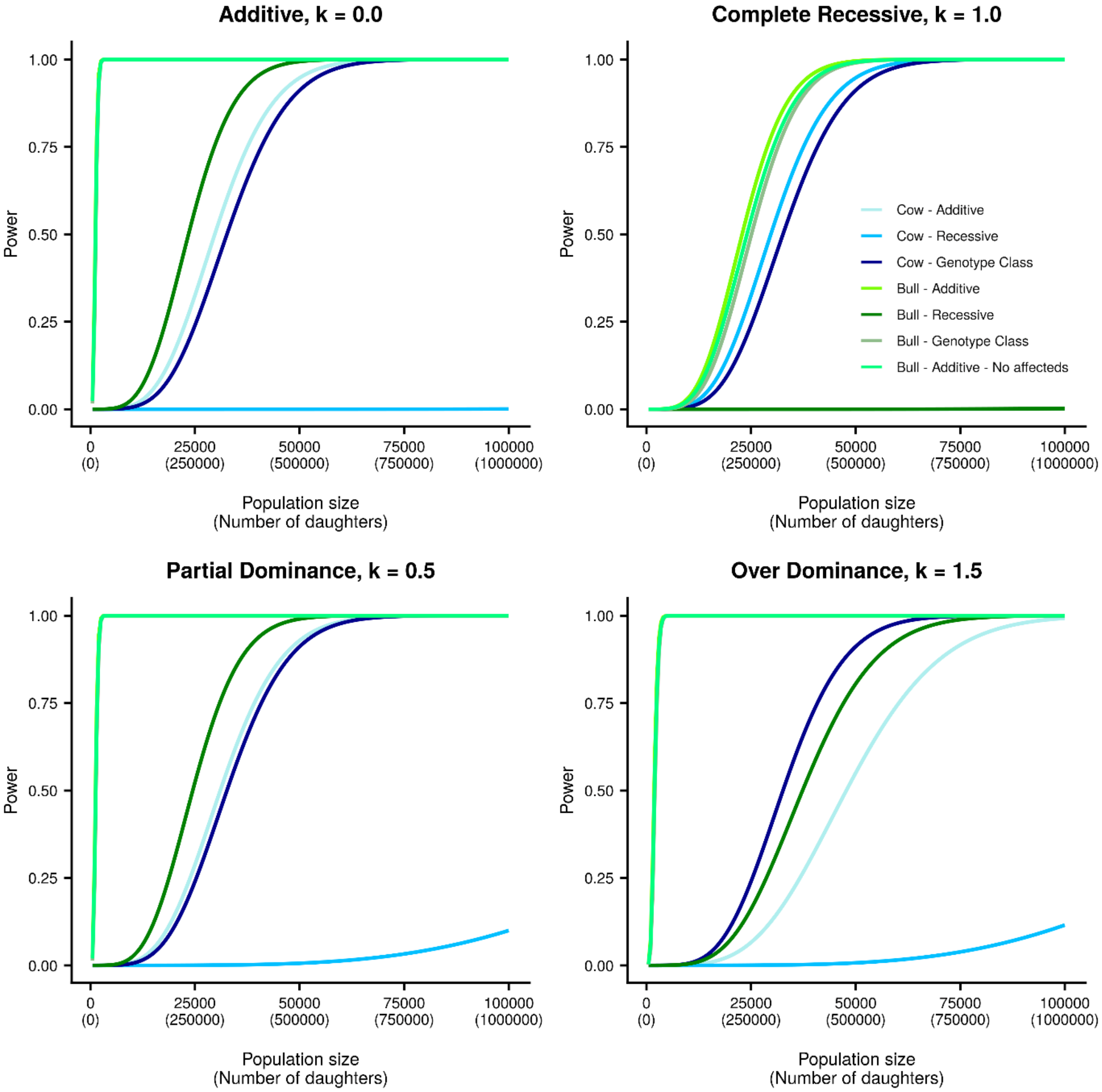


**Figure 3.3 | Power comparisons across genetic mechanisms.**

Plots comparing the power of different models across different modes of effect (additive, complete recessive, partial dominance, over dominance). Each colour indicates a different model. The parameter k indicates the dominance coefficient of the mutation. Population size indicates the number of cows or sires in the model, Number of daughters indicates the number of non-genotyped daughters simulated to generate the sire phenotypes in the model. Power indicates the probability of detection as determined via a non-central F distribution and a critical value equivalent to a p-value of 5x10$^{-8}$.

### 3.4.2 Heritability results

We were interested how much genetic variance could be attributed to additivity and dominance. The heritability estimates in Table 3.1 and Table 3.2 show additive heritabilities are typically far greater than dominance heritabilities across the growth and developmental traits investigated here. The largest additive and dominance heritabilities were 0.429 and 0.067, respectively, in Holstein Friesian liveweight. Body condition score exhibited low dominance heritabilities with estimates ranging from 0.00 to 0.026. These findings suggest additivity is the primary mode of inheritance for these traits.

We aimed to evaluate the differences in estimates of genetic variance components and heritabilities between SNP-chip panels. We used Chebyshev's inequality to calculate the maximum probability that the heritabilities estimated differed between the BovineSNP50 and BovineHD SNP-chips (Table 3.1 and 3.2). While heritabilities based on the BovineHD SNP-chip were typically numerically higher than those based on the BovineSNP50 SNP-chip, we did not find any significant differences across estimates between these panels, suggesting the 50K panel accounts for as much genetic variance as the HD panel. As such, further heritability comparison was only performed on the BovineSNP50 SNP chip.

**Table 3.1 | Heritability estimates for Holstein-Friesian cattle**

| Holstein-Friesian | 50K panel | | HD panel | | 50K vs HD | |
|---|---|---|---|---|---|---|
| Phenotype (N) | $h^2$ (se) | $\delta^2$ (se) | $h^2$ (se) | $\delta^2$ (se) | max. p ($h^2$) | max. p ($\delta^2$) |
| Liveweight<br>(12149) | 0.3904<br>(0.0131) | 0.0378<br>(0.0119) | 0.4291<br>(0.0136) | 0.0667<br>(0.0174) | 0.24 | 0.53 |
| Stature<br>(10753) | 0.3411<br>(0.0141) | 0.0488<br>(0.0147) | 0.3696<br>(0.0147) | 0.0589<br>(0.0212) | 0.51 | 1.00 |
| Body condition score<br>(10840) | 0.2497<br>(0.0138) | 0.0184<br>(0.0152) | 0.2741<br>(0.0146) | 0.0265<br>(0.0219) | 0.68 | 1.00 |

**Table 3.2 | Heritability estimates for Jersey Cattle**

| Jersey | 50K panel | | HD panel | | 50K vs HD | |
|---|---|---|---|---|---|---|
| Phenotype (N) | $h^2$ (se) | $\delta^2$ (se) | $h^2$ (se) | $\delta^2$ (se) | max. p ($h^2$) | max. p ($\delta^2$) |
| Liveweight<br>(7502) | 0.3294<br>(0.0170) | 0.0593<br>(0.0158) | 0.3554<br>(0.0176) | 0.0631<br>(0.0182) | 0.89 | 1.00 |
| Stature<br>(7088) | 0.2644<br>(0.0174) | 0.0540<br>(0.0180) | 0.2882<br>(0.0181) | 0.0247<br>(0.0184) | 1.00 | 0.77 |
| Body condition score<br>(7132) | 0.2259<br>(0.0167) | 0.0165<br>(0.0163) | 0.2406<br>(0.0173) | 0.0000<br>(0.0139) | 1.00 | 1.00 |

Table 3.1 and 3.2 present additive ($h^2$) and dominance ($\delta^2$) heritability estimates for Holstein-Friesian and Jersey cattle, respectively, for various phenotypes and the different SNP-chip panels densities BovineSNP50 (50K) and BovineHD (HD). We calculated the maximum probability (max. p) that the distributions of the heritability estimates from HD and 50K were the same as calculated using Chebyshev's Inequality.

We investigated whether breed impacted the heritability estimates (Table 3.3). Holstein-Friesian additive heritability estimates were significantly higher than Jersey additive heritability estimates for stature. However, this was not the case for additive heritabilities of the other traits or for dominance heritabilities. Investigating stature with the 50K panel, we observe Holstein Friesian $h^2 = 0.3411$ versus Jersey $h^2 = 0.2644$ while HF $\delta^2 = 0.0488$ versus J $\delta^2 = 0.0540$.

**Table 3.3 | Maximum probability of breed differences in heritability estimates**

| 50K panel | HF vs J | |
|---|---|---|
| Phenotype | max. p ($h^2$) | max. p ($\delta^2$) |
| Liveweight | 0.12 | 0.85 |
| Stature | 0.09 | 1.00 |
| Body condition score | 0.83 | 1.00 |

Table 3.3. presents the maximum probabilities (max. p) that there was no difference between the distributions of the Holstein-Friesian (HF) and Jersey (J) heritability estimators calculated using Chebyshev's Inequality.

We noted a non-zero negative sampling covariance between the additive and dominance variance components across breeds and phenotypes. To investigate whether the dominance heritabilities were underestimated due to the inclusion of the additive variance component, we calculated the dominance heritability estimates without including the additive GRM in the model (Table 3.4). Dominance heritability estimates for liveweight were significantly greater when dominance was fitted alone, than when fitted with an additive GRM for both breeds. The same was not true for analysis of stature or body condition score for which the estimates were not significantly different between the two models.

**Table 3.4 | Dominance heritability estimates.**

| | HF 50K | J 50K | HF 50k | J 50K |
|---|---|---|---|---|
| Phenotype ($N_{HF}$, $N_J$) | $\delta^2$ (se) | $\delta^2$ (se) | max. p ($\delta^2$) | max. p ($\delta^2$) |
| Liveweight<br>(12149, 7502) | 0.0989<br>(0.155) | 0.1391<br>(0.0202) | 0.10 | 0.10 |
| Stature<br>(10753, 7088) | 0.097<br>(0.0177) | 0.1145<br>(0.0212) | 0.23 | 0.21 |
| BCS<br>(10840, 7132) | 0.0288<br>(0.0174) | 0.0366<br>(0.0182) | 1.00 | 1.00 |

Table 3.4 presents dominance heritabilities ($\delta^2$) estimated within breed (Holstein-Friesian (HF), Jersey (J)) across different phenotypes. We calculated the maximum probability that there was no difference between the distributions of the $\delta^2$ estimates from the dominance-only model and the additive and dominance model using Chebyshev's Inequality.

### 3.4.3 Dominance GWAS model

The foremost aim of fitting this association model was to detect deleterious recessive alleles while also providing insight on other non-additive mechanisms. To this end, we decided to investigate cow phenotypes using the genotype class means model due to the consistent power this model provides across genetic mechanisms and the increased resolution the extra parameter provides for inferring the mechanism underlying a QTL. We developed a two-step GWAS model which fits marker effects using a LOSO approach to avoid proximal contamination, then fits each tested sequence variant one-at-a-time.

These criteria result in the following extended description of (1) as provided in Chapter 2 - Methods (Reynolds et al. 2021).

$$y = Tb + \mathrm{M\_\alpha\_} + \mathrm{e} \tag{5}$$

Where $\boldsymbol{y}$ is a vector of phenotypes for one trait, $\boldsymbol{b}$ is a vector of genotype class effects for the tested sequence variant, $\boldsymbol{T}$ is a design matrix relating records to the genotype classes at the tested sequence variant. $\boldsymbol{\alpha\_}$ is a vector of random SNP chip marker effects with coverage across the whole genome except for the 10Mbp segment of interest such that $\boldsymbol{\alpha\_} \sim \mathrm{N}\ (\boldsymbol{0}, \boldsymbol{I}\sigma_\alpha^2)$, where $\boldsymbol{I}$ is an identity matrix of order equal to the number of marker effects and $\sigma_\alpha^2$ represents the marker effect variance, $\boldsymbol{M\_}$ is a matrix obtained from $\boldsymbol{M}$ (a matrix relating records to markers (encoded [0, 1, 2])), by deleting the columns corresponding to the region neighbouring the tested sequence variant, $\boldsymbol{e}$ is an error vector with $\boldsymbol{e} \sim \mathrm{N}\ (\boldsymbol{0}, \boldsymbol{D})$, where diagonal elements of $\boldsymbol{D}$ vary according to the number of observations in the aggregated phenotypic record.

### 3.4.4 Genotypic parameterisation implemented for biological inference

Genotype class means are not as interesting as some of the contrasts between the means of different genotype classes. We used an equivalent model to overcome these challenges. A set of models are equivalent if any one of the formulations can be fitted and the estimated effects transformed to generate the estimated effects that would have been obtained from any of the other models (Henderson 1985). We considered three equivalent models, as explained in the Supplementary Note of Chapter 2 (Reynolds et al. 2021). Here, we implemented an association model with the primary aim of detecting deleterious recessive alleles, therefore we chose the genotypic parameterisation. Briefly, the genotypic parameterisation fits a genotypic additive effect ($\boldsymbol{a}$) and a genotypic dominance effect (***d***) representing the genotypic values described in Figure 3.1. Where ***a*** represents half the difference between homozygote genotype classes, and ***d*** represents the deviation of the heterozygote genotype class from the mean of the homozygote genotype classes.

### 3.4.5 Modelling association of non-additive effects package

We implemented a package called Modelling Association of Non-Additive effects (MANA) in the Julia programming language (https://github.com/egmreynolds/MANA.git, (Reynolds et al. 2021)). That software incorporates marker effect samples from GenSel and implements part of Step 1 and Step 2 of our association model utilising Julia's high performance computing. The GWAS model was run on a large computing cluster called the New Zealand eScience Infrastructure (NeSI) and our software can utilise the many CPUs available to reduce the real time it took for it to run, making it possible to rapidly investigate over 16 million variants in datasets of over 100,000 individuals.

### 3.4.6 Accounting for population structure on the ARS-UCD1.2 reference genome

Reynolds et al. (2021) details how the GWAS model was used in the discovery of novel recessive mutations affecting growth and developmental phenotypes in cattle.  This method appeared to accurately account for population structure presenting several plausible candidate causal mutations associated with genetic disorders. In transitioning from UMD to ARS, we investigated how differing marker sets and marker specifications captured population structure and accounted for spurious associations.

We observed concordance between GWAS results from the ARS-based marker set performed compared to the UMD-based marker set, suggesting our analyses on the new reference genome would account for spurious associations at least as well as the previous study presented in Chapter 2. Figure 3.4 contrasts the significance of genotypic additive and genotypic dominance effects of 396,241 imputed-to-sequence variants on chromosome 1 between UMD-based and the ARS-based marker sets. We observe strong concordance between marker effect panels. We therefore chose to use the ARS-50k set of 31,451 marker positions to account for population structure on the ARS assembly.

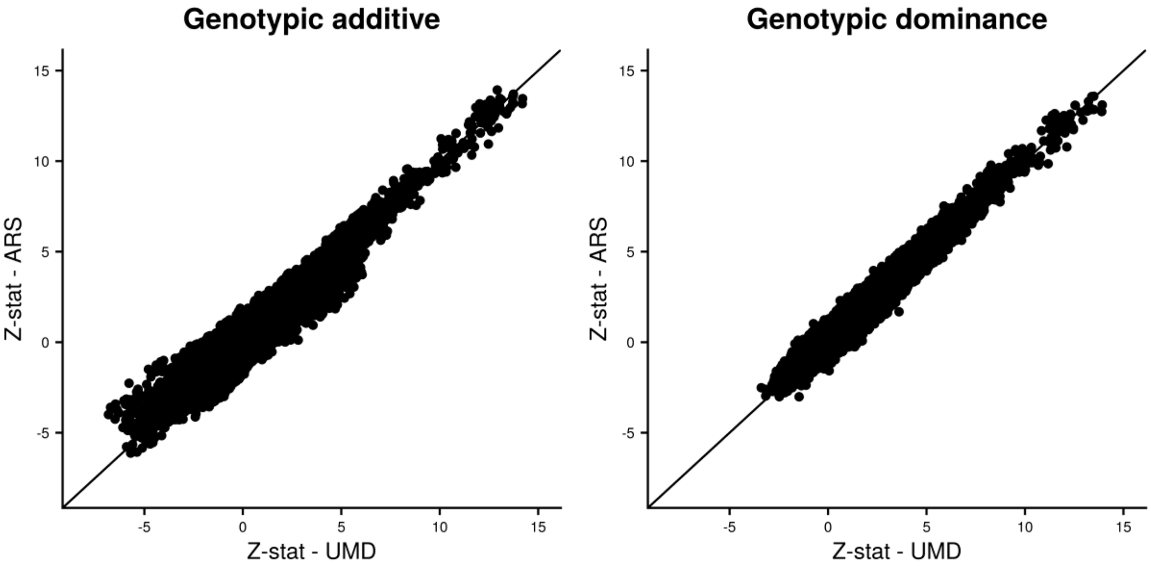


**Figure 3.4 | Scatter plots comparing genotypic effects between reference datasets**
Plots comparing significance (Z-statistics) of genotypic additive and genotypic dominance effects when using different marker sets to account for population structure.

### 3.4.7 Investigating dominance population structure

We were interested in the impact of dominance population structure and investigated how it might cause spurious associations. We performed GWAS on the milk-fat yield phenotype, which had not been pre-adjusted for pairwise heterosis, using two different scenarios for account for population structure. First, we fit additive marker effects (additive-only) for the 31,451 markers, and second, we fit additive and dominance marker effects (additive and dominance) for the same marker set. Figure 3.5 presents

GWAS results for genotypic additive and genotypic dominance effects for the two scenarios. We observe that while the significance of genotypic additive effects appeared moderately consistent across scenarios, the significance of genotypic dominance effects did not. In the additive-only scenario statistics (Figure 3.5a) appear far inflated relative to the additive and dominance scenario (Figure 3.5b). To investigate the inflation of each scenario on chromosome 1, we contrasted significance statistics against those we would expect from their assumed normal distribution with null mean and variance of 1 (Figure 3.6). We calculated a genomic control inflation factor to quantify the inflation or deflation of each scenario. The genotypic additive effects had an inflation factor of 1.47 in the additive-only scenario and 1.02 in the additive and dominance scenario. The genotypic dominance effects had an inflation factor of 2.37 in the additive only scenario suggesting strong inflation of test statistics, and 0.83 in the additive and dominance scenario suggesting slight deflation of test statistics.

This observation suggests accounting for dominance marker effects is important to avoid spurious associations. Therefore, in order to use this GWAS model on milk production and milk composition phenotype as in Chapter 5, we chose to additionally fit random dominance marker effects. This modified equation (5) to include this additional term.

$$y = Tb + M\_\alpha\_ + M_{\delta\_}\delta\_ + e \qquad (6)$$

Where all terms are as previously discussed in (5) as well as the addition of $\boldsymbol{\delta}\_$, a vector of random SNP chip dominance effects with coverage across the whole genome except for the segment of interest such that $\boldsymbol{\delta}\_ \sim \mathrm{N}\ (\boldsymbol{0}, \boldsymbol{I}\sigma_{\delta}^{2})$, where $\boldsymbol{I}$ is an identity matrix of order equal to the number of dominance effects and $\sigma_{\delta}^{2}$ represents the marker effect variance, $\boldsymbol{M}_{\boldsymbol{\delta}\_}$ is a matrix obtained from $\boldsymbol{M}_{\boldsymbol{\delta}}$ (a matrix relating records to markers (encoded [0, 1, 0])), by deleting the columns corresponding to the region neighbouring the tested sequence variant.

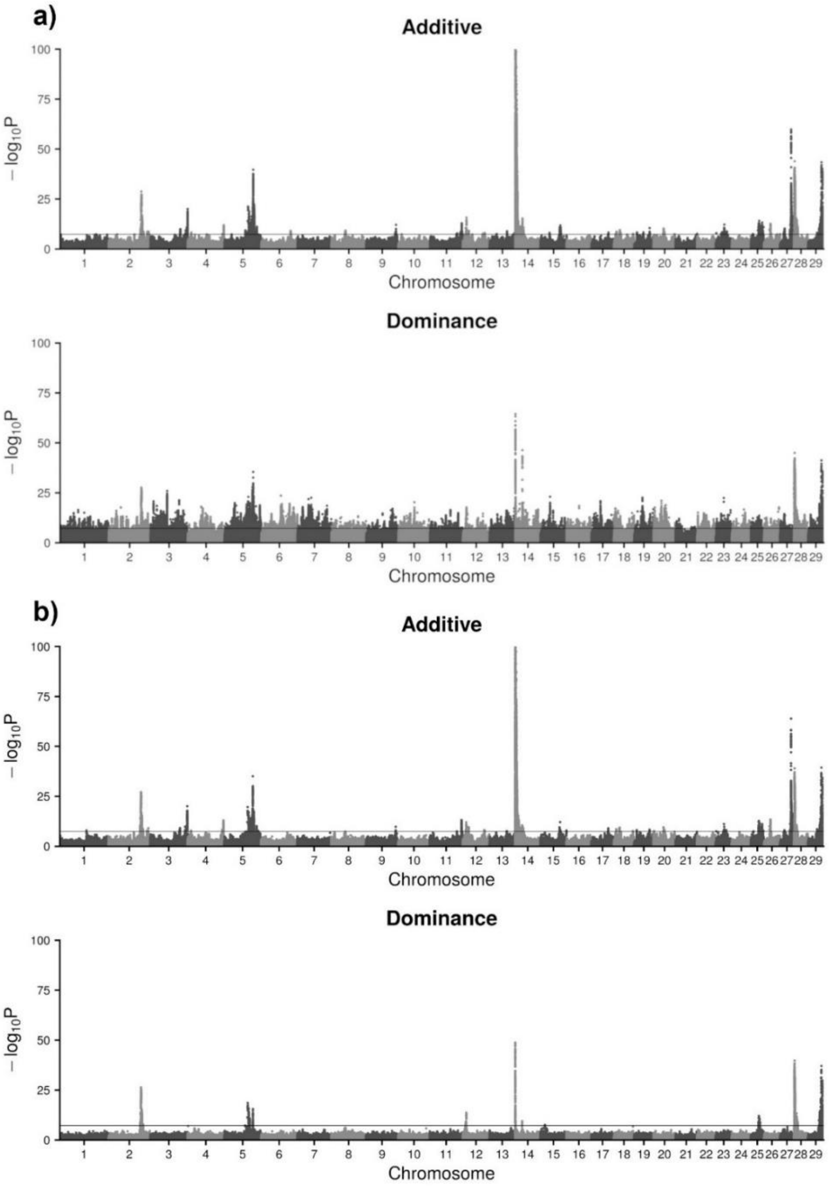


**Figure 3.5 | Manhattan plots of genotypic additive and dominance effects for contrasting model implementations.**

Genotypic additive and genotypic dominance Manhattan plots of milk fat yield when (a) fitting additive-only marker effects and (b) fitting both additive and dominance marker effects. Alternating colours are used to demarcate chromosomes. A multiple testing correction threshold is indicated by a horizontal grey line at P = 5 × $10^{-8}$.

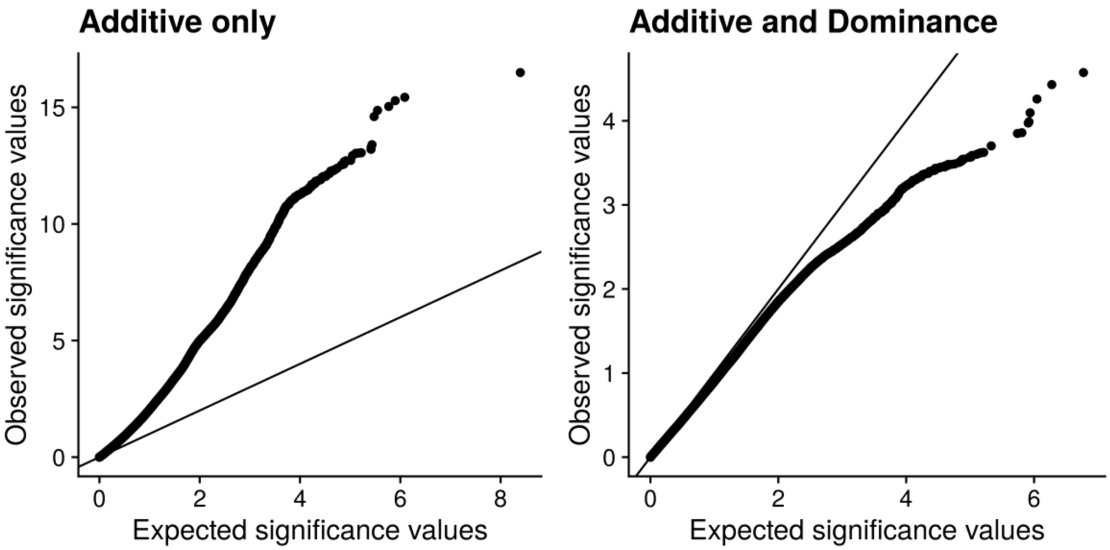


**Figure 3.6 | Plots contrasting the expected vs observed GWAS significance values.**

Scatter plots indicating the distribution of observed significance values for genotypic dominance effects against their expected distribution for both the additive-only model and the additive and dominance model. The black diagonal line indicates x = y.

## 3.5 Discussion

### 3.5.1 Cow phenotypes more readily present recessive effects

In this research, we developed a GWAS model and algorithm to detect non-additive intra-locus effects. A notable observation from the simulation analysis was that sire phenotypes were consistently more powerful than cow phenotypes. However, it is important to note that each sire phenotype was based on the mean of 100 daughter phenotypes and these daughter counts were presented as 'Number of daughters' in Figure 3.3. The dairy cattle industry promotes a small number of bulls to be used across a large proportion of the national herd. Between the years 2000 and 2015 an average of 439 bulls/year were proven (Livestock Improvement Corporation 2020), each proof typically comprising of less than 100 daughters.  Only about 10% of those sires are used subsequently to their progeny test. This means attaining large numbers of sires with many (>100) offspring is difficult and the power advantages demonstrated by simulation would be impractical to achieve in real terms. It is interesting that additive models can detect a rare recessive QTL in sires, a result that differs from that attainable using cow models. This difference indicates how affected daughters' phenotypes blend with the non-affected daughter phenotypes when aggregating the sire's phenotype. This means that a recessive mutation may be detectable as an additive QTL, though the effects may be missed unless further dissection of the underlying data is performed. Together, these limitations of sample size and the phenomenon of 'phenotype blending' make cow phenotypes easier to work with than sire models.  As such, we decided to investigate cow phenotypes in subsequent analyses.

### 3.5.2 The BovineSNP50 panel sufficiently captures the genomic relationships of the population

We aimed to examine the variance components of growth and developmental traits, identifying substantial dominance heritabilities for liveweight and stature. These findings presented the opportunity to investigate how dominance plays a role in these traits, where we detected dominance QTL as described in Chapter 2. This knowledge may also be exploited in future selection strategies and prediction models. Through incorporating dominance in genomic prediction models the accuracy of estimated genomic breeding values can be increased (Varona et al. 2018), and through mate allocation strategies breeders can achieve an improvement in selection response when including dominance to guide their decisions (Toro and Varona 2010).

We investigated whether a medium or a high density SNP chip would be sufficient to capture the genetic variance attributable to these traits. Although the heritability estimates derived from the BovineHD panel were typically numerically higher than those of the BovineSNP50 panel, we did not observe a significant change in heritability estimates between these panels and the increase in marker density from 34,738 to 280,570 independently distributed SNPs did not capture any additional genetic variance when fitted simultaneously. The lack of difference in heritability estimates between genotyping panels suggests the BovineSNP50 panel can adequately capture the relationship structures between individuals, and this indicates it may be used to represent the population structure in our sample.

We used Chebyshev's Inequality (Chebyshev 1867) because it is a generalised theorem that can apply when the probability distributions are unknown. Due to its generality, this test was very conservative and meant we had reduced power to detect a difference in heritability estimates between panels. However, in cases where significant differences were identified (like the breed and model comparisons discussed below), we can be more comfortable with our results.

For stature, we observed Holstein-Friesians had significantly higher additive heritability estimates than Jerseys. Additive heritability gives an indication of the future opportunity for genetic gain via selection (Hill, Goddard, and Visscher 2008), as there is more genetic variance for selection to act on. Therefore, this finding suggests (using the data available) there is increased opportunity for selection on stature in Holstein-Friesians than Jerseys. We did not observe differences in dominance heritabilities between breeds which may indicate dominance plays a similar role in these traits across breeds.

Huang & Mackay (2016) have shown additive genetic variance can appear to contribute the majority of the total genetic variance regardless of the underlying genetic mechanisms. We identified a non-zero sampling covariance between additive and dominance variance components, as such we investigated dominance genetic variance in the absence of an additive genetic variance term. We observed a significant increase in dominance heritability between the two models for liveweight. While this dominance genetic variance estimation is not the same as the alternative model described in Huang & Mackay (2016), it does indicate that the order the variance components are fitted matters, as each GRM can capture at least some proportion of the variation explained by the other. The finding also shows that while heritability estimates can be useful for genetic selection, they do not define the genetic architecture of a phenotype.

### 3.5.3 The non-additive GWAS model was implemented

We chose to fit the genotype class means model to represent the sequence variant being tested because it appeared to best capture the variation of different non-additive genetic mechanisms. While this parameterisation can capture the variation of a locus of interest, it is not particularly useful for interpretation. Instead, we can transform the effect estimates to those of equivalent models (Henderson 1985), as discussed in Chapter 2 – Supplementary note. We chose one of these equivalent models, the genotypic model, which fits a genotypic additive and a genotypic dominance term. Other researchers have applied different approaches to detecting dominance effects. Instead of fitting the genotypic model, these studies fit a more traditional additive or breeding value term as derived from an additive-only model and a dominance deviation term (Zhu et al. 2015; Jiang et al. 2019). An important consideration when modelling additivity in this way is that instead of modelling biological additivity, instead one is modelling the breeding value, $\alpha$, where α = a + (1-2p) d. This term depends on the allele frequency of the variant ($p$) and the dominance effect of the variant, so while it is useful for selection purposes it may be less useful for understanding the biological effect of the mutation. The genotypic dominance and dominance deviation terms fitted in the different model parameterisations are identical in effect estimate so when identifying dominance QTL through GWAS, the models are equivalent. However, when interpreting the biological mechanism underlying the QTL, we found the genotypic model more appropriate (Reynolds et al. 2021).

In implementing Step 1 and Step 2 of the GWAS model, we observed the difficulties of working with a programming language that was in its infancy. Many updates to the Julia language during its beta stages made it difficult to develop stable tools and syntax changes often broke tools as future updates were released. The 8th of August 2018 saw the release of Julia 1.0 and a massive update which resulted in the requirement of several bug fixes, fortunately this did help stabilise the language development and made it easier to develop future tools.

### 3.5.4 Managing a new marker set and accounting for dominance population structure

We developed a GWAS model to scan millions of sequence variants for dominance and recessive mechanisms. We applied this model to growth and developmental traits in Chapter 2 and discovered several candidate causal mutations with recessive effects. Through these discoveries and through simulation, we suggest the GWAS model developed is fit for purpose. The transition between reference genome assemblies and application to new phenotypes presented a new challenge.

During this research, the world of bovine genomics made a major transition between reference genomes from UMD 3.1.1 to ARS-UCD1.2 (Zimin et al. 2009; Rosen et al. 2020). Although updated genome assemblies can provide novel insights, this change can also cause disruption and required major updates and changes to sequence mapping, genotype panel specifications, sequence imputation, and other genomic pipelines. This change promised improvements in all these areas as well due to increased continuity and accuracy of the reference assembly (Rosen et al. 2020). To account for these changes, we filtered the ARS-based BovineSNP50 panel to get a set of 31,451 markers to account for population structure. Despite fewer genomic positions in the marker set, it appeared to sufficiently remove spurious associations.

Instead of adjusting phenotypes for the pedigree-derived pairwise heterosis term in the national genetic evaluation model we wanted to see what impact dominance population structure had on our GWAS results. We observed increased inflation in the significance of genotypic additive and dominance effects when we only fit additive marker effects. This inflation appeared to decrease when we fit both additive and dominance marker effects, where the genotypic additive effects were very weakly inflated, and the dominance effects were somewhat deflated. As a result of these observations, to account for the spurious associations caused by dominance population structure, we extended the model to include random dominance effects which appeared to sufficiently account for spurious dominance associations although may be somewhat conservative in dominance effect estimation. This was implemented and applied in an investigation of lactation phenotypes presented in Chapter 5.

## 3.6 Conclusion

Elucidating the genetic architecture of complex traits is a key consideration in the genetic improvement of species and prevention to genetic disease.  Many previous experiments have largely ignored the contributions from non-additive genetic mechanisms, owing partly to the lack of statistical models capable of doing so.  We have addressed these challenges and successfully implemented an intra-locus non-additive GWAS model in a Julia package called Modelling Association of Non-Additive effects (MANA).  This method accounts for population structure using additive or additive and dominance random marker effects to reduce spurious associations. This tool can be used to detect non-additive QTL and locate genomic regions of interest which may contain causal mutations useful for selection and discovery of novel genetic disease.

# Chapter 4 An investigation of a *MUS81* nonsense mutation on recombination rate

## 4.1 Abstract

The *MUS81* gene has an important role in DNA damage repair and the resolution of errors during replication, whereby knockout of the gene can lead to deleterious consequences in both human and mice cells. Recently, a knockout mutation in the *MUS81* gene was discovered in cattle presenting a large deleterious recessive effect on growth and production traits. Here, we investigated whether recombination rate might act as a proxy phenotype for *MUS81*'s role in DNA damage repair. While we corroborate differences in recombination rate between sexes and between breeds, we could not detect a significant relationship between the *MUS81* mutation and recombination rate. These findings suggest more specialised DNA damage repair phenotypes may be required to elucidate the molecular basis of the inherited genetic disorder.

## 4.2 Introduction

In a recent non-additive GWAS study we discovered several recessive loci and mutations for growth and developmental traits (Reynolds et al. 2021)[Chapter 2]. One of the most prominent loci identified was a signal on chromosome 29 that presented two plausible candidate causative mutations for this effect, consisting of a conserved missense variant in the *DPF2* gene (double PHD fingers 2), and a premature stop codon in the *MUS81* gene (*MUS81* structure-specific endonuclease subunit). To attempt to differentiate which of these candidates might be responsible, we assessed the biology of these genes to perform targeted analyses that might differentiate the effects of both candidate mutations. Chapter 2 presented hoof anatomical data to explore gene-specific phenotypes related to *DPF2,* a known regulator of finger and toenail characteristics in humans (Vasileiou et al. 2018).

Here, we present an association analysis of recombination rate, a process with potential relevance to the function of *MUS81.*

An organism's DNA may be damaged through naturally occurring cellular and environmental processes, that can compromise the genomic integrity of the cell and lead to premature cell death, rapid ageing, or cancer. DNA damage repair mechanisms exist to prevent and resolve this damage and act through multiple biological pathways (Hakem 2008). While DNA damage and its subsequent repair can occur at checkpoints throughout the cell cycle, during meiosis the repair of double stranded DNA breaks can cause crossover events to occur through a process called homologous recombination (Hakem 2008). Crossover events facilitate the exchange of genetic information between parental homologous chromosomes (Coop and Przeworski 2007). When recombination catastrophically fails, through inconsistencies in double stranded DNA break repair or through other processes, chromosomal instability can occur whereby gametes may have an atypical number of chromosomes (aneuploidy) or may have chromosomal rearrangements due to non-homologous recombination (ectopic exchange) (Coop and Przeworski 2007).

Recombination plays a significant role in genomic diversity because it is the process that dictates the erosion of linkage disequilibrium through the breakage and formation of haplotypes (Baudat, Imai, and de Massy 2013). The recombination rate of an individual, often quantified by the number of crossover events, is a heritable complex quantitative phenotype and is modulated by many genes, several of which have been identified to have major effects in both humans and model organisms (Baudat, Imai, and de Massy 2013; Stapley et al. 2017). The deleterious consequences resulting from recombination failures can have a highly deleterious impact on the health of offspring and makes meiotic recombination an important mechanism to study.

The *MUS81* gene is involved in DNA damage repair and can prevent genomic aberrations by resolving replication complications (Hanada et al. 2007; Fu et al. 2015). Null mutations of *MUS81* in human cell lines result in slower DNA replication rates (Fu et al. 2015), and in mice increase the cells' susceptibility to DNA damage and chromosomal aberrations (Dendouga et al. 2005). *MUS81* holds a complex, multi-faceted role in cancer cells with instances of both promoting and obstructing cell proliferation (Chen et al. 2021). In one case, *MUS81* was shown to aid recombination in telomerase negative cancer cells (Zeng et al. 2009). While the viability and fertility of *MUS81* knockout mice suggests *MUS81* is not essential for meiotic recombination (McPherson et al. 2004), the role of *MUS81* in DNA repair and maintenance of genomic integrity in humans and mice cells suggests it may play a similar role in cattle. We propose recombination rate may act as a proxy for the efficacy of DNA damage repair in cattle. As such, we hypothesise the deleterious mutation in *MUS81* (p.Gly70*) may increase the number of crossing over events that occur in an individual's cells. Here, we estimated a recombination rate phenotype in over 28,000 animals to investigate the impact of the premature stop mutation in *MUS81*.

## 4.3 Methods

### 4.3.1 Animal population and genotyping

Our study consisted of 28,053 animals from the New Zealand dairy cattle population. These cattle included 3,578 males and 24,475 females and the breeds of these animals were primarily 16/16ths Holstein-Friesian (HF), 16/16ths Jersey (J), or crossbred (HFxJ).

An individual's breed may be coded as 16/16ths; however, this does not preclude the possibility that an ancestor may be crossbred.

Study animals were genotyped on a variety of medium- and high-density genotyping platforms, previously described in Chapter 2 (Reynolds et al. 2021). For recombination rate estimation, we used physical genotypes of autosomal SNPs which had not been imputed. Genotypes of the *MUS81* mutation were extracted from the imputed to sequence genotype set described in Chapter 2.

### 4.3.2 Phenotype generation

We estimated the number of crossover events per animal to represent recombination rate. To do so, we identified crossover events using LINKPHASE3 (Druet and Georges 2015), in a similar way to Kadri et al. (2016). LINKPHASE3 uses SNP-chip and pedigree information to reconstruct haplotypes and identify crossover events between parent-offspring pairs. Autosomal SNP genotypes which provide utility in identifying crossover events are termed informative markers and the number of informative markers used for each parent-offspring pair is reported by LINKPHASE3. Informative markers must be phased in the parent, parentally phased in the offspring, and heterozygous in the parent. An increased density of informative markers in a genomic interval increases the likelihood of identifying crossover events in that interval if they exist.  As such those animals that are genotyped on lower density SNP chips or have a higher degree of homozygosity will have fewer informative markers and this may mask crossover events from being detected and result in an artificially lower recombination rate than their true recombination rate.

Our study consisted of 131,464 sire-offspring pairs and 40,950 dam-offspring pairs. In the case of trios, an offspring will contribute to both the sire's and dam's phenotype as the offspring's haplotypes have been parentally phased so the crossover event estimates for the two parental gametes that formed the individual are independent.

### 4.3.3 Single locus model

To analyse this dataset, we used a similar approach to the Single Locus Models presented in Chapter 2. We fit a mixed linear model including pedigree to account for population structure, and repeated measures using the Julia packages, JWAS (Cheng, Fernando, and Garrick 2018) and MANA (Reynolds et al. 2021).

The model is represented by

$$y = Xb + Zu + Wp + e \tag{1}$$

where $\boldsymbol{y}$ is a vector of the number of crossing over events identified across all autosomes, $\boldsymbol{b}$ is a vector of fixed effects including the variant of interest (encoded as genotype class effects), the number of informative markers (a quantitative covariate), breed (an animals Holstein-Friesian breed proportion in 16ths), and sex (a class effect), $\boldsymbol{X}$ is a design matrix relating records to respective fixed effects, $\boldsymbol{u}$ is a vector of random breeding value effects such that $\boldsymbol{u} \sim \mathrm{N}\ (\boldsymbol{0}, \boldsymbol{A}\sigma_u^2)$ where $\sigma_u^2$ represents the additive genetic variance and $\boldsymbol{A}$ is the additive relationship matrix conditional on the pedigree and, $\boldsymbol{Z}$ is a design matrix relating records to breeding values, $\boldsymbol{p}$ is a vector of random permanent environment effects such that $\boldsymbol{p} \sim \mathrm{N}\ (\boldsymbol{0}, \boldsymbol{I}\sigma_p^2)$ where $\sigma_p^2$ represents the permanent environment variance, and $\boldsymbol{I}$ is an identity matrix of order equal to the number of phenotypic records, $\boldsymbol{W}$ is a design matrix relating records to permanent environment effects, and $\boldsymbol{e}$ is a random error term where $\boldsymbol{e}$

$\sim$ N ($\boldsymbol{0}$, $\boldsymbol{I}\sigma_e^2$) where $\sigma_e^2$ represents the residual error variance. The random variables $\boldsymbol{u}$, $\boldsymbol{p}$ and $\boldsymbol{e}$ are assumed to be uncorrelated (Reynolds et al. 2021).

JWAS implements a Gibbs sampler to draw plausible samples of effects and variance components from their posterior distributions at each iteration (Cheng, Fernando, and Garrick 2018). We ran the chain for 100,000 iterations with a burn in of 50,000, keeping every 10th iteration thereafter. Through this computation, we obtained vectors of plausible effect estimates from the posterior distribution of each effect fit in the model, x. These vectors were then summarised by their posterior mean ($\beta_x$), standard deviation ($\sigma_x$), and z-statistic ($z_x$). The statistical significance of these effects was evaluated using a Z-test, where $z_x = \frac{\beta_x}{\sigma_x}$ and $z_x \sim N(0, 1)$.

## 4.4 Results

### 4.4.1 Estimation of a crossing over phenotype

We used data from our mixed breed population to estimate a recombination rate phenotype using LINKPHASE3. We identified 3,418,097 crossover events in gametes transmitted from 3,578 males to 131,464 offspring, and 937,604 crossover events in gametes transmitted from 24,475 females to 40,950 offspring. The average global recombination rate (GRR) was 25.4 in males and 22.7 in females (Figure 4.1). Previous estimates of male GRR are 27.4 in Angus, and 26.9 in Limousin cattle which are consistent with our findings (Weng et al. 2014). These recombination rates are also consistent with those observed in Ma et al. (2015), where that study estimated GRRs at approximately 25.5 in males and 23.2 in females. These findings demonstrated the ability of LINKPHASE3 to detect crossover events.

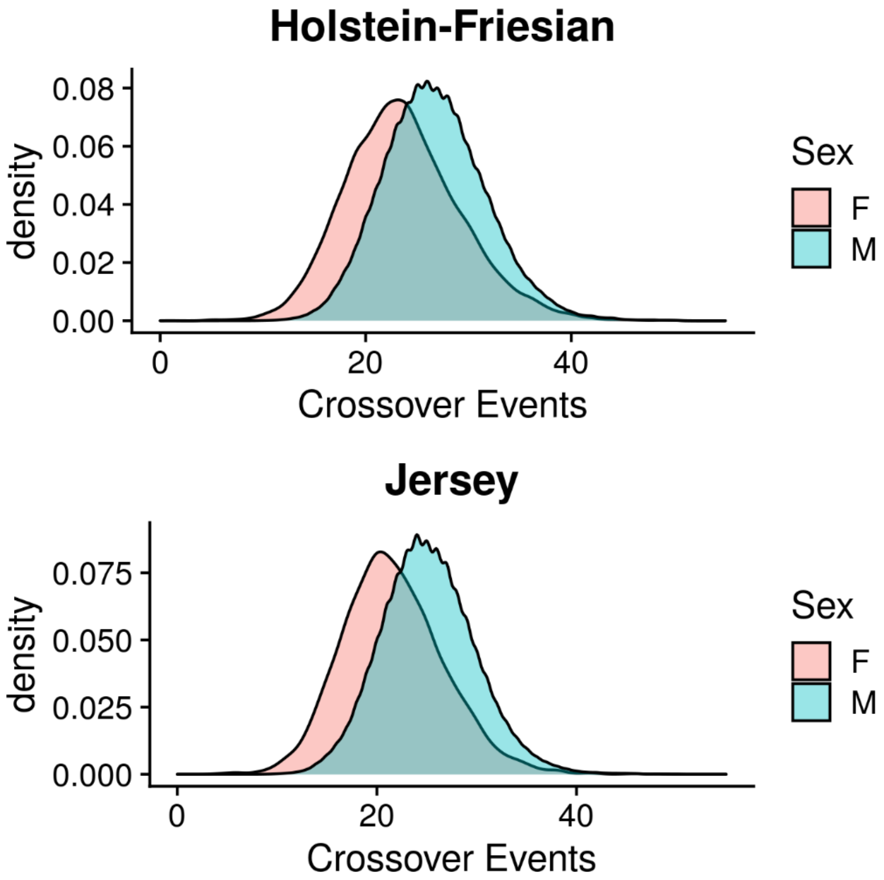


**Figure 4.1 | Crossover events between parent offspring pairs.**

Density plots comparing the number of crossover events for male (M) and female (F) animals and for each of the Holstein Friesian and Jersey breeds. An animal was considered to be Holstein-Friesian or Jersey if they were at least 13/16$^{th}$s purebred.

### 4.4.2 MUS81 is not associated with recombination rate

We fit a linear mixed model with pedigree and random permanent environment terms to assess the effect of *MUS81* on recombination rate. While we observe fewer crossover events in affected individuals for the *MUS81* mutation, this is not a significant result (Table 4.1). As such, there is no evidence from these data that the *MUS81* mutation has an impact on recombination rate.

**Table 4.1 | Summary statistics for recombination rate analysis.**

| Covariate | Posterior mean | Standard deviation | p-value |
|---|---|---|---|
| $A_1A_2$ – $A_1A_1$ (*MUS81*) | 0.104 | 0.094 | 0.27 |
| $A_2A_2$– $A_1A_1$ (*MUS81*) | -0.731 | 0.487 | 0.13 |
| Holstein-Friesian (per 16th) | 0.087 | 0.018 | 2.36e-6 |
| Sex (Male/Female) | 2.618 | 0.051 | 9.81e-575 |
| Informative Markers (per 10,000) | 6.54E-05 | 8.94E-07 | 4.69e-1164 |

Table 4.1 presents the posterior means, standard deviations, and p-values derived from the Markov chains of effects fitted in the single locus recombination rate model. The effects of the MUS81 locus are presented as differences between the three genotype classes ($A_1A_1$, $A_1A_2$, and $A_2A_2$).

We estimated the heritability and repeatability of the recombination rate phenotype using the pedigree and permanent environment terms in the model. Heritability was estimated as 0.106 ± 0.007 and repeatability estimated as 0.224 ± 0.005, these values are similar to those reported in Kadri et al. (2016) who used the same LINKPHASE3 software. This suggests that our recombination rate phenotype is characterising real crossover events and that there is a genetic component to the phenotype.

We note a significant impact of breed on the number of crossover events transmitted between parents and their offspring (Table 4.1). This result indicates a purebred (16/16ths) Holstein Friesian has 1.4 additional crossover events compared to a purebred Jersey. This effect appears to be consistent across sexes, where purebred Holstein-Friesian GRRs are 26.2 for males and 23.5 for females whereas purebred Jersey GRRs are 24.4 for males and 21.3 for females. Breed-specific recombination rates have been previously observed in Shen et al. (2018) where the authors compared Holstein-Friesian, Jersey, Ayrshire, and Brown Swiss breeds.

We detect a large effect of sex on recombination rate, where males have 2.6 more crossover events in their gametes than females. This finding is consistent with previous research that observed an average of 1.9 (Kadri et al. 2016) and 2.3 (Ma et al. 2015) more crossover events in males than females. We observe the number of informative sites is strongly significant in the number of crossover events. This confirms that recombination will be more definitively resolved at an increased density of informative markers.

## 4.5 Discussion

### 4.5.1 MUS81 mutation does not act through recombination rate

The *MUS81* gene has been implicated in DNA damage repair pathways and the resolution of complexities during replication in humans and mice (Hanada et al. 2007; Dendouga et al. 2005; Fu et al. 2015). A premature-stop mutation in *MUS81* was identified as a candidate causal variant resulting in a large recessive effect on growth, developmental, and production traits in New Zealand dairy cattle (Reynolds et al. 2021). Given *MUS81*s role in preventing chromosomal aberrations, we hypothesised loss-of-function of the protein would result in increased chromosomal structural changes that might present as an increased recombination rate. We did not identify a significant difference in recombination rate between genotype classes, although individuals homozygous for the mutation present 0.73 fewer crossover events compared to their wild type counterparts. These findings suggest the deleterious impact caused by the loss-of-function mutation does not act through increased recombination rate, similar to previous findings (McPherson et al. 2004) in which mice homozygous fora *MUS81* knockout mutation were still viable and fertile. Instead, the deleterious impact may manifest through another yet to be determined cellular mechanism or the *MUS81* mutation might not be the causal mutation responsible for the growth and developmental trait QTL. Assuming that the mutation is involved, it may operate through failure to repair DNA damage leading to more holistic health complications such as immunodeficiency, neurological disorders, and cancer (Hakem 2008). Future work might investigate structural variant or de novo mutation frequencies as alternative proxies of DNA damage and chromosomal aberrations, however these experiments may require higher resolution genotypic datasets, and additional, more specialised genomic data such as accurately called structural variants datasets.

### 4.5.2 Further evidence for effect of sex and breed on recombination rate

We identified differences in the number of crossover events between sexes and between breeds. Males had 2.6 additional crossover events over females, and purebred Holstein Friesians had 1.4 more than purebred Jerseys. These differences between sexes in cattle have been identified previously, where Ma et al. (2015) identified the recombination map of male cattle was 10% longer than females for every chromosome, a result that was confirmed in a separate population (Kadri et al. 2016). Ma et al. (2015) suggested the greater artificial selection intensity on males over females has led to this discrepancy in recombination map length. The difference between breeds has been presented in Shen et al. (Shen et al., 2018), where the authors identified several SNP intervals with differing recombination rates between Holsteins and Jerseys. We observed comparable but numerically fewer crossover events in Holstein-Friesian and Jersey bulls compared to that observed in Angus and Limousin bulls (Weng et al. 2014). These differences between breeds might suggest recombination rate QTL are segregating between breeds, or that the increased homozygosity or inbreeding in some breeds such as Jerseys (Pryce et al. 2014) may mask crossover events resulting in an artificially lower estimate of recombination rate.

While we have generated a recombination rate phenotype that was consistent with those reported by others (Weng et al. 2014; Ma et al. 2015; Kadri et al. 2016; Shen et al. 2018), it is important to note that our dataset may be limited by the density of SNP interrogated. We observed a highly significant association between the number of informative markers and the number of crossover events. Several recent reports suggest there are many more crossover events in the telomeres of cattle autosomes compared to their centres (Ma et al. 2015; Kadri et al. 2016; Shen et al. 2018). This may suggest having more individuals genotyped on higher density SNP-chip panels with an increased density in telomere regions would allow us to observe crossover events more accurately using LINKPHASE3.

## 4.6 Conclusion

Here, we have generated a dataset to examine recombination rate in cattle, and validated recombination rate differences between sexes and between breeds in the New Zealand dairy cattle population. We found no evidence that a deleterious loss-of-function mutation in the *MUS81* gene has an impact on recombination rate, suggesting that, if the mutation is indeed responsible for the large recessive QTL detailed in Chapter 2, it likely acts through some other biological mechanism. These findings suggest more detailed chromosomal and molecular phenotypes will be required to highlight such as function.

# Chapter 5 Non-additive QTL mapping of lactation traits in 124,000 cattle reveals novel recessive loci

Edwardo GM Reynolds[1], Thomas Lopdell[2], Yu Wang[2], Kathryn M Tiplady[1,2], Chad S Harland[2], Thomas JJ Johnson[2], Catherine Neeley[2], Katie Carnie[2], Richard G Sherlock[2], Christine Couldrey[2], Stephen R Davis[2], Bevin L Harris[2], Richard J Spelman[2], Dorian J Garrick[1], Mathew D Littlejohn[1,2]

[1]Massey University, New Zealand

[2]Livestock Improvement Corporation, Hamilton, New Zealand

Correspondence should be addressed to EGMR (edwardo.reynolds.1@uni.massey.ac.nz)

## 5.1 Abstract

### 5.1.1 Background

Deleterious recessive conditions have been primarily studied in the context of Mendelian diseases. Recently, several deleterious recessive mutations with large effects were discovered via non-additive genome-wide association studies (GWAS) of quantitative growth and developmental traits in cattle, which showed that quantitative traits can be used as proxies of genetic disorders when such traits are indicative of whole-animal health status. We reasoned that lactation traits in cattle might also reflect genetic disorders, given the increased energy demands of lactation and the substantial stresses imposed on the animal. In this study, we screened more than 124,000 cows for recessive effects based on lactation traits.

### 5.1.2 Results

We discovered five novel quantitative trait loci (QTL) that are associated with large recessive impacts on three milk yield traits, with these loci presenting missense variants in the *DOCK8*, *IL4R*, *KIAA0556*, and *SLC25A4* genes or premature stop variants in the *ITGAL*, *LRCH4*, and *RBM34* genes, as candidate causal mutations. For two milk composition traits, we identified several previously reported additive QTL that display small dominance effects. By contrasting results from milk yield and milk composition phenotypes, we note differing genetic architectures. Compared to milk composition phenotypes, milk yield phenotypes had lower heritabilities and were associated with fewer additive QTL but had a higher non-additive genetic variance and were associated with a higher proportion of loci exhibiting dominance.

### 5.1.3 Conclusions

We identified large-effect recessive QTL which are segregating at surprisingly high frequencies in cattle. We speculate that the differences in genetic architecture between milk yield and milk composition phenotypes derive from underlying dissimilarities in the cellular and molecular representation of these traits, with yield phenotypes acting as a better proxy of underlying biological disorders through presentation of a larger number of major recessive impacts.

## 5.2 Background

Non-additive genetic effects are best known from studies of Mendelian diseases, where recessive conditions have been shown to have major deleterious impacts on health and performance. These studies have mostly used a 'forward genetics' approach, where the observation of a disease phenotype precedes fine mapping and sequencing to highlight the mutation (Charlier et al. 2012; Littlejohn, Henty, et al. 2014; Bourneuf et al. 2017). However, the reverse approach has also been applied, which first identifies candidate loss-of-function genotypes and subsequently performs phenotyping on traits likely to reflect the impact of the mutation (VanRaden et al. 2011; Charlier et al. 2016; Michot et al. 2016). Genome-wide association studies (GWAS) have been used to investigate non-additive effects in quantitative traits, but the number of findings remains limited in comparison to additive effects, where most such analyses fit an additive model only. Recent studies of non-additive effects include the investigation of complex traits in both humans (Zhu et al. 2015) and cattle (Bolormaa et al. 2015; Sun et al. 2014; Aliloo et al. 2016; Jiang et al. 2019; Reynolds et al. 2021). In cattle, Reynolds et al. (2021) identified several recessive mutations with major negative impacts on growth and developmental traits, where some of these effects were found to be due to underlying genetic syndromes.

The concept of using routinely gathered, quantitative traits as proxies of genetic disorders is based on the idea that phenotypes such as growth or liveweight might be indicative of the overall health status of the animal, e.g. reduced growth could be caused by an underlying genetic disorder, in which case such effects could be detected via GWAS. Thus, it is relevant to investigate whether other easily measured traits might also serve as proxies of animal fitness, with a view to extend the scope of this approach. Lactation traits such as milk volume comprise one of the most commonly targeted classes of quantitative traits studied in cattle, where additive analyses of these traits have identified numerous candidate causative genes such as *DGAT1* (Grisart et al. 2002), *GHR* (Blott et al. 2003), *ABCG2* (Cohen-Zinder et al. 2005), *GPAT4* (Littlejohn, Tiplady, et al. 2014), and *MGST1* (Littlejohn et al. 2016). Lactation traits might also reflect genetic disorders, given the increased energy demands of lactation and the substantial metabolic and physiological stresses imposed on the animal (Bauman and Bruce Currie 1980). Thus, we were interested in investigating whether the application of non-additive models to lactation data might identify recessive mutations in addition to those found for growth traits, and to this end, have conducted non-additive GWAS for milk traits on 124,000 animals. We contrast the additive and non-additive genetic architectures of milk yield traits and milk composition traits. Finally, we describe the discovery of several novel major effect recessive loci and highlight candidate mutations that could underlie these undiagnosed recessive disorders.

## 5.3 Methods

### 5.3.1 Animal populations

The dataset reported in this study consists of 124,364 New Zealand dairy cattle. These animals come from a mixed breed population, where 20,893 were recorded as 16/16th's Holstein–Friesian (HF), 13,184 were recorded as 16/16th's Jersey (J), 67,520 were crosses with varying proportions of the two breeds (HFXJ), and 22,767 were HF or J crossbreeds with minor proportions of other breeds including Ayrshire, Brown Swiss, or Hereford (and other crosses). The breed of an individual may be coded as 16/16ths, however, this does not preclude the possibility that an ancestor may have been crossbred since matings between 15/16ths and 16/16ths animals are recorded as producing 16/16ths offspring. The animals were born between 1990 and 2018 with a mean birth year of 2010.

### 5.3.2 Phenotypes

We analysed five first-lactation yield deviation phenotypes: three milk yield traits, i.e. milk volume (L/lactation; a lactation refers to a standardised 268-day lactation period; N = 124,356), milk protein yield (kg/lactation; N = 124,356), and milk fat yield (kg/lactation; N = 124,356); and two milk composition traits, i.e. milk protein percentage (%; N = 124,363), and milk fat percentage (%; N = 124,363). Milk protein yield and milk fat yield are calculated on individual herd tests and are the product of the herd test milk volume multiplied by the herd test milk protein percentage or milk fat percentage, respectively.

Prior to genetic analysis, the phenotypes were adjusted for non-genetic effects obtained from the national genetic evaluation of the entire cattle population (30 ×

$10^6$ animals), which fits a mixed linear model, including effects for: contemporary group, age at calving, stage of lactation, and record type (i.e. am milkings, or pm milkings, or both). Since the number of herd-test measurements varies for each animal, these adjusted test day phenotypes were aggregated to a first lactation phenotypic deviation such that each animal has a single record and a corresponding weighting that reflects the amount of information contributing to the record (Garrick, Taylor, and Fernando 2009).

### 5.3.3 Reference population for sequence-based imputation

Whole-genome sequencing was performed on 1300 animals that were mostly ancestral sires and represented the reference population for sequence-based imputation. These animals: HF (N = 306), J (N = 219), HFXJ (N = 717), or other breeds and crossbreeds (N = 58); were sequenced on Illumina HiSeq 2000 instruments targeting 100-bp paired-end reads. The sequence data were aligned to the ARS-UCD1.2 reference genome assembly using the Burrows–Wheeler alignment algorithm (BWA) version 0.7.17 (Li 2013), which resulted in a mean read depth of 15×. For variant calling, we used the Genome Analysis ToolKit (GATK) v4.0.6.0 (DePristo et al. 2011), followed by filtering of the variants with the variant quality score recalibration technique (DePristo et al. 2011). Based on the animals with a read depth > 10× (N = 850), variants that were singletons or were multi-allelic, had a map quality score lower than 50, or a Mendelian error rate higher than 5%, were filtered out leaving 21,005,869 whole-genome sequence variants. The genotypes at the positions of these filtered variants were extracted from the sequence data of all 1300 animals and were phased using the software Beagle 5.0 (Browning, Zhou, and Browning 2018) to generate the sequence-based imputation reference panel.

### 5.3.4 Genotyping

DNA was extracted either from ear-punch tissue samples or blood samples for the 124,364 animals included in our study. These samples were processed to extract DNA at GeneMark (Hamilton, New Zealand) using Qiagen BioSprint kits, or at GeneSeek (Lincoln, NE, USA) using the Life Technologies' MagMAX system. Genotyping was performed using a variety of single nucleotide polymorphism (SNP) arrays including GeneSeek GGPv1 (8729 SNPs), GGPv2 (20,012 SNPs), GGPv2.1 (20,015 SNPs), GGPv3 (31,813 SNPs), GGPv3.1 (31,945 SNPs), GGPv4 (37,092 SNPs), GGP50kv1 (48,156 SNPs), GGP50kv1.1 (48,161 SNPs), Illumina BovineSNP50v1 (53,126 SNPs), Illumina BovineSNP50v2 (53,629 SNPs), or the BovineHD (772,235 SNPs) chips.

### 5.3.5 Consolidation of SNP-chip panels for sequence imputation

Imputation from the genotyping panels to sequence resolution was performed as described in Wang et al. (2020). The various genotyping panels were grouped into four sets: GGP panels (GGPv1, GGPv2, GGPv2.1, GGPv3, GGPv3.1, and GGPv4), 50K panels (BovineSNP50v1 and BovineSNP50v2), GGP50k panels (GGP50kv1 and GGP50kv1.1), and the BovineHD panel. Animals genotyped on the GGP panels were imputed to the BovineSNP50v1 panel, then combined with the physically genotyped 50K panel animals and successively imputed to the BovineHD panel. Animals genotyped on the GGP50k panels were separately imputed to the BovineHD panel in a single step. In order to incorporate the custom content that had been genotyped on the GGPv3 platform, we conducted similar imputation steps to impute all animals to GGPv3. Then, we combined the imputed and physically genotyped panels (imputed HD, imputed GGPv3, and physically genotyped HD), and finally imputed the resulting animals to sequence resolution using the sequence-based imputation reference population, described above. LINKPHASE3 (Druet and Georges 2015) and Beagle 5.0 (Browning, Zhou, and Browning 2018) were used for all phasing and imputation steps. In Beagle 5.0, we applied the default parameters except for effective population size that was set at 400, and a window size of 20 Mb was used except for chromosomes 7, 10, 12, 14, and 23, for which a 7-Mb window size was applied because of the greater computational demands for these chromosomes, probably due to assembly and structural complexities (as previously reported (Pausch et al. 2017)). Very rare variants (homozygous alternate count ≤ 5) were removed by post-imputation filtering and poorly imputed variants based on the dosage $R^2$ statistic ($DR^2$; $DR^2 < 0.7$) were also filtered out. In total, 16,640,294 variants remained for the GWAS and further analyses.

### 5.3.6 Genotypes for the adjustment of population structure

We used the genotyping data from the Bovine SNP50 chip platforms to account for spurious effects due to population structure. From the initial 54,708 autosomal SNPs, markers with a high missing genotype rate (> 0.01), a low minor allele frequency (< 0.02), or that deviated from the expected Hardy–Weinberg equilibrium (> 0.15, calculated within breed) were excluded. An additional filtering step was carried out to remove poorly imputed markers ($DR^2 < 0.9$) and markers in high linkage disequilibrium (LD) with another marker on the panel (pairwise $R^2 > 0.9$, within 1 Mb). After these edits, a set of 31,451 SNPs remained for subsequent analyses.

### 5.3.7 Heritability estimates

We estimated breed-specific additive and dominance heritabilities based on genomic relationship matrices (GRM) using the GCTA software (Yang et al. 2011; Zhu et al. 2015).

Additive and dominance variance components were estimated simultaneously from purebred individuals (HF = 20,893 and J = 13,184), using the same set of 31,451 filtered BovineSNP50 SNPs as for population structure adjustment (see previous section). The GCTA software estimates the variance components using a restricted maximum likelihood (REML) approach. It estimates the additive heritability ($h^2$) as the ratio of additive genetic variance to phenotypic variance, and dominance heritability ($\delta^2$) as the ratio of dominance genetic variance to phenotypic variance. We analysed yield deviations which aggregate the herd test records that are described in the 'Phenotypes' section, thus no additional records not already described were used in this analysis.

### 5.3.8 GWAS

*Overview of the model*

We applied a non-additive GWAS approach that is similar to that described in Reynolds et al. (2021) to identify non-additive QTL for milk traits. This approach is a two-step method that leaves-one-segment-out (LOSO) and fits all other genomic SNP effects among the 31,451 SNPs to adjust for population structure, and then applies a Markov chain Monte Carlo (MCMC) method to test the effects of all imputed-to-sequence variants in the segment that had been left out, one at a time. In general, for each sequence variant the method fits the following model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{T}\mathbf{b} + \mathbf{M}_{\boldsymbol{\alpha}}\boldsymbol{\alpha} + \mathbf{M}_{\boldsymbol{\delta}}\boldsymbol{\delta} + \mathbf{e}, \qquad (1)$$

where $\mathbf{y}$ is the vector of one of the five phenotypes of interest that were pre-adjusted as described in the 'Phenotypes' section; $\mu$ is the overall mean; $\mathbf{1}$ is a vector of 1s; $\mathbf{b}$ is a vector of genotype class effects for the sequence variant of interest; $\mathbf{T}$ is the design matrix relating records to genotype class for the sequence variant; $\boldsymbol{\alpha}$ is a vector of random additive effects of SNPs spanning the whole genome except the segment of interest such that $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}\sigma_{\alpha}^{2})$, and $\mathbf{I}$ is an identity matrix of order equal to the number of SNP effects and $\sigma_{\alpha}^{2}$ is the additive variance of the SNP effects; $\boldsymbol{\delta}$ is a vector of random dominance effects of SNPs spanning the whole genome except the segment of interest such that $\boldsymbol{\delta} \sim N(\mathbf{0}, \mathbf{I}\sigma_{\delta}^{2})$, and $\sigma_{\delta}^{2}$ is the dominance variance of the SNP effects; $\mathbf{M}_{\boldsymbol{\alpha}}$ and $\mathbf{M}_{\boldsymbol{\delta}}$ are matrices in which each column represents the covariate values for a marker locus ([0, 1, 2] and [0, 1, 0], respectively); and $\mathbf{e}$ is the vector of residual errors with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{R})$, such that for a simple model based on single observations $\mathbf{R} = \mathbf{I}\sigma_{\mathrm{e}}^{2}$, and $\mathbf{I}$ is an identity matrix of order equal to the number of phenotypic records and $\sigma_{\mathrm{e}}^{2}$ is the residual error variance. Since the traits investigated here are represented by the mean of a variable number of repeated test day observations, the diagonal elements of $\mathbf{R}$ varied according to the number of observations contributing to the yield deviation. One notable contrast to the model previously implemented in Reynolds et al. (2021), is that here, we fit both additive ($\mathbf{M}_{\boldsymbol{\alpha}}$) and dominance ($\mathbf{M}_{\boldsymbol{\delta}}$) effects of the genomic markers to adjust for population structure. This modification was made to better control the inflation that was observed when analysing milk traits in a population larger than that studied in Reynolds et al. (2021).

*Adjustment of population structure*

Five hundred samples of vectors of plausible additive and dominance SNP effects, $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\delta}}$, were generated for the 31,451 SNPs using single-site Gibbs sampling based on the BayesC0 algorithm implemented in the GenSel program using standard priors (Fernando and Garrick 2013). The fitted model omitted the $\mathbf{Tb}$ term from Eq. (1) and the convergence of the Markov chain of plausible SNP effects was determined using the

Geweke diagnostic (Geweke 1991). The LOSO approach was used to avoid fitting effects of nearby SNPs that are in linkage disequilibrium with the sequence variant being tested. The genome was partitioned into 10-Mb LOSO intervals and, for each interval, phenotypes were adjusted for the samples of SNP effects except for those within the relevant LOSO interval. This produced distinct LOSO-adjusted phenotypic deviations for each 10-Mb interval for each sample of plausible SNP effects.

*Association analysis*

We sampled the effects of genotype classes for each sequence variant separately, for every plausible sample of LOSO-adjusted phenotypic deviations. We obtained MCMC chains of additive and dominance genotypic effects, and standard-additive effects as contrasts of these plausible effects of genotype classes. The posterior distributions were summarised in terms of their posterior means, posterior standard deviations, and z-statistics that assumed a standard normal distribution (Bernal Rubio et al. 2015). The statistical significance of standard-additive, additive, and dominance genetic effects were evaluated using a Z-test.

### 5.3.9 QTL identification, significance criteria, and annotation

Our primary aim was to detect non-additive QTL, thus we declared variants as significant if the dominance genotypic effect, $d$, passed a false discovery rate (FDR) threshold of $1 \times 10^{-3}$. For each phenotype, this FDR threshold was calculated using q-values (Storey and Tibshirani 2003) as implemented in the *qvalue* package in R (Storey et al. 2020). Since we were particularly interested in medium- to large-effect QTL, only the loci with effect sizes ($a$ or $d$) greater than 5% of the phenotypic standard deviation of the trait were considered for further downstream analyses. We calculated the dominance coefficient $k =$

$\frac{d}{|a|}$ for each significant QTL to characterise the underlying non-additive mechanism where $k \approx 0$ represents a completely additive locus, $k \approx 1$ a completely recessive locus, $k < 1$ a partially dominant locus, and $k > 1$ an over-dominant locus. For standard additive effects, α, we used GCTA-COJO (Yang et al. 2012) to detect tag variants for QTL identified in our standard additive GWAS. The GCTA-COJO routine uses LD structure and GWAS summary statistics to iteratively identify significant QTL at the FDR threshold of $1\times10^{-3}$.

We used sequence annotations from variant effect predictor (VEP; Ensembl 97, (McLaren et al. 2016)) to highlight mutations that might be responsible for the non-additive QTL identified, and then used SIFT scores to evaluate the potential impact of any missense mutations on protein function (Ng and Henikoff 2003). To assess the quality of VEP-derived variant annotations and ensure that the predicted missense and nonsense variants intersected expressed exons, we manually visualised mammary RNA-seq alignments as described in Reynolds et al. (2021) using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011). These analyses confirmed that, for the three nonsense candidate mutations identified in *ITGAL*, *LRCH4*, and *RBM34*, all appeared to encode valid premature stop variants, and in the case of the *LRCH4* mutation, its position that is adjacent to the exon 3 splice acceptor boundary suggested that the variant might also have splicing consequences.  We also manually inspected genome sequence alignments representing the non-additive QTL regions in animals of contrasting QTL genotypes (i.e. those carrying opposing alleles of the QTL tag SNPs),  to look for possible gene-disrupting structural variants in these regions.

### 5.3.10 Iterative GWAS

We were interested in determining if multiple dominance QTL might segregate at associated loci, thus we implemented an iterative GWAS approach to differentiate QTL. First, we identified on each chromosome the variants with an FDR lower than the threshold. Next, we adjusted the phenotype for the effects of the genotype classes of the most significant variant (or candidate causal variant if identified) and then re-ran the GWAS model on the chromosome of interest using the adjusted phenotype. This process was iteratively repeated until no significant QTL remained on the chromosome.

## 5.4 Results

### 5.4.1 Heritabilities of lactation traits

First, we estimated the additive and dominance heritabilities for each phenotype within each breed to investigate the additive and non-additive genetic architecture of each trait. These results (Table 5.1) show that the dominance heritabilities were far outweighed by the additive heritabilities. This was not surprising as the values presented are of similar magnitude to those reported for other traits and populations in the literature (Sun et al. 2014; Jiang et al. 2017). Milk fat yield in Jersey cows had the highest dominance heritability at 0.074, and milk protein percentage in Holstein–Friesian cows had the lowest dominance heritability at 0. It should be noted that there was a clear contrast between the relative heritabilities of milk composition and milk yield traits, with milk composition traits displaying high additive heritabilities but near to zero dominance heritabilities, whereas milk yield traits displayed lower additive heritabilities but higher dominance heritabilities (Table 5.1).

**Table 5.1 | Heritability estimates for lactation traits**

| Trait | $h^2_{\mathrm{HF}}$ (SE) | $\delta^2_{\mathrm{HF}}$ (SE) | $h^2_{\mathrm{J}}$ (SE) | $\delta^2_{\mathrm{J}}$ (SE) |
|---|---|---|---|---|
| Milk volume | 0.296 (0.010) | 0.044 (0.007) | 0.312 (0.012) | 0.064 (0.009) |
| Milk fat yield | 0.261 (0.010) | 0.059 (0.008) | 0.232 (0.012) | 0.074 (0.010) |
| Milk protein yield | 0.235 (0.009) | 0.053 (0.008) | 0.236 (0.012) | 0.073 (0.010) |
| Milk fat percentage | 0.700 (0.007) | 0.006 (0.004) | 0.616 (0.010) | 0.015 (0.006) |
| Milk protein percentage | 0.642 (0.008) | 0 (0.005) | 0.636 (0.010) | 0.005 (0.005) |

$h^2$: additive heritability: $\delta^2$: dominance heritability; HF: Holstein-Friesian, J: Jersey; SE: standard error

### 5.4.2 GWAS for lactation traits

We performed GWAS across the five milk traits of interest, namely milk volume, milk protein yield, milk fat yield, milk protein percentage, and milk fat percentage, to identify non-additive QTL (Figure. 5.1). Both additive and dominance effects are included in these plots, and the iterative analysis identified 23 dominance QTL signals that were above the FDR threshold of 1×10$^{-3}$. Some of the QTL were identified for multiple traits. These dominance QTL included 10, 11, 12, 8, and 7 QTL from 4618, 2706, 8525, 8987, and 5800 significant variants for milk volume, milk protein yield, milk fat yield, milk protein percentage, and milk fat percentage, respectively. The QTL spanned 13 discrete autosomes. After the iterative COJO analysis, standard additive GWAS identified 217, 152, 142, 673, and 457 QTL for milk volume, milk protein yield, milk fat yield, milk protein percentage, and milk fat percentage, respectively.

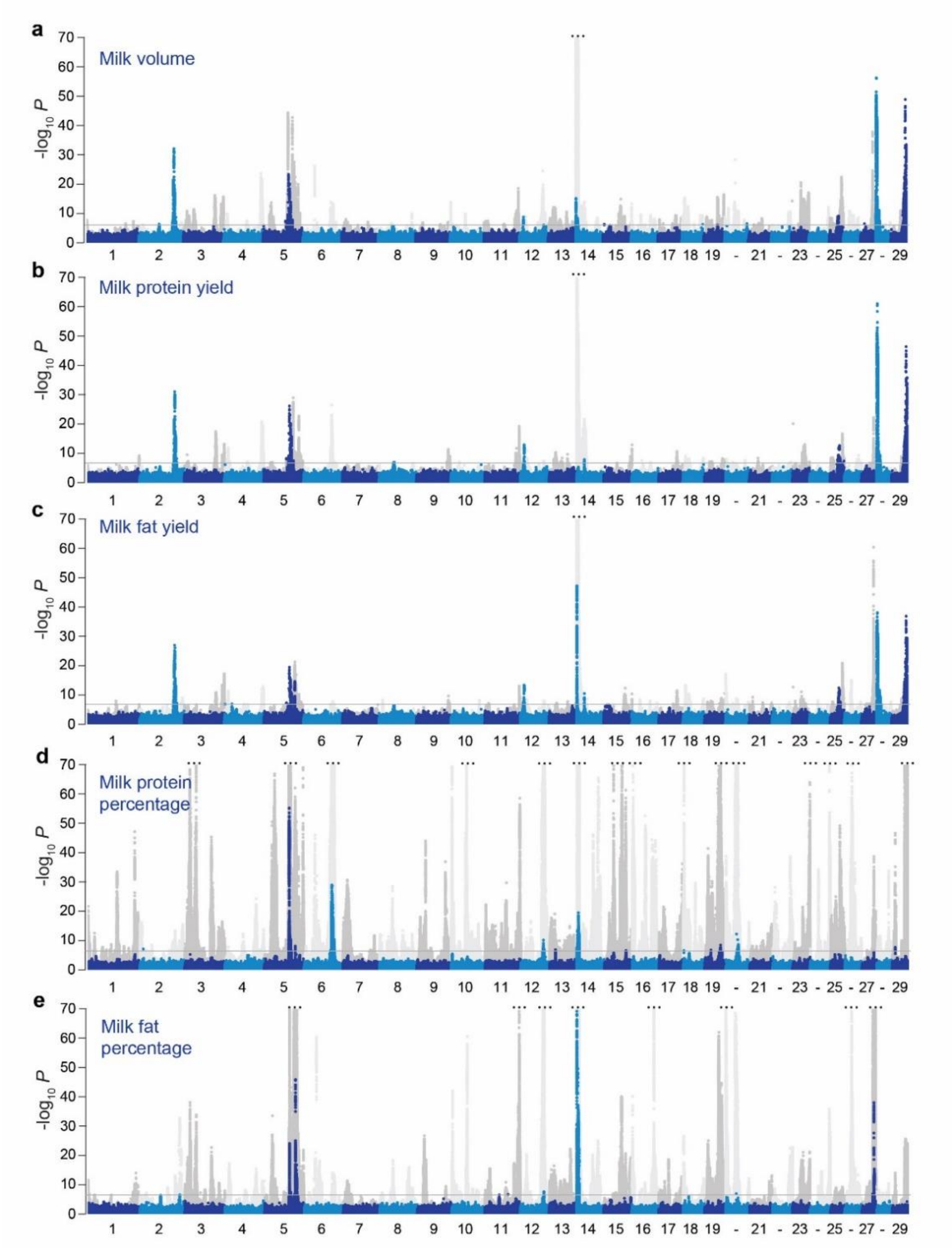


**Figure 5.1 | Manhattan plots of genotypic additive and dominance effects on lactation traits**
Manhattan plots for milk volume (**a**), milk protein yield (**b**), milk fat yield (**c**), milk protein percentage (**d**), and milk fat percentage (**e**) showing significance of genotypic dominance (blue and light blue), and additive (grey and light grey) estimates for ~16.6 million imputed sequence variants. Chromosomes are differentiated by alternating colours and a grey line indicates the false discovery rate of 1×10$^{-3}$, used to account for multiple testing. The y-axes are truncated for display purposes (indicated by 3 dots); chromosome numbers are shown on the x-axis (labels for chromosomes 20, 22, 24, 26 and 28 are not shown for clarity of display).

### 5.4.3 Dominance QTL

We identified 15 significant dominance QTL for milk yield traits, and 11 for milk composition traits (Table 5.2 and see Additional file 1: Table S1 at https://doi.org/10.1186/s12711-021-00694-3). Twelve of the 15 milk yield dominance QTL had recessive effects and were located on chromosomes 2, 4, 5, 8, 12, 25, 28, or 29. Seven of these signals did not appear to have been previously reported, whereas the remainder were highlighted in our analysis (Reynolds et al. 2021) of growth and developmental traits in a population that overlapped with that described here. Eight of the 11 milk composition dominance QTL presented partial dominance effects of which six were identified in our previously published additive GWAS (see Additional file 1: Table S1). Figure 5.2a compares the minor allele frequency and the size of the effect of the dominance components for all these loci. Interestingly, milk composition QTL appeared to be tagged by high minor allele frequency variants with comparatively small effects, whereas milk yield QTL were tagged by variants that had low minor allele frequencies and large effects. The type of effects also appeared to differ between traits (Figure. 5.2b), where milk yield traits were mostly impacted by recessive QTL, whereas milk composition traits near-exclusively presented QTL showing partial dominance.

**Table 5.2 | Association statistics for candidate mutations at recessive loci.**

| Trait | **QTL** | Chr8_44 Mb | Chr25_24-27 Mb | | | Chr25_35 Mb | Chr27_15 Mb | Chr28_6-7 Mb |
|---|---|---|---|---|---|---|---|---|
| | **Position** | g.8.44119667T>A | g.25.24904939C>T | g.25.25161613G>A | g.25.26689392G>A | g.35975573C>T | g.27.15491451C>T | g.28.7922207G>A |
| | **rsID** | rs483207034 | rs453138457 | rs471945767 | rs1116814780 | NA | rs523126258 | NA |
| | **Candidate Gene** | *DOCK8* | *IL4R* | *KIAA0556* | *ITGAL* | *LRCH4* | *SLC25A4* | *RBM34* |
| | **VEP** | AA substitution | AA substitution | AA substitution | Premature stop | Premature stop | AA substitution | Premature stop |
| | **Protein Impact** | p.His649Leu | p.Pro151Leu | p.Arg158His | p.Trp731* | p.Arg123* | p.Thr197Met | p.Arg55* |
| | **SIFT score** | 0 | 0.02 | 0.14 | NA | NA | 0.01 | NA |
| | **MAF (HF /J /ALL)** | 0.013 / 0.059 / 0.03 | 0.001 / 0.043 / 0.017 | 0.001 / 0.042 / 0.016 | 0.002 / 0.049 / 0.019 | 0.034 / 0.001 / 0.031 | 0.046 / 0.001 / 0.027 | 0.044 / 0.004 / 0.043 |
| Milk (L/Lactation) | *a* ± sd | -129.181 ± 23.604 | -218.249 ± 39.988 | -279.656 ± 49.108 | -169.491 ± 37.441 | -153.832 +/- 24.201 | -123.607 ± 25.598 | -106.454 ± 17.786 |
| | p | 4.43E-08 | 4.82E-08 | 1.24E-08 | 5.99E-06 | 2.05E-10 | 1.38E-06 | 2.16E-09 |
| | *d* ± sd | 109.644 ± 23.905 | 215.668 ± 40.648 | 269.952 ± 49.887 | 161.062 ± 37.587 | 97.084 +/- 245.537 | 120.056 ± 25.895 | 106.246 ± 17.929 |
| | p | 4.51E-06 | 1.12E-07 | 6.26E-08 | 1.83E-08 | 7.60E-05 | 3.55E+06 | 3.10E-09 |
| | *k* | 0.849 | 0.988 | 0.965 | 0.95 | 0.63 | 0.971 | 0.998 |
| Fat (kg/Lactation) | *a* ± sd | -5.643 ± 1.177 | -11.827 ± 2.109 | -15.569 ± 2.359 | -9.708 ± 1.870 | -6.849 +/- 1.137 | -7.075 ± 1.201 | -5.170 ± 0.866 |
| | p | 1.66E-06 | 2.05E-08 | 4.10E-11 | 2.09E-07 | 1.71E-09 | 3.84E-09 | 2.40E-09 |
| | *d* ± sd | 5.110 ± 1.181 | 11.339 ± 2.087 | 14.744 ± 2.372 | 9.022 ± 1.910 | 4.412 +/- 1.133 | 5.729 ± 1.249 | 5.546 ± 0.859 |
| | p | 1.51E-05 | 5.56E-08 | 5.08E-10 | 2.33E-06 | 9.82E-05 | 4.48E-06 | 1.06E-10 |
| | *k* | 0.906 | 0.959 | 0.947 | 0.929 | 0.64 | 0.809 | 1.073 |
| Protein (kg/Lactation) | *a* ± sd | -4.981 ± 0.870 | -9.226 ± 1.616 | -11.885 ± 1.834 | -7.847 ± 1.374 | -5.498 +/- 0.838 | -5.008 ± 0.944 | -3.539 ± 0.587 |
| | p | 1.05E-08 | 1.12E-08 | 9.23E-11 | 1.11E-08 | 5.49E-11 | 1.14E-07 | 1.60E-09 |
| | *d* ± sd | 4.308 ± 0.897 | 9.023 ± 1.631 | 11.435 ± 1.829 | 7.497 ± 1.389 | 4.067 +/- 0.844 | 4.595 ± 0.949 | 3.695 ± 0.592 |
| | p | 1.56E-06 | 3.14E-08 | 4.02E-10 | 6.77E-08 | 1.43E-06 | 1.30E-06 | 4.29E-10 |
| | *k* | 0.865 | 0.978 | 0.962 | 0.955 | 0.74 | 0.917 | 1.044 |

VEP: variant effect predictor; NA: Not applicable or unknown; AA substitution: amino-acid substitution; *a*: genotypic additive effect; *d*: genotypic dominance effect; *k*: dominance coefficient; sd: standard deviation; p: p-value; MAF: minor allele frequency; HF: Holstein-Friesian; J: Jersey; ALL: all animals. Linkage values with top variants are in Additional file 1: Table S1 at https://doi.org/10.1186/s12711-021-00694-3.

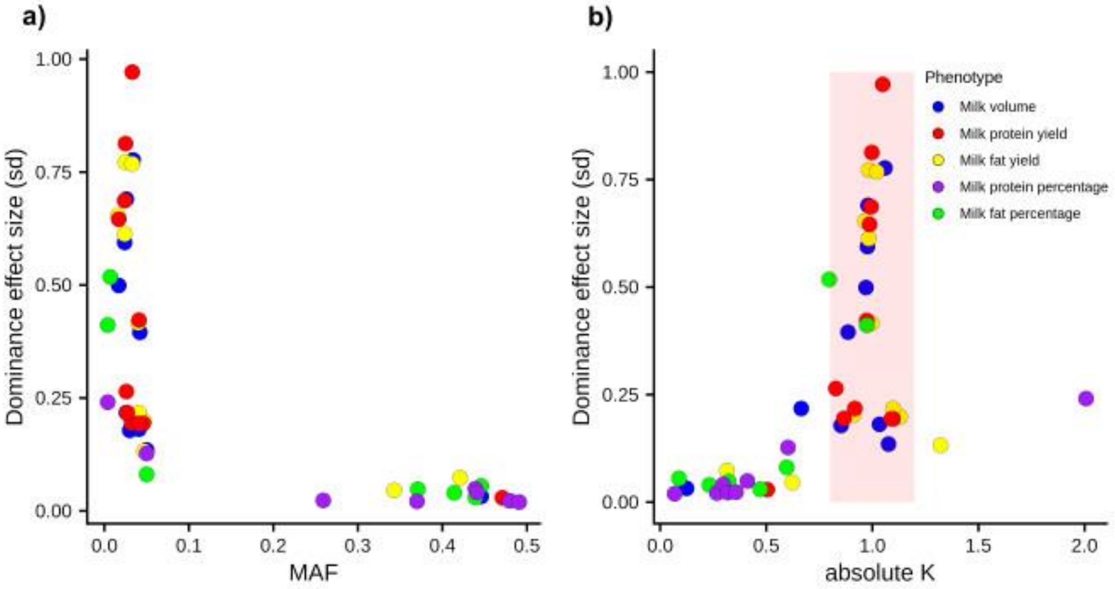


**Figure 5.2 | Plots presenting the genetic architecture of significant dominance QTL from GWAS on lactation traits.**

The plots contrast the minor allele frequency (MAF) against the dominance effect size (**a**), and the absolute value of k, where $k = d/|a|$ against the dominance effect size (**b**) across five lactation traits: milk volume, milk protein yield, milk fat yield, milk protein percentage, and milk fat percentage.

### 5.4.4 Identification of candidate causal mutations

Given that the recessive milk yield QTL potentially represented novel bovine disorders, we prioritised these QTL for further investigation and selected those for which the dominance coefficient ($k$) was near 1 (0.7 < $k$ < 1.3). We used sequence annotations from VEP to highlight the mutations that might be responsible for these effects (Ensembl 97, (McLaren et al. 2016)), i.e. pinpointing variants that were in strong to moderate LD ($R^2$ > 0.7) with the lead variant *per locus*, and that were also predicted to alter or disrupt protein function. Furthermore, we manually investigated each QTL by visualising the whole-genome sequence alignments that corresponded to animals with contrasting QTL genotypes. This step was performed to identify obvious structural mutations that were not detected by automated variant calling, i.e. those intersecting genes that could be similarly expected to modify or ablate gene function.

However, we did not identify any structural variants that tagged QTL. It should be noted that these methods focussed only on protein-coding variants as candidates since, for recessive signals at least, we consider that protein altering mutations are primary candidates given the loss of function mechanism assumed to underlie recessive QTL. However, this does not preclude the involvement of regulatory variants, which we did not consider in our study. We identified five novel recessive QTL (including one near-significant recessive QTL), and several other previously identified recessive effects attributed to mutations in the *PLCD4*, *FGD4*, *MTRF1*, *GALNT2*, *DPF2*, and *MUS81* genes (Reynolds et al. 2021). Figure 5.3 presents the position, regional LD, and association statistics for the QTL that are novel to this paper. Additional File 1: Table S1 shows all significant QTL identified, including those that are not described in detail here.

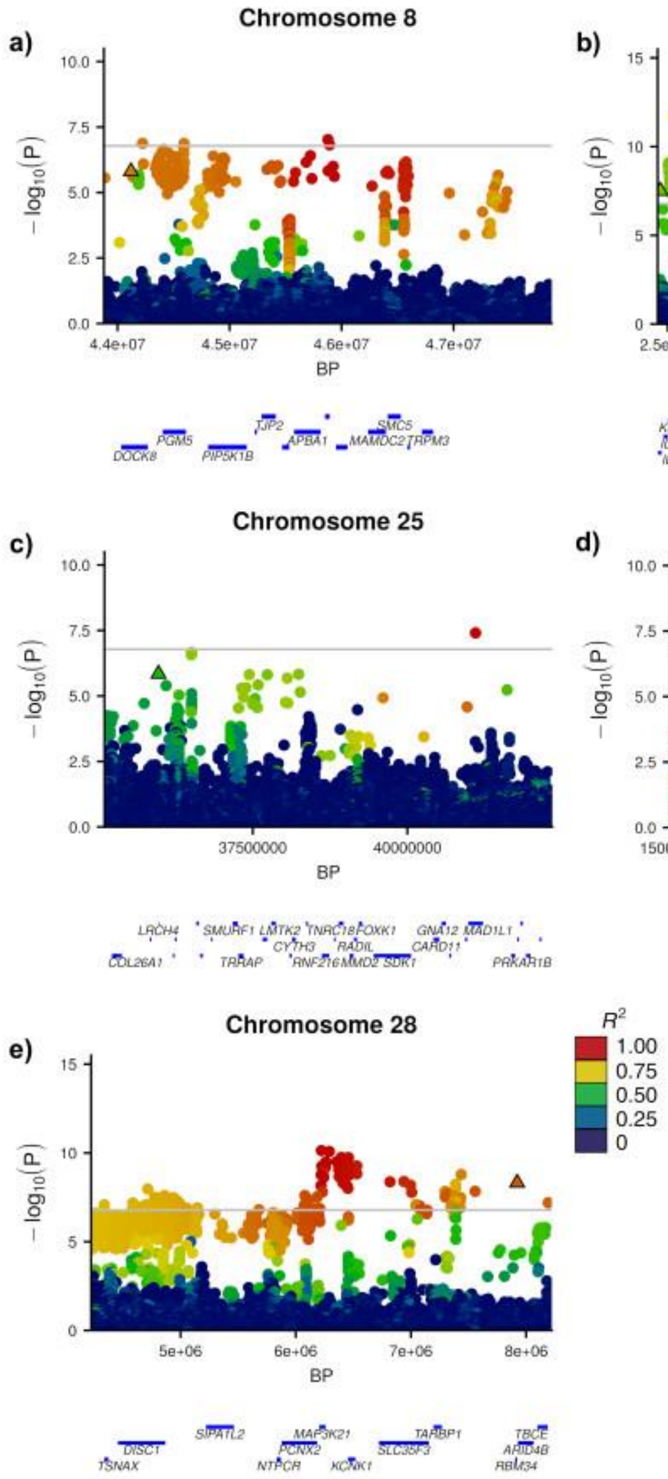


**Figure 5.3 | Manhattan plots for the five novel milk protein yield QTL representing the chr8:44Mbp (a), chr25:24-27Mbp (b), chr25:35Mbp (c), chr27:15Mbp (d), and chr28:7Mbp (e) loci.**

Variants are coloured by LD ($R^2$) values with the top tag variant per locus, protein coding variants are shown as outlined triangles. Gene tracks are presented below each plot based on Ensembl 97, where gene names have been filtered on size.

*Chromosome 8*

Chromosome 8 presented a significant signal at 45 Mb for milk protein yield and milk fat yield. The most significant variants for these signals (g.45878531A>C and g.45880948C>T) were in strong LD ($R^2$ = 0.99), and an annotated missense variant (g.44119667T>A, rs483207034) was in high LD with both of the top-associated variants ($R^2$ = 0.85 and 0.85, respectively; Fig. 3a). This variant in the *DOCK8* gene results in an amino acid (p.His649Leu) change and has a predicted deleterious impact (SIFT = 0).

*Chromosome 25*

A dispersed QTL signal was found on chromosome 25 spanning 24-27 Mb for the three lactation yield traits. The region presented different top-associated variants for milk fat yield (g.25921991AT>T) and milk protein yield and volume traits (g.27868969C>T). Variant effect prediction highlighted three candidate causal mutations in the region. These included a p.Pro151Leu substitution in the *IL4R* gene (g.24904939C>T, rs453138457) with $R^2$ = 0.74, and 0.62, for the milk fat and milk protein/milk volume top variants, respectively, another missense variant (p.Arg158His) in the *KIAA0556* gene (g.25161613G>A, rs471945767) with $R^2$ = 0.89, and 0.74, respectively, and a nonsense variant (p.Trp731*) in the *ITGAL* gene (g.26689392G>A, rs1116814780) with $R^2$ = 0.76, and 0.70, respectively (Figure 5.3b). Although all these variants represented plausible candidates to explain the QTL, we were not able to distinguish between the candidates through iterative analysis, since when any one of these candidates was fitted, the majority of the association for any of the other candidates was removed at this locus.

A second signal for protein yield on chromosome 25 was observed at 35 Mb. That locus maintained its significance after accounting for the QTL on chromosome 25 at

24-27 Mb through iterative analysis, suggesting that it was a different effect. The locus presented a strong candidate causative mutation that could underlie the effect, i.e. a stop gain mutation (g.35975573C>T; Arg123*) in the *LRCH4* gene that was the third most highly associated variant at this locus overall (Figure 5.3c). We observed a mostly recessive effect for this variant ($k$ = 0.74), with the animals that carried the heterozygous and homozygous alternate genotypes producing 1.44 kg, and 11.21 kg less milk protein per lactation compared to the homozygous reference genotype. When g.35975573C>T was fitted as a fixed effect, the significance of the QTL was removed, and no other QTL was detected on the chromosome (Extended Data Figure 5.1).

*Chromosome 27*

We observed a signal at 15 Mb on chromosome 27 for milk protein yield. Although this did not exceed our q-value FDR threshold of $1\times10^{-3}$ (equivalent to p-value = $1.65\times10^{-7}$), this signal was notable given that the top variant (g.15491451C>T; rs523126258, p-value = $1.30\times10^{-6}$) is a predicted deleterious missense mutation (p.Thr197Met) in the *SLC25A4* gene. Figure 5.3d shows a Manhattan plot for this region.

*Chromosome 28*

We previously reported a major recessive bodyweight QTL on chromosome 28 that corresponds to a likely causative splice acceptor mutation in the *GALNT2* gene (g.2281801G>A) (Reynolds et al. 2021). This QTL was observed in the current analysis and impacted all three milk yield traits. However, iterative association analysis revealed a secondary QTL that is located approximately 4 Mb downstream of the *GALNT2* mutation at Chr28:6-7 Mb (top variant at g.6223350G>A). This residual signal highlighted a stop-gain nonsense mutation (g.7922207G>A) that is strongly linked to the g.6223350G>A variant ($R^2 = 0.89$; Figure 5.3e). This stop-gain mutation (p.Arg55*) is located in the *RBM34* gene, appears to be in linkage equilibrium with the *GALNT2* causal mutation ($R^2 < 0.001$), and was not associated with bodyweight in our previous analysis (p = 0.37;(Reynolds et al. 2021)). A second GWAS iteration on chromosome 28 (fitting both *GALNT2* and *RBM34* mutations as fixed effects) did not reveal any other significant QTL on the chromosome (Extended Data Figure 5.2).

### 5.4.5 Comparison between lactation and growth trait recessive QTL

We were interested in determining whether the novel recessive candidate causal mutations identified here had effects on the growth and developmental traits investigated in our previous study (Reynolds et al. 2021). Here, we assessed the association statistics of these variants reported in that study, and while none of the novel mutations reached statistical significance (and would have thus been reported as part of that analysis), some did display apparent recessive mechanisms of moderate effect size. This suggests that, with increased sample sizes, these variants may present significant effects on growth traits. Notably, the mutation in *KIAA0556* was one of the most strongly associated variants for body condition score in that study, presenting the 10th smallest dominance p-value of the ~16 million variants tested in that analysis. Supplementary Table 5.1 includes the association statistics for five of the seven candidate causal mutations presented above (the *ITGAL* and *SLC25A4* mutations were not captured in the genotype dataset reported by Reynolds et al. (2021)). All of the novel candidate mutations highlighted in Reynolds et al. (2021) were also associated with lactation traits (Additional file 1: Table S1) except for the *MYH1*-disrupting structural variant which was only associated with body condition score in that study.

### 5.4.6 Dominance QTL for composition traits

In addition to the recessive QTL identified for milk yield traits, we also identified dominance QTL for milk composition traits. We investigated these effects and observed several partial dominance QTL that are in close proximity to previously described additive loci. The tag variants of these QTL were adjacent to the following genes: *CSF2RB* (Lopdell et al. 2019), *MGST1* (Littlejohn et al. 2016), *DGAT1* (Grisart et al. 2002), *GHR* (Blott et al. 2003), *GPAT4* (Littlejohn, Tiplady, et al. 2014), and *PICALM* (Lopdell et al. 2017) and, in each case, these variants were in high linkage disequilibrium ($R^2 > 0.8$) with previously identified causal and/or tag variants (Additional file 1:Table S1).

Milk protein percentage presented multiple dominance QTL on chromosome 6 within the 80 to 85 Mb region (Additional file 1: Table S1). Among these QTL, the most significant variant (g.84112451C>A) showed a partial dominance effect. Unlike in the above examples, we did not identify any very strongly linked candidate mutation although this variant was in moderate LD with a previously proposed causative variant in the *CSN1S1* gene ($R^2 = 0.53$; p.Glu192Gly mutation; g.85427427A>G) (Caroli, Chessa, and Erhardt 2009). Chromosome 12 presented a significant dominance QTL, for which we observed a partial dominance effect at 68 Mb for milk protein percentage with the top variant at g.68763031T>TG. As observed for the chromosome 6 locus, no particularly obvious candidate causal variant or gene was identified that might account for that signal.

### 5.4.7 Comparison between the additive and dominance GWAS results

Figure 5.4 compares the minor allele frequency (MAF) and the effect sizes between homozygous genotypes across all traits and genetic mechanisms. As expected, we observed many more additive QTL than dominance QTL across all traits. On the one hand, it is noteworthy that the mutations detected via dominance GWAS for milk yield traits had very large effects compared to the additive QTL detected for these traits, and most of them had a recessive effect. On the other hand, the largest effects observed for the two milk composition traits were mostly additive QTL, and dominance effects tended to have high MAF and presented mostly partial dominance effects.

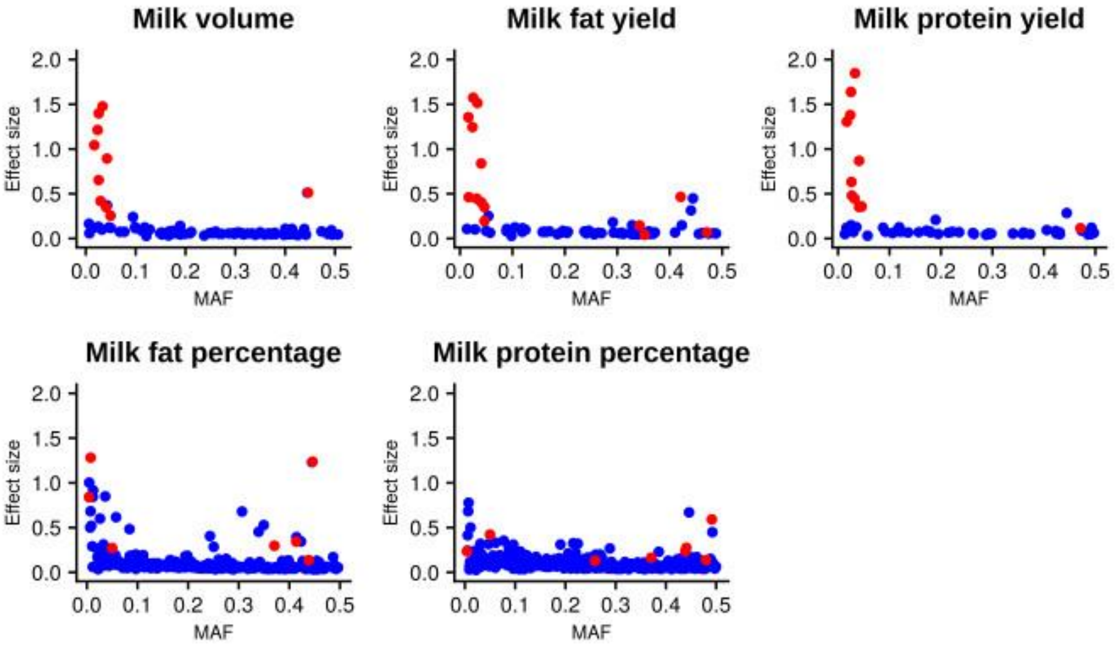


**Figure 5.4 | Plots contrasting minor allele frequency and effect size for different genetic mechanisms.**

Plots of minor allele frequency (MAF) and the absolute effect size between homozygote genotype classes (effect size) for additive (blue) and dominance (red) QTL detected via GWAS across lactation traits.

## 5.5 Discussion

Our results highlight the presence of many non-additive QTL for milk traits in cattle. The majority of these signals for milk yield traits present recessive QTL, that involve five novel loci and several previously described recessive loci (Reynolds et al. 2021). Although the milk protein percentage and milk fat percentage traits also yielded many dominance GWAS signals, most of them correspond to partially dominant effects that are attributable to previously reported additive QTL.

### 5.5.1 Different trait classes present contrasting additive and non-additive genetic architectures

One remarkable observation from our study is the apparent difference in additive and non-additive genetic architectures between milk yield traits and milk composition traits. Dominance heritabilities for the milk yield traits ranged from 3 to 7%, whereas for the milk composition traits they were zero or near zero. In contrast, the additive heritabilities ranged from 23 to 31% for the milk yield traits and from 64 to 70% for the milk composition traits. These findings are consistent with those of Sun et al. (2014) who report similar additive and dominance heritabilities, and suggest that dominance, in particular recessive mechanisms, may play a bigger role in the regulation of milk yield traits than that of composition traits.

These differences in the genetic architecture of the milk traits investigated in this study were also observed when the properties of individual dominance QTL were compared between milk yield and milk composition traits. The majority of the dominance QTL identified for milk yield traits had recessive genetic effects, while the majority of the milk composition traits had partial dominance effects. Furthermore, the dominance QTL for milk yield traits were characterised by low MAF and large effect sizes, whereas those for milk composition traits were characterised by high MAF and comparatively smaller effect sizes. We hypothesize that these observations reflect the way in which different traits may represent underlying recessive syndromes—i.e., their utility as proxies for genetic disorders. Among all the recessive QTL detected in our study, a subset of these had previously been validated as representing new genetic disorders (Reynolds et al. 2021). Although we did not investigate the novel recessive loci in this study with the same rigour as those analysed in Reynolds et al. (2021), their very large, uniformly negative effects suggest that at least some of them will be similarly validated. Notably, none of these loci (new or old) show substantial effects on milk composition, suggesting that milk fat and protein percentage traits do not readily reflect recessive effects. This finding can be rationalised by the comparatively broad range of biological processes expected to impact milk yield traits (or the growth and development traits investigated in Reynolds et al. (2021)), where the energy demands of lactation (or growth) may manifest a wide range of other organismal stresses. In contrast, the relative composition of milk components likely represents a narrower spectrum of mammary-specific biological mechanisms, and thus we hypothesise that these traits are less able to serve as proxies of animal fitness.

It should be acknowledged that given that protein yield and fat yield are the products of milk volume and their respective percentages, these traits are not independent. We observed that the variance components and the genetic architectures of milk fat yield and milk protein yield are more comparable to milk volume than their respective composition traits.

### 5.5.2 Previous studies highlighting recessive effects on quantitative traits

As discussed above, we recently reported an investigation of growth and developmental traits that identified non-additive QTL using similar approaches to those presented here (Reynolds et al. 2021). That study demonstrated how quantitative traits can be used as proxies to map genetic disorders without prior disease identification. In doing so, we highlighted several recessive QTL represented by variants in the *PLCD4*, *FGD4*, *MTRF1*, *GALNT2*, *DPF2*, and *MUS81* genes, each with large effects on bodyweight and other quantitative traits. The work presented in the current paper builds on those findings; we identified many of the same recessive mutations as well as several additional recessive QTL. Some of these additional QTL displayed moderate but not significant recessive effects for growth traits and their discovery may be assumed to reflect the increased sample sizes leveraged in the current study. These findings suggest that milk yield traits might also be used to represent whole-animal health, and since lactation measurements are more routinely derived than bodyweight phenotypes (at least in bovine dairy systems), these likely represent a more accessible phenotype relevant to a larger number of international evaluation systems.

Few studies other than Reynolds et al. (2021) have highlighted major recessive effects using quantitative trait data. Although non-additive GWAS with large sample sizes have been performed in cattle (Jiang et al. 2017; 2019), the low density of the SNP arrays used in those earlier studies may have hampered the ability to directly resolve candidate causative variants (Reynolds et al. 2021). This challenge arises due to the different linkage disequilibrium (LD) properties between causal and observed variants for additive and non-additive QTL, such that the variance that an observed variant can explain decreases by $R^2$ for additive QTL, and by $R^4$ for dominant or recessive QTL. This means that the observed tag variants need to be more closely linked to the causal dominance variants to capture the QTL (Wei, Hemani, and Haley 2014; Visscher et al. 2017). The fact that major deleterious alleles are also likely to be infrequent compounds this problem. Under Hardy–Weinberg expectations where $p^2 + 2pq + q^2 = 1$, the number of rare allele homozygotes ($q^2$) decreases exponentially as allele frequency decreases. Practically, this means very large sample sizes are needed to represent rare allele homozygotes, where at 1% MAF, 10,000 individuals would be expected to present a single homozygote (with 1,000,000 individuals required at MAF = 0.1%). However, as sample sizes and high-density genotyping platforms begin to permit such analyses, we anticipate similar such studies in other populations to begin to appear. One recent, noteworthy such study has suggested the importance of recessive variants in the context of male fertility and semen traits in cattle (Hiltpold et al. 2021). In that study, recessive QTL and candidate causal mutations were identified in several genes including a missense variant in the *SPATA16* gene. That discovery was based on imputed genotypes at high density (the Illumina BovineHD platform), but the size of the studied population was quite small (N = 3736 bulls). It is likely that the discovery of these QTL was partly aided by the remarkable frequency of the deleterious haplotypes identified in that study, presenting allele frequencies ranging from 9-34% (Hiltpold et al. 2021).

### 5.5.3 Recessive QTL of interest

Although many non-additive signals were identified in our study, we were particularly interested in the recessive QTL with large effects, given that these might represent underlying genetic disorders. We highlighted protein-coding variants as candidates because we considered these to be the most probable causal variants, but we acknowledge this is a relatively simple approach and that regulatory or unidentified structural variants may alternatively underlie these recessive QTL. These caveats aside, the five novel recessive QTL on chromosomes 8, 25, 27, and 28 are presented and discussed below.

*Chromosome 8 – DOCK8*

Our results present a missense mutation in the *DOCK8* gene as potentially having a deleterious recessive impact on milk yield traits. The QTL appears to operate in a completely recessive manner, with the *DOCK8* variant present at low allele frequencies in each breed (Holstein-Friesian MAF = 0.013 and Jersey MAF = 0.059).The *DOCK8* gene encodes dedicator of cytokinesis 8, a member of the DOCK180 family of guanine nucleotide exchange factors, which influences intracellular signalling networks and is important in immune responses and lymphocyte regulation in humans and mice (Kearney, Randall, and Oliaro 2017). Recessive mutations in *DOCK8* have been associated with the hyper immunoglobulin E syndrome which leads to the onset of an immunodeficiency disease combined with other health complications (Engelhardt et al. 2009). In mice, compromised immune responses are also observed including negative impacts on B cell migration (Randall et al. 2009), and T cell migration and viability (Lambe et al. 2011; Qian Zhang et al. 2014). *DOCK8* variants have not previously been associated with cattle performance traits, but if this missense mutation underlies the QTL on chromosome 8, we hypothesized that it could act through similar negative impacts on the immune system. Under this hypothesis, it is unknown whether the effects on lactation are due to mammary immune function or secondary impacts. However, given that higher levels of circulating immunoglobulins E and lymphocyte profiling can indicate *DOCK8* deficiency in humans (Janssen et al. 2014; Engelhardt et al. 2009), it would be interesting to sample and profile homozygous animals to definitively establish the causality of the *DOCK8* missense mutation for this QTL.

*Chromosome 25 – IL4R, KIAA0556, ITGAL*

The QTL identified on chromosome 25 at 24-27 Mb presented three candidate mutations in the *IL4R*, *KIAA0556*, and *ITGAL* genes. The *IL4R* gene encodes the interleukin 4 receptor, which is a transmembrane protein involved in immune responses in humans (Shirakawa et al. 2000). The *KIAA0556* gene is associated with microtubule regulation in humans, and *KIAA0556* knockout mutations in humans and mice have been associated with Joubert syndrome, a neurological disorder (Sanders et al. 2015). The *ITGAL* gene encodes the integrin alpha L chain, and loss of function variants in this gene have been associated with compromised immunity including increased susceptibility to infection to Salmonella in mice (J. Zhang et al. 2019). Given that the iterative association analysis failed to prioritise one of these variants over the other, it is unknown which of these variants might be responsible for the QTL, and our focus on protein-coding variants as candidates may have also overlooked alternative non-coding or structural mutations. These variants are nevertheless in moderately strong, though not in perfect LD (maximum pairwise $R^2 = 0.79$), thus physical genotyping for fine mapping and future functional testing should help to resolve the identity of the gene (or genes) underpinning this QTL.

*Chromosome 25 – LRCH4*

Although iterative GWAS did not resolve candidates in the above example, this approach did highlight a second QTL on chromosome 25 represented by a nonsense mutation in the *LRCH4* gene, which encodes leucine-rich repeats and calponin homology containing protein 4. It regulates the signalling of toll-like receptors (TLR) and has been shown to influence innate immune responses in mice (Aloor et al. 2019). In that study, researchers showed that *LRCH4*-silenced cells presented a reduced expression across pro-inflammatory cytokines produced in the TLR4 pathway, most notably in that of IL-10 and MCP-1. We hypothesise that the *LRCH4* knockout mutation identified in our study may have negative impacts on the innate immunity of cattle and those impacts could lead to the recessive effects we observed on milk volume, milk fat yield, and milk protein yield.

*Chromosome 27 – SLC25A4*

While non-significant at the genome-wide level (cf. $P = 1.65\times10^{-7}$ vs $P = 1.30\times10^{-6}$), the locus on chromosome 27 at 15.5 Mb presented a conserved amino acid mutation in the *SLC25A4* gene as the lead associated variant and was therefore of interest. This variant demonstrated a complete recessive effect on all three lactation yield traits. The *SLC25A4* (*solute carrier family 25 member 4*) gene encodes the adenine nucleotide translocator (Ant1) protein, responsible for the translocation of ATP and ADP between the cytoplasm and mitochondria. In mice, *SLC25A4* knockouts result in mitochondrial myopathy and cardiomyopathy, and severe intolerance to exercise (Graham et al. 1997). Similarly, in humans, childhood-onset mitochondrial disease and exercise intolerance have been observed for both dominant (Kaukonen et al. 2000) and recessive mutations (Palmieri et al. 2005) in *SLC25A4*. Given the implication that mitochondrial functional deficits might underlie the negative lactation effects highlighted in the current study, it would be intriguing to examine the phenotypes of homozygous cows further in this context.

*Chromosome 28 – RBM34*

At first glance, the strong associations with the lactation yield traits on chromosome 28 might reasonably be attributed to the previously reported splice site mutation in *GALNT2* (Reynolds et al. 2021). However, when this mutation was fitted as a covariate in our iterative GWAS approach, a secondary QTL was observed, highlighting a nonsense mutation in the *RBM34* gene as potentially responsible for the effect. The *RBM34* gene encodes an RNA recognition motif protein with an RNA-binding domain. The literature on *RBM34* in humans or model organisms is scarce with limited implication of the gene in embryonic stem cell differentiation (X. Wang et al. 2021). Here, we observed a predicted homozygous knockout of *RBM34* that may influence milk volume, milk protein yield, and milk fat yield in a recessive manner, although its status as a largely uncharacterised RNA-binding protein leaves little room for speculation as to how these effects might manifest. Mechanism aside, the identification of two co-locating, yet uncorrelated recessive QTL demonstrates the utility of using iterative GWAS approaches, given that conventional analysis would likely fail to differentiate these effects. We note that other researchers have observed effects on lactation at the 6-10 Mb locus (Raven et al. 2016). However, the LD ($R^2$ with RBM34 =

0.04, GALNT2 = 0.02) between the tag variant identified by Raven et al. (2016) (rs41607517) and the nonsense mutations identified here is very low, which suggests that they are different effects.

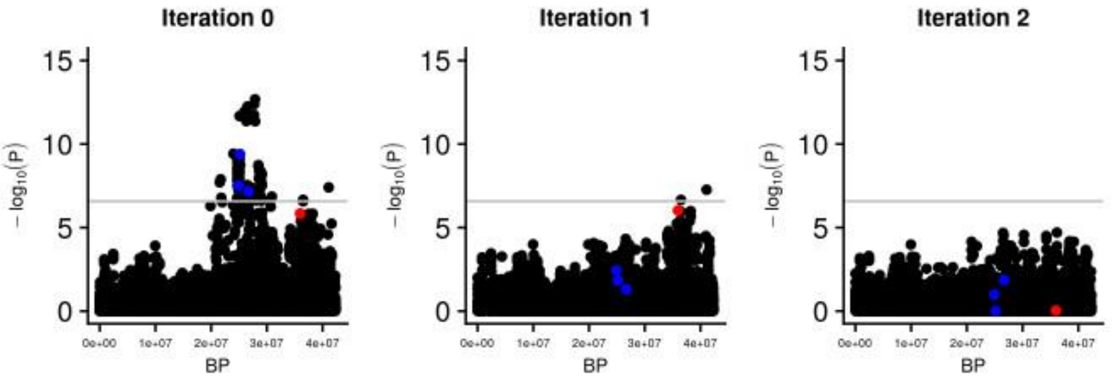
### 5.5.4 Previously described additive QTL present partial dominance

We observed several partial dominance QTL that are closely linked to previously described QTL identified from standard additive analyses. As presented in Additional file 1: Table S1 we identified dominance components in high LD with variants associated with the *CSF2RB* (Lopdell et al. 2019), *MGST1* (Littlejohn et al. 2016), *DGAT1* (Grisart et al. 2002), *GHR* (Blott et al. 2003), *AGPAT6* (Littlejohn, Tiplady, et al. 2014), *PLAG1* (Karim et al. 2011; Fink et al. 2017), and *PICALM* (Lopdell et al. 2017) genes (and in moderate LD with a variant in the *CSN1S1* gene (Caroli, Chessa, and Erhardt 2009)). These partial dominance associations were mostly identified in milk composition traits. These observations suggest that many well-known major-effect QTL that are identified in additive GWAS' incorporate some level of non-additivity, in agreement with the analyses of milk traits reported by Jiang et al. (2017; 2019).

## 5.6 Conclusions

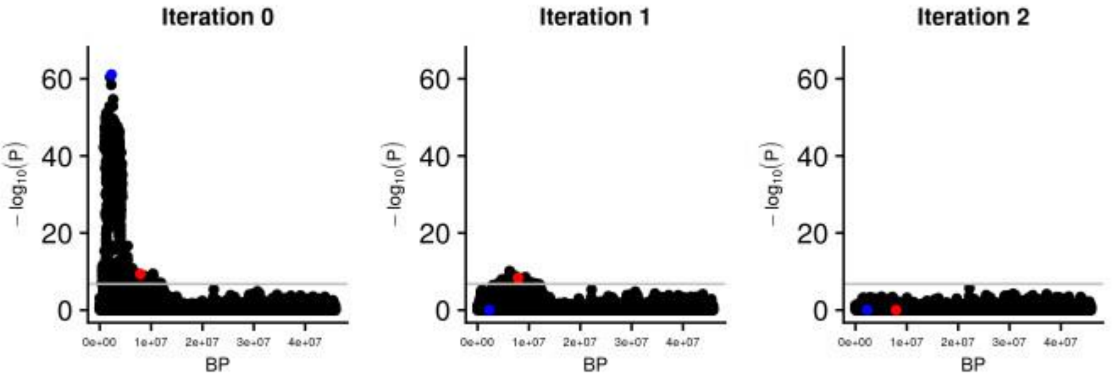
In this study, we have highlighted that different classes of lactation traits (yield compared to composition traits) present different additive and non-additive genetic architectures. We speculate, that these differences derive from dissimilarities in the cellular and molecular manifestation of these traits, and although milk yield traits have comparatively low additive heritabilities, these traits may better reflect whole-animal energy and fitness status and be a better proxy of a wider range of underlying biological disorders. At the single locus level, we identified five QTL presenting seven candidate causative variants in the *DOCK8*, *IL4R*, *KIAA0556*, *ITGAL*, *LRCH4*, *SLC25A4*, and *RBM34* genes, highlighting medium- to large-effect recessive variants that may provide future opportunity for diagnostic testing and animal improvement.

## 5.7 Extended Data Figures



**Extended Data Figure 5.1 | Iterative Manhattan plots for milk protein yield on chromosome 25.**

Blue indicates the candidate causal variants in genes; *IL4R*, *KIAA0556*, and *ITGAL*, and red indicates the candidate causal variant in the *LRCH4* gene. A grey line indicates the false discovery rate of $1\times10^{-3}$, used to account for multiple testing.



**Extended Data Figure 5.2 | Iterative Manhattan plots for milk protein yield on chromosome 28.**

Blue indicates the candidate causal variant in the *GALNT2* gene, and red indicates the candidate causal variant in the *RBM34* gene. A grey line indicates the false discovery rate of $1\times10^{-3}$, used to account for multiple testing.

**Supplementary Table 5.1 | Association statistics for growth traits**

| Phenotype | QTL | Chr8_44Mbp | Chr25_24-26Mbp | | Chr25_35Mbp | Chr28_7Mbp |
|---|---|---|---|---|---|---|
| | Position | g.8.44119667T>A | g.25.24904939C>T | g.25.25161613G>A | g.35975573C>T | g.28.7922207G>A |
| | Candidate Gene | *DOCK8* | *IL4R* | *KIAA0556* | *LRCH4* | *RBM34* |
| Bodyweight | *a* ± sd | -12.262 ± 2.671 | -9.309 ± 4.677 | -24.211 ± 5.640 | -18.669 ± 3.016 | -1.812 ± 1.916 |
| | p | 4.41E-06 | 0.047 | 1.76E-05 | 6.00E-10 | 0.344 |
| | *d* ± sd | 10.861 ± 2.726 | 10.182 ± 4.703 | 24.847 ± 5.693 | 12.903 ± 3.019 | 1.692 ± 1.892 |
| | p | 6.75E-05 | 0.03 | 1.27E-05 | 1.93E-05 | 0.371 |
| Stature | *a* ± sd | -0.874 ± 0.204 | -0.490 ± 0.371 | -1.100 ± 0.474 | -0.831 ± 0.240 | -0.125 ± 0.149 |
| | p | 1.85E-05 | 0.186 | 0.02 | 5.37E-04 | 0.401 |
| | *d* ± sd | 0.754 ± 0.202 | 0.471 ± 0.377 | 1.074 ± 0.478 | 0.593 ± 0.246 | 0.089 ± 0.149 |
| | p | 1.87E-04 | 0.212 | 0.025 | 0.016 | 0.551 |
| Body condition score | *a* ± sd | -0.042 ± 0.019 | -0.076 ± 0.003 | -0.206 ± 0.041 | -0.038 ± 0.021 | -0.003 ± 0.014 |
| | p | 0.029 | 0.024 | 5.18E-07 | 0.065 | 0.835 |
| | *d* ± sd | 0.037 ± 0.019 | 0.084 ± 0.034 | 0.211 ± 0.042 | 0.034 ± 0.021 | -0.019 ± 0.014 |
| | p | 0.057 | 0.013 | 4.05E-07 | 0.102 | 0.178 |

ITGAL and SLC25A4 mutations were not examined in Reynolds et al. (2021). *a*: genotypic additive effect; *d*: genotypic dominance effect; sd: standard deviation; p: p-value.

# Chapter 6 General discussion and conclusion

## 6.1 Introduction

Dairy cattle are an important socio-economic resource across the globe. Technologies such as artificial insemination (AI) and genetic selection have led to tremendous gains in the productivity and efficiency of cattle (Dekkers and Hospital 2002). A key driver for this gain has been garnered by AI facilitating the use of a relatively small number of superior bulls to sire subsequent generations. However, through this intensive selection and past population bottlenecks, the effective population sizes of cattle breeds were small in 2008 small (de Roos et al. 2008) and are likely smaller again today, in 2022. Furthermore, although these superior bulls pass on advantageous gene variants, deleterious variants carried by the bulls may also be transmitted. A widely used bull carrying a novel deleterious allele will lead to a substantial increase in its allele frequency in just one generation, much more than that which occurs in other naturally selected species (Charlier et al. 2016; Georges, Charlier, and Hayes 2018).

The genetic architecture of a trait represents the number, frequency, effect size, and mechanism of all the causal mutations that contribute to variation for a given phenotype (Mackay 2001). Most research on the genetic architecture of complex traits has focussed on additive genetic mechanisms where each allele at each causal locus has some independent effect on the trait of interest. Non-additive genetic mechanisms such as dominance (intra-locus interactions) and epistasis (inter-loci interactions) can also influence traits but have largely been overlooked, principally because they are harder to estimate, more difficult to exploit in an outbreeding programme and appear to contribute a seemingly negligible amount of variance to trait variation (Hill, Goddard, and Visscher 2008). However, as the work described in this thesis has shown, non-additive effects exist and can have large consequences on the phenotypes of individuals, even if contributions to population variances are small.

This thesis has aimed to elucidate the non-additive genetic architecture of growth, developmental, and production traits in dairy cattle. To achieve this goal, we developed and implemented a non-additive GWAS model and algorithm to assess dominance and recessive effects across quantitative traits while accounting for complex population structures (Chapter 2, Chapter 3). We applied this model to growth and developmental traits (Chapter 2), and to lactation traits (Chapter 5) to detect dominance QTL in our dairy cattle population. Through this approach, we identified several recessive QTL that presented candidate causal mutations underlying probable genetic disorders (Chapter 2, Chapter 5). We further investigated these candidates for anatomical, molecular, and metabolic phenotypes to understand how these disorders manifest (Chapter 2, Chapter 4). Here, I discuss the overarching findings of that research including description of the physiological effects of the highlighted mutations and their linkages to analogous disorders in other species. Furthermore, I discuss the key impacts the work may have on farmers and breeding schemes, limitations to the work, and the future of genomics in agriculture in these contexts.

## 6.2 Mutation discoveries linked to genetic disorders

Through GWAS we identified thirteen novel recessive mutations with deleterious effects on growth, production, and welfare traits.  These mutations included premature-stop (*MUS81*, *ITGAL*, *LRCH4*, *RBM34*), splice disrupting (*FGD4*, *GALNT2*), and missense (*PLCD4*, *MTRF1*, *DPF2*, *DOCK8*, *SLC25A4*, *KIAA0556*, *IL4R*) variants (Reynolds et al. 2021; Reynolds et al. 2022). These primarily breed-specific variants, now occurring at surprisingly high frequencies in the New Zealand dairy herd, provide an insight into the potential to account for non-additive effects in the improvement of livestock selection.

Deleterious mutations in *FGD4*, *DPF2*, *GALNT2*, *DOCK8*, *KIAA0556*, and *SLC25A4* gene mutations have been proposed to underlie genetic disorders in humans and mice, where some of the defining physiological features of these conditions appear to be shared in cattle bearing analogous mutations. Variants in *FGD4* can cause Charcot Marie Tooth disease (CMT), the most commonly inherited neurological disorder in humans (Delague et al. 2007; Stendel et al. 2007). Symptoms of CMT include nerve degeneration and muscle wastage that can lead to difficulties in motor control, resulting in individuals being prone to injury (Stendel et al. 2007).  In humans, deleterious mutations in the *DPF2* gene are associated with a developmental disorder called Coffin Siris Syndrome (CSS), with symptoms including intellectual disability and nail abnormalities (Kosho, Miyake, and Carey 2014). Small Calf Syndrome (Cronshaw 2013) is a recessive genetic disorder in cattle caused by a nonsense splice-site mutation in the *GALNT2* gene (Charlier et al. 2016).  Recently, analogous mutations in *GALNT2* in humans have been associated with a novel congenital disorder of glycosylation called GALNT2-CDG (Zilmer et al. 2020), similarly defined by small stature and a host of other abnormalities.

In humans, hyper Immunoglobulin E (hyper-IgE) syndrome can result from analogous recessive mutations in the *DOCK8* gene (Engelhardt et al. 2009).  Resulting from negative impacts on B cell and T cell migration, hyper-IgE syndrome can lead to combined immunodeficiency disease (Randall et al. 2009; Lambe et al. 2011). Knockout mutations in *KIAA0556* have been associated with the recessively inherited cilia-based neurological disorder, Joubert syndrome, in both humans and mice (Sanders et al. 2015).  Finally, knockout mutations in *SLC25A4* cause mitochondrial disease in mice and humans where individuals display a severe intolerance to exercise (Graham et al. 1997; Palmieri et al. 2005). These diseases and disorders in other species present the potential impacts and biological mechanisms through which the discovered mutations might be causing reduced growth and production phenotypes in our cattle population.

For some of the variants, we conducted research farm studies to investigate whether these mutations manifested in a similar way to analogous mutations in humans and model organisms. *FGD4* mutants were 50kgs lighter than controls at 2 years of age, validating our effect estimation from GWAS. Through histology of peripheral nerves, we observed axonal degeneration and Schwann cell demyelination consistent with muscular atrophy and nerve damage, key characteristics of CMT (Bird 1993). Although *DPF2/MUS81* mutants presented lower mature bodyweights, growth rates, and qualitative differences in hoof characteristics, these findings were not significantly different from controls and thus we could not confirm the CSS phenotype. At discovery, the *GALNT2* mutation did not have an analogous disorder in other species and had been previously termed Small Calf Syndrome in cattle (Charlier et al. 2016). However, recently GALNT2-CDG has been proposed as a new human syndrome where individuals typically have lower bodyweights and developmental delays (Zilmer et al. 2020). *GALNT2* mutant cattle displayed slower growth rates and reduced levels of circulating triglycerides and creatine, consistent with observations of human GALNT2-CDG. These results suggest the *FGD4* and *GALNT2* recessive mutations in cattle have similar symptoms and consequences to those displayed in human and mouse disorders.

Several additional recessive mutations were identified which did not have a known analogous disorder in another species. The variant with the largest effects, a mutation in the *PLCD4* gene, caused deleterious impacts across growth, milk yield and conformation traits including 'farmer opinion', a subjective 1 to 10 score of the overall like or dislike of the animal by the farmer (Reynolds et al. 2021). While we could not identify an analogous disorder in another species, the mutation presented some of the largest effect sizes observed in our studies, being similar in magnitude to those seen in individuals affected by Small Calf Syndrome (*GALNT2* mutation). We validated these effects via a research farm trial and so believe *PLCD4* mutant animals have a relatively severe underlying genetic disorder causing these symptoms.

## 6.3 Proxy phenotypes can be used to detect genetic disorders

We proposed that phenotypes that indicate an animal's performance can be used to detect mutations underlying genetic disorders detrimental to individual health. Previous work to detect causal recessive mutations had required disease diagnoses and had focussed on Mendelian disorders such as Brachyspina (Charlier et al. 2012) or retinal degeneration (Michot et al. 2016). However, Reed et al. (2008) presented work showing that the knockout of any of thousands of genes led to reduced bodyweight in mice. That indicated bodyweight might act as a latent variable or proxy phenotype in genetic disorders and might be used to detect the presence of deleterious knockout mutations themselves.

We conducted GWAS on bodyweight and discovered six deleterious recessive mutations. These results suggested bodyweight could be used as a proxy of inherited disease diagnosis to identify genetic disorders in cattle without prior knowledge that the disorders existed. We also investigated how other phenotypes might serve as proxies for genetic disorders and we identified several additional recessive QTL in traits such as stature and milk volume, suggesting that these traits might similarly act as proxies for animal health and wellbeing. While the GWAS approach requires much larger sample sizes, its utility comes from the availability of genotypes and phenotypes gathered via routine commercial activities.  Furthermore, by identifying undiagnosed genetic disorders before they become endemic, we can act to generate a healthier, more productive national herd.

## 6.4 The population structures of cattle present high frequency deleterious mutations

Domestication and breed differentiation bottlenecks have shaped the population history of cattle to where it is today (Felius et al. 2014), resulting in a widely different structure compared to other highly studied species such as humans and mice. The recent intensification of breeding technologies such as artificial insemination has had a dramatic impact on the relatedness structures in cattle populations, where some highly ranked sires have been used for over one million inseminations (Tacon 2002). While this dispersion of gametes from a small number of sires can rapidly increase genetic gain, it also means every deleterious variant carried by such widely used sires propagate throughout the national herd at a concerningly high rate.  In simulation analyses of recessively lethal variants, Charlier et al. (2016) described how cattle, despite carrying fewer embryonic lethal mutations than humans, present mutations that will segregate at a much higher frequency on average. Given that the power to detect a causal mutation is dependent on its allele frequency, these population structures likely empowered our mutation discoveries over what would otherwise be achievable in less inbred species.

The majority of the recessive candidate causal mutations identified in this thesis were breed-specific.  The within-breed allele frequencies ranged from 3.6% (*LRCH4* – Holstein-Friesian) to 11.4% (*MTRF1* – Jersey).  Notably, even the most severe mutations identified had relatively high frequencies (e.g. *GALNT2* at 5.5% in Holstein-

Friesian). Recently, recessive mechanisms impacting male fertility in cattle were discovered at surprising allele frequencies as well, ranging from 9 to 34% in Brown Swiss cattle (Hiltpold et al. 2021). Both studies indicate how the current selection methodology can overlook deleterious variants and allow recessive mutations to reach common frequencies. The summation of these findings presents the importance of investigating non-additive mechanisms to improve selection decisions.

As study sample sizes increase, exploration of the effects of lower frequency variants becomes possible. While we initially believed human populations wouldn't present deleterious variants at detectable frequencies, very recently, Guindo-Martinez et al. (2021) identified three novel recessive effects across disease-related traits including cardiovascular disease and type-II diabetes at allele frequencies of 0.9 to 3.6%. That work suggests that while we expect deleterious variants to segregate at lower allele frequencies in human populations (Charlier et al. 2016), other factors such as increases in sample sizes and disease-specific phenotypes may make these approaches tenable in a human disease context as well.

## 6.5 Different genetic architectures require different models

Genetic architecture is often summarised and studied based on variance component estimation or GWAS techniques (Mackay 2001). While authors have suggested the contribution of non-additive genetic variation is negligible (Hill, Goddard, and Visscher 2008; Crow 2010), recent studies have disputed that claim and significant dominance variance components have been identified in economically important traits in cattle (Sun et al. 2014; Jiang et al. 2017). Furthermore, that dominance genetic variance may be underestimated due to the parameterisation of commonly used models (Huang and Mackay 2016). Across several traits such as bodyweight, stature and milk production traits, we present significant estimates of dominance genetic variance similar to those previously observed (Sun et al. 2014; Jiang et al. 2017). These findings, along with the discovery of large effect recessive mutations, present further evidence that non-additive genetic variants exist, and that we should not ignore looking for them.

Comparing milk yield traits (milk volume, milk-fat yield, and milk protein yield) to milk composition traits (milk-fat percentage and milk protein percentage) we observe differences in the estimate of variance components and the distribution of effect mechanisms identified through GWAS (Chapter 5, (Reynolds et al. 2022)). Milk yield traits presented a much higher proportion of recessive mechanisms compared to milk composition traits, while composition traits present much more additive QTL. The higher heritabilities observed in composition traits compared to yield traits suggest the variation in these phenotypes is more closely regulated by genetics and less perturbed by environmental factors or measurement error. That may explain why composition traits present a greater quantity of additive QTL as the effect of allele substitutions is not masked by environmental interactions in the same way as they are in yield traits. We did not observe QTL in composition traits representing the discovered recessive mutations, which may support the hypothesis that the recessive QTL identified in yield traits represent holistic impacts on animal health rather than direct effects on milk-fat yield or milk protein yield. Different genetic architectures require different models to best understand their underlying genetic mechanisms and to better appreciate what these phenotypes represent. More precise selection protocols and an increased rate of genetic gain may thus be achieved through more accurate description and measurement of the phenotypes we wish to select for.

## 6.6 Limitations

Winner's curse is the phenomenon whereby the effect sizes of significant QTL are overestimated compared to the true genetic effects in the population (Beavis 1994)

(Göring, Terwilliger, and Blangero 2001; Lohmueller et al. 2003). Winner's curse can lead to false positives or failure in replicating results; however, its impact is reduced as sample sizes increase and the problem is alleviated through successful replication of results (Beavis 1994) (Göring, Terwilliger, and Blangero 2001). We were concerned the QTL we discovered were affected by winner's curse and their effect sizes were inflated. Due to the non-random way in which commercially gathered genotypes and phenotypes are obtained (namely that a discrete number of farms is targeted recurrently each year), we were also aware that allele frequencies might differ between our sample and the New Zealand dairy cattle population, and we would have greater power to detect QTL with inflated frequencies.

We made three attempts to investigate these possible biases. First, in Chapter 2 we attempted to overcome the winner's curse problem by validating the effect size and allele frequencies of the discovered candidate causal mutations in a separate, independent dataset. A caveat to this analysis is that it was conducted on a sample closely related to the discovery dataset and, therefore, did not represent a truly random sample from the New Zealand dairy cattle population. Second, research farm trials for affected animals representing the *PLCD4*, *DPF2*, *FGD4*, and *GALNT2* mutations demonstrated similarly significant effects for the *PLCD4*, *FGD4*, and *GALNT2* mutants but not the *DPF2* mutants where the estimated effect size from the research trial on bodyweight was half that of the discovery dataset (Reynolds et al. 2021). Our power to validate effect size may have been limited due to the research farm sample sizes (N = 9 individuals per genotype class) and a larger study may be required to have the power to validate the effect size of the *DPF2* mutation. Finally, Livestock Improvement Corporation (LIC) has recently developed a 1K SNP-chip for parentage testing of calves in the national herd, a panel that includes several of the recessive candidate causal mutations described in this thesis (LIC 2020). Over 400,000 individuals have been genotyped on this panel in the past year and differences in allele frequencies have been observed between those genotyped on the 1K panel and our discovery sample (M. Littlejohn, personal communication, 2021). If the mutant alleles identified are less abundant or have a different impact in the national population, the mutations discovered in this thesis may have reduced relevance. This difference would not discount our findings but is an important reminder that the dramatic effects we observe might not be the same in another independent sample of the same population. Critically, these findings suggest there may be recessive mutations in the greater population that are underrepresented in our sample, and therefore suggests future discoveries will be made as samples accumulate and new populations are targeted.

## 6.7 Future Work

Genetics companies and research organisations have developed customised genotyping panels targeting variants specific to their population. In New Zealand, LIC has recently developed 1K and 50K SNP panels that can be used across hundreds of thousands of animals each year (M. Littlejohn, personal communication, 2021). While these panels have been developed for commercial products such as confirming a calf's parentage or personalised genomic breeding values, variants of interest (including the recessive mutations outlined in this thesis) have been added to the panels (LIC 2020). Ideally, we will be able to avoid the conception of affected calves in the first place, however, preventing dozens of deleterious matings across millions of inseminations is not a trivial problem in practice. In the meantime, the mutation status of animals for the most severe of these variants (*PLCD4*, *FGD4*, *DPF2*/*MUS81*, *GALNT2*, and *KIAA0556)* (LIC. 2021) can be communicated to farmers. This approach will allow farmers to make more informed rearing and selection decisions on young animals, ultimately reducing the prevalence of genetic disorders in their herds.

Several of the mutations described in this thesis have not been validated through a research trial, however, physiological validation of the deleterious *FGD4* and *GALNT2* mutations suggests we might also validate some of the other variants. Future animal trials of the discovered mutations in genes such as *DOCK8*, *KIAA0556*, and *SLC25A4* to test phenotypes relevant to hyper Immunoglobulin E syndrome, Joubert disease, or mitochondrial diseases would indicate if the recessive effects we observe are due to disorders analogous to those observed in humans and mice. These mutations were only detected in milk yield traits so a trial would require milking cows to validate the effects, though, more direct analyses such as blood or muscle biopsy sampling might similarly be applied to validate and explore these effects.

This thesis has focused on identifying intra-locus interactions; therefore, a natural extension of these approaches is to attempt to the detect epistatic inter-locus interactions. Epistatic mechanisms are considered to be important contributors to biological processes (Phillips 2008), yet their contribution to genetic variance is often captured by additive genetic variance and few causal loci have been identified in mammals (Crow 2010; Huang and Mackay 2016). Insufficient statistical and computational power are often cited as barriers to further discovery of epistasis (Mackay 2013; Wei, Hemani, and Haley 2014; Varona et al. 2018) but dramatic increases in sample sizes and continued algorithmic advances may alleviate these roadblocks such that these analyses may now be possible. We propose a targeted QTL approach investigating epistasis between known QTL for quantitative traits. A similar study design in drosophila identified many significant pairwise interactions and showed epistatic effects associated with many previously identified additive QTL (Huang et al. 2012). Exhaustive GWAS' investigating pairwise epistasis across milk production traits in dairy cattle would also be an ambitious (yet intriguing) study and may lead to the discovery of causal mutations. The discovery of epistatic mechanisms might provide additional motivation to advancing non-additive genetic methodology such that we can exploit these mechanisms in breeding and selection decisions, and it might also lead to a better understanding of biological systems.

While we may have discovered primarily negative mutations in this study, it is important to consider the potential impact of desirable mutations as well. Breeding programmes have steadily resulted in genetic gain year on year for decades (Dekkers and Hospital 2002) through the indirect selection of positive genetic variation. Known large-effect genetic variants can be used to directly select desirable phenotypes as well such as polled variants and the A2 β-casein variant. The process of dehorning raises animal welfare concerns which may be alleviated by selecting animals without horns carrying the dominant polled variant (Georges et al. 1993). A2 milk is the product of herds that are homozygous for the A2 β-casein variant (Truswell 2005) and is marketed at a premium, incentivising farmers to select for it. Other novel phenotypes may be discovered to create niche dairy products or improve milk processing efficiency. A recent example is a dominant mutation in *DGAT1* that significantly alters the saturated fat profile in a cow's milk that was discovered using high-throughput phenotypic outlier screening techniques (Lehnert et al. 2015). Although this *DGAT1* mutation caused a disorder in homozygous animals, similar methods could be applied to detect other variants. The examples described here make it desirable to search for genetic mutations which present novel phenotypes.

Beyond the selection of a handful of desirable or deleterious variants with major non-additive effects, non-additive selection in dairy cattle is difficult. Genomic selection methods that leverage non-additive effects offer increased accuracy in genomic breeding value estimation but come with added complexity in calculation and implementation (Varona et al. 2018). These complications are exacerbated by the admixed structure of the New Zealand dairy cattle population and the seasonality of the dairy system.  As more of the national population is genotyped on medium-density SNP chip panels like the 50K panel, strategies for mate allocation may be able to better integrate non-additive contributions and improve the prediction of future offspring. In any case, further research in this area is required to get there.

## 6.8 Conclusion

This thesis has aimed to dissect the dominance relationship between genotype and phenotype and elucidate some of the non-additive genetic architecture of growth, developmental, and production traits in dairy cattle. Through the development and implementation of a non-additive GWAS model, we discovered several deleterious recessive mutations with moderate to high impacts across a range of economically important traits.  In some cases, these mutations were analogous to disorder-causing mutations in other species, and by analysing their physical, metabolic, and molecular impacts, we validated a subset of these effects and confirmed the likely causality of some of the candidate mutations highlighted. As increasing numbers of animals are genotyped and phenotyped, our capability to detect causal mutations will continue to improve.  Overall, these are important findings that can be used to improve the health and productivity of dairy cattle in New Zealand.

# References

Adams, Heather A., Tad S. Sonstegard, Paul M. VanRaden, Daniel J. Null, Curt P. van Tassell, Denis M. Larkin, and Harris A. Lewin. 2016. "Identification of a Nonsense Mutation in APAF1 That Is Likely Causal for a Decrease in Reproductive Efficiency in Holstein Dairy Cattle." *Journal of Dairy Science* 99 (8): 6693–6701. https://doi.org/10.3168/JDS.2015-10517.

Advisory Committee on Traits Other than Production. 2020. "Evaluation System for Traits Other than Production TOP for Dairy Cattle in New Zealand."

Affymetrix Inc. 2005. "Affymetrix MegAllele GeneChip Bovine 10K SNP Array." South San Francisco, CA.: Affymetrix Inc.

Akiyama, Masato, Yukinori Okada, Masahiro Kanai, Atsushi Takahashi, Yukihide Momozawa, Masashi Ikeda, Nakao Iwata, et al. 2017. "Genome-Wide Association Study Identifies 112 New Loci for Body Mass Index in the Japanese Population." *Nature Genetics* 49 (10): 1458–67. https://doi.org/10.1038/ng.3951.

Aliloo, Hassan, Jennie E. Pryce, Oscar González-Recio, Benjamin G. Cocks, and Ben J. Hayes. 2016. "Accounting for Dominance to Improve Genomic Evaluations of Dairy Cows for Fertility and Milk Production Traits." *Genetics Selection Evolution* 48. https://doi.org/10.1186/s12711-016-0186-0.

Aloor, Jim J., Kathleen M. Azzam, John J. Guardiola, Kymberly M. Gowdy, Jennifer H. Madenspacher, Kristin A. Gabor, Geoffrey A. Mueller, et al. 2019. "Leucine-Rich Repeats and Calponin Homology Containing 4 (Lrch4) Regulates the Innate Immune Response." *Journal of Biological Chemistry* 294 (6): 1997–2008. https://doi.org/10.1074/jbc.RA118.004300.

Amberger, Joanna S., Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. 2015. "OMIM.Org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders." *Nucleic Acids Research* 43 (D1): D789–98. https://doi.org/10.1093/NAR/GKU1205.

Andres, Robert H., Angélique D. Ducray, Uwe Schlattner, Theo Wallimann, and Hans Rudolf Widmer. 2008. "Functions and Effects of Creatine in the Central Nervous System." *Brain Research Bulletin* 76 (4): 329–43. https://doi.org/10.1016/J.BRAINRESBULL.2008.02.035.

Avery, O T, C M Macleod, and M McCarty. 1944. "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Tranformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III." *The Journal of Experimental Medicine* 79 (2): 137–58. https://doi.org/10.1084/JEM.79.2.137.

Barton, N. H., A. M. Etheridge, and A. Véber. 2017. "The Infinitesimal Model: Definition, Derivation, and Implications." *Theoretical Population Biology* 118: 50–73. https://doi.org/10.1016/j.tpb.2017.06.001.

Baudat, Frédéric, Yukiko Imai, and Bernard de Massy. 2013. “Meiotic Recombination in Mammals: Localization and Regulation.” *Nature Reviews Genetics* 14: 794-806. https://doi.org/10.1038/nrg3573.

Bauman, Dale E., and W. Bruce Currie. 1980. “Partitioning of Nutrients during Pregnancy and Lactation: A Review of Mechanisms Involving Homeostasis and Homeorhesis.” *Journal of Dairy Science* 63 (9): 1514–29. https://doi.org/10.3168/JDS.S0022-0302(80)83111-0.

Beavis, W D. 1994. “The Power and Deceit of QTL Experiments: Lessons from Comparative QTL Studies.” In *Proceedings of the Forty-Ninth Annual Corn and Sorghum Industry Research Conference*, 250:266.

Becker, Jutta, Oliver Semler, Christian Gilissen, Yun Li, Hanno Jörn Bolz, Cecilia Giunta, Carsten Bergmann, et al. 2011. “Exome Sequencing Identifies Truncating Mutations in Human SERPINF1 in Autosomal-Recessive Osteogenesis Imperfecta.” *The American Journal of Human Genetics* 88 (3): 362–71. https://doi.org/10.1016/J.AJHG.2011.01.015.

Bennewitz, Jörn, Christian Edel, Ruedi Fries, Theo H. E. Meuwissen, and Robin Wellmann. 2017. “Application of a Bayesian Dominance Model Improves Power in Quantitative Trait Genome-Wide Association Analysis” *Genetics Selection Evolution* 49. https://doi.org/10.1186/s12711-017-0284-7.

Bernal Rubio, Y. L., J. L. Gualdrón Duarte, R. O. Bates, C. W. Ernst, D. Nonneman, G. A. Rohrer, A. King, et al. 2015. “Meta-Analysis of Genome-Wide Association from Genomic Prediction Models.” *Animal Genetics* 47 (1): 36–48. https://doi.org/10.1111/age.12378.

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2017. “Julia: A Fresh Approach to Numerical Computing.” *SIAM Review* 59 (1): 65–98. https://doi.org/10.1137/141000671.

Bird, Thomas D. 1993. *Charcot-Marie-Tooth (CMT) Hereditary Neuropathy Overview*. *GeneReviews®*. University of Washington, Seattle. https://www.ncbi.nlm.nih.gov/books/NBK1358/.

Blott, Sarah, Jong Joo Kim, Sirja Moisio, Anne Schmidt-Küntzel, Anne Cornet, Paulette Berzi, Nadine Cambisano, et al. 2003. “Molecular Dissection of a Quantitative Trait Locus: A Phenylalanine-to-Tyrosine Substitution in the Transmembrane Domain of the Bovine Growth Hormone Receptor Is Associated with a Major Effect on Milk Yield and Composition.” *Genetics* 163 (1): 253–66. https://doi.org/10.1093/genetics/163.1.253.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20. https://doi.org/10.1093/BIOINFORMATICS/BTU170.

Bolormaa, Sunduimijid, Ben J. Hayes, Julius H.J. van der Werf, David Pethick, Michael E. Goddard, and Hans D. Daetwyler. 2016. “Detailed Phenotyping Identifies Genes with Pleiotropic Effects on Body Composition.” *BMC Genomics* 17. https://doi.org/10.1186/S12864-016-2538-0.

Bolormaa, Sunduimijid, Jennie E Pryce, Yuandan Zhang, Antonio Reverter, William Barendse, Ben J Hayes, and Michael E Goddard. 2015. “Non-Additive Genetic Variation in Growth, Carcass and Fertility Traits of Beef Cattle.” *Genetics Selection Evolution* 47. https://doi.org/10.1186/s12711-015-0114-8 .

Bourneuf, E., P. Otz, H. Pausch, V. Jagannathan, P. Michot, C. Grohs, G. Piton, et al. 2017. “Rapid Discovery of de Novo Deleterious Mutations in Cattle Enhances the Value of Livestock as Model Species.” *Scientific Reports* 7. https://doi.org/10.1038/s41598-017-11523-3.

Bouwman, Aniek C., Hans D. Daetwyler, Amanda J. Chamberlain, Carla Hurtado Ponce, Mehdi Sargolzaei, Flavio S. Schenkel, Goutam Sahana, et al. 2018. “Meta-Analysis of Genome-Wide Association Studies for Cattle Stature Identifies Common Genes That Regulate Body Size in Mammals.” *Nature Genetics* 50 (3): 362–67. https://doi.org/10.1038/s41588-018-0056-5.

Brito, Fernanda v., José B. Neto, Mehdi Sargolzaei, Jaime A. Cobuci, and Flavio S. Schenkel. 2011. “Accuracy of Genomic Selection in Simulated Populations Mimicking the Extent of Linkage Disequilibrium in Beef Cattle.” *BMC Genetics* 12. https://doi.org/10.1186/1471-2156-12-80.

Browning, Brian L., and Sharon R. Browning. 2009. “A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals.” *American Journal of Human Genetics* 84 (2): 210–23. https://doi.org/10.1016/j.ajhg.2009.01.005.

Browning, Brian L., Ying Zhou, and Sharon R. Browning. 2018. “A One-Penny Imputed Genome from next-Generation Reference Panels.” *American Journal of Human Genetics* 103 (3): 338–48. https://doi.org/10.1016/j.ajhg.2018.07.015.

Browning, Sharon R., and Brian L. Browning. 2007. “Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering.” *The American Journal of Human Genetics* 81 (5): 1084–97. https://doi.org/10.1086/521987.

Buniello, Annalisa, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. “The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019.” *Nucleic Acids Research* 47 (D1): D1005–12. https://doi.org/10.1093/NAR/GKY1120.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, et al. 2018. “The UK Biobank Resource with Deep Phenotyping and Genomic Data.” *Nature* 562 (7726): 203–9. https://doi.org/10.1038/s41586-018-0579-z.

Cai, Zexi, Bernt Guldbrandtsen, Mogens Sandø Lund, and Goutam Sahana. 2019. “Weighting Sequence Variants Based on Their Annotation Increases the Power of Genome-Wide Association Studies in Dairy Cattle.” *Genetics Selection Evolution* 51. https://doi.org/10.1186/S12711-019-0463-9.

Caroli, A. M., S. Chessa, and G. J. Erhardt. 2009. “Invited Review: Milk Protein Polymorphisms in Cattle: Effect on Animal Breeding and Human Nutrition.” *Journal of Dairy Science* 92 (11): 5335–52. https://doi.org/10.3168/JDS.2009-2461.

Champely, Stephane, Claus Ekstrom, Peter Dalgaard, Jeffrey Gilland Stephan Weibelzahl, Aditya Anandkumar, Clay Ford, Robert Volcic, and Helios de Rosario. 2020. “Pwr: Basic Functions for Power Analysis.” https://cran.r-project.org/web/packages/pwr/.

Charlier, Carole, Jorgen Steen Agerholm, Wouter Coppieters, Peter Karlskov-Mortensen, Wanbo Li, Gerben de Jong, Corinne Fasquelle, et al. 2012. “A Deletion in the Bovine FANCI Gene Compromises Fertility by Causing Fetal Death and Brachyspina.” *PLoS ONE* 7 (8): e43085. https://doi.org/10.1371/journal.pone.0043085.

Charlier, Carole, Wanbo Li, Chad Harland, Mathew Littlejohn, Wouter Coppieters, Frances Creagh, Steve Davis, et al. 2016. “NGS-Based Reverse Genetic Screen for Common Embryonic Lethal Mutations Compromising Fertility in Livestock.” *Genome Research* 26 (10): 1333–41. https://doi.org/10.1101/gr.207076.116.

Chebyshev, Pafnutii Lvovich. 1867. “Des Valeurs Moyennes.” *J. Math. Pures Appl* 12 (2): 177–84.

Chen, Sisi, Xinwei Geng, Madiha Zahra Syeda, Zhengming Huang, Chao Zhang, and Songmin Ying. 2021. “Human MUS81: A Fence-Sitter in Cancer.” *Frontiers in Cell and Developmental Biology* 9: 415. https://doi.org/10.3389/FCELL.2021.657305.

Cheng, Hao, Dorian J Garrick, and Rohan L Fernando. 2016. “JWAS: Julia Implementation of Whole-Genome Analyses Software Using Univariate and Multivariate Bayesian Mixed Effects Model.” *QTL. Rocks*.

Cheng, Hao, Rohan L Fernando, and Dorian J Garrick. 2018. “JWAS: Julia Implementation of Whole-Genome Analysis Software.” In *Proceedings of the World Congress of Genetics Applied to Livestock Production*, 11:859. Auckland, New Zealand.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff.” *Fly* 6 (2): 80–92. https://doi.org/10.4161/FLY.19695.

Cockerham, C Clark. 1954. “An Extension of the Concept of Partitioining Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present.” *Genetics* 39 (6): 859–82. https://doi.org/10.1093/GENETICS/39.6.859.

Cockett, Noelle E., Sam P. Jackson, Tracy L. Shay, Frédéric Farnir, Stéphane Berghmans, Gary D. Snowder, Dahlia M. Nielsen, and Michel Georges. 1996. “Polar Overdominance at the Ovine Callipyge Locus.” *Science* 273 (5272): 236–38. https://doi.org/10.1126/SCIENCE.273.5272.236.

Cohen-Zinder, Miri, Eyal Seroussi, Denis M. Larkin, Juan J. Loor, Annelie Everts-Van Der Wind, Jun Heon Lee, James K. Drackley, et al. 2005. “Identification of a Missense Mutation in the Bovine ABCG2 Gene with a Major Effect on the QTL on Chromosome 6 Affecting Milk Yield and Composition in Holstein Cattle.” *Genome Research* 15 (7): 936–44. https://doi.org/10.1101/gr.3806705.

Coop, Graham, and Molly Przeworski. 2007. “An Evolutionary View of Human Recombination.” *Nature Reviews Genetics* 8:23-34. https://doi.org/10.1038/nrg1947.

Cortes, Adrian, Sara L. Pulit, Paul J. Leo, Jenny J. Pointon, Philip C. Robinson, Michael H. Weisman, Michael Ward, et al. 2015. “Major Histocompatibility Complex Associations of Ankylosing Spondylitis Are Complex and Involve Further Epistasis with ERAP1” *Nature Communications* 6. https://doi.org/10.1038/ncomms8146.

Cressie, Noel, and Soumendra Nath Lahiri. 1993. “The Asymptotic Distribution of REML Estimators.” *Journal of Multivariate Analysis* 45 (2): 217–33. https://doi.org/10.1006/JMVA.1993.1034.

Cronshaw, Tim. 2013. “Small-Calf Gene to Be Erased from NZ Herd.” *Press, The*, A13.

Crow, James F. 2010. “On Epistasis: Why It Is Unimportant in Polygenic Directional Selection.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1544): 1241–44. https://doi.org/10.1098/RSTB.2009.0275.

Daetwyler, Hans D, Aurélien Capitan, Hubert Pausch, Paul Stothard, Rianne van Binsbergen, Rasmus F Brøndum, Xiaoping Liao, et al. 2014. “Whole-Genome Sequencing of 234 Bulls Facilitates Mapping of Monogenic and Complex Traits in Cattle.” *Nature Genetics* 46 (8): 858–65. https://doi.org/10.1038/ng.3034.

DairyNZ. 2021. “All about BW.” 2021. https://www.dairynz.co.nz/animal/animal-evaluation/interpreting-the-info/all-about-bw/.

Dekkers, Jack C. M., and Frédéric Hospital. 2002. “The Use of Molecular Genetics in the Improvement of Agricultural Populations.” *Nature Reviews Genetics* 3 (1): 22–32. https://doi.org/10.1038/nrg701.

Delague, Valérie, Arnaud Jacquier, Tarik Hamadouche, Yannick Poitelon, Cécile Baudot, Irène Boccaccio, Eliane Chouery, et al. 2007. “Mutations in FGD4 Encoding the Rho GDP/GTP Exchange Factor FRABIN Cause Autosomal Recessive Charcot-Marie-Tooth Type 4H.” *American Journal of Human Genetics* 81. https://doi.org/10.1086/518428.

Dendouga, Najoua, Hui Gao, Dieder Moechars, Michel Janicot, Jorge Vialard, and Clare H. McGowan. 2005. “Disruption of Murine Mus81 Increases Genomic Instability and DNA Damage Sensitivity but Does Not Promote Tumorigenesis.” *Molecular and Cellular Biology* 25 (17): 7569. https://doi.org/10.1128/MCB.25.17.7569-7579.2005.

DePristo, Mark A., Eric Banks, Ryan Poplin, Kiran v. Garimella, Jared R. Maguire, Christopher Hartl, Anthony A. Philippakis, et al. 2011. "A Framework for Variation Discovery and Genotyping Using Next-Generation DNA Sequencing Data." *Nature Genetics* 43 (5): 491–501. https://doi.org/10.1038/ng.806.

Devlin, B., and Kathryn Roeder. 1999. "Genomic Control for Association Studies." *Biometrics* 55 (4): 997–1004. https://doi.org/10.1111/J.0006-341X.1999.00997.X.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/BIOINFORMATICS/BTS635.

Druet, Tom, and Michel Georges. 2015. "LINKPHASE3: An Improved Pedigree-Based Phasing Algorithm Robust to Genotyping and Map Errors." *Bioinformatics* 31 (10): 1677–79. https://doi.org/10.1093/bioinformatics/btu859.

Engelhardt, Karin R, Sean McGhee, Sabine Winkler, Atfa Sassi, Cristina Woellner, Gabriela Lopez-herrera, Andrew Chen, et al. 2009. "Large Deletions and Point Mutations Involving DOCK8 in the Autosomal Recessive Form of the Hyper-IgE Syndrome." *Journal of Allergy and Clinical Immunology* 124 (6): 1289–1302. https://doi.org/10.1016/j.jaci.2009.10.038.Large.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, and M. E. Goddard. 2012. "Improving Accuracy of Genomic Predictions within and between Dairy Cattle Breeds with Imputed High-Density Single Nucleotide Polymorphism Panels." *Journal of Dairy Science* 95 (7): 4114–29. https://doi.org/10.3168/JDS.2011-5019.

Eu-ahsunthornwattana, Jakris, E. Nancy Miller, Michaela Fakiola, Selma M. B. Jeronimo, Jenefer M. Blackwell, Heather J. Cordell, and Heather J. Cordell. 2014. "Comparison of Methods to Account for Relatedness in Genome-Wide Association Studies with Family-Based Data." *PLOS Genetics* 10 (7): e1004445. https://doi.org/10.1371/journal.pgen.1004445.

Evans, David M, Chris C A Spencer, Jennifer J Pointon, Zhan Su, David Harvey, Grazyna Kochan, Udo Oppermann, et al. 2011. "Interaction between ERAP1 and HLA-B27 in Ankylosing Spondylitis Implicates Peptide Handling in the Mechanism for HLA-B27 in Disease Susceptibility." *Nature Genetics* 43 (8): 761–67. https://doi.org/10.1038/ng.873.

Evershed, Richard P., Sebastian Payne, Andrew G. Sherratt, Mark S. Copley, Jennifer Coolidge, Duska Urem-Kotsu, Kostas Kotsakis, et al. 2008. "Earliest Date for Milk Use in the Near East and South-eastern Europe Linked to Cattle Herding." *Nature* 455 (7212): 528–31. https://doi.org/10.1038/NATURE07180.

Falconer, Douglas S. 1960. *Introduction to Quantitative Genetics*. Edinburgh/London: Oliver & Boyd.

Fan, Bin, Suneel K. Onteru, Zhi-Qiang Du, Dorian J. Garrick, Kenneth J. Stalder, and Max F. Rothschild. 2011. "Genome-Wide Association Study Identifies Loci for Body Composition and

Structural Soundness Traits in Pigs." *PLOS One* 6 (2): e14726. https://doi.org/10.1371/journal.pone.0014726.

Felius, Marleen, Marie-Louise Beerling, David S. Buchanan, Bert Theunissen, Peter A. Koolmees, and Johannes A. Lenstra. 2014. "On the History of Cattle Genetic Resources." *Diversity* 6 (4): 705–50. https://doi.org/10.3390/D6040705.

Fernando, Rohan L, and Dorian Garrick. 2013. "Genome-Wide Association Studies and Genomic Prediction." In *Genome-Wide Association Studies and Genomic Prediction*, 1019:237–74. https://doi.org/10.1007/978-1-62703-447-0.

Fink, Tania, Thomas J. Lopdell, Kathryn Tiplady, Renee Handley, Thomas J. J. Johnson, Richard J. Spelman, Stephen R. Davis, Russell G. Snell, and Mathew D. Littlejohn. 2020. "A New Mechanism for a Familiar Mutation – Bovine DGAT1 K232A Modulates Gene Expression through Multi-Junction Exon Splice Enhancement." *BMC Genomics* 21. https://doi.org/10.1186/S12864-020-07004-Z.

Fink, Tania, Kathryn Tiplady, Thomas Lopdell, Thomas Johnson, Russell G. Snell, Richard J. Spelman, Stephen R. Davis, and Mathew D. Littlejohn. 2017. "Functional Confirmation of PLAG1 as the Candidate Causative Gene Underlying Major Pleiotropic Effects on Body Weight and Milk Characteristics." *Scientific Reports* 7. https://doi.org/10.1038/srep44793.

Finno, Carrie J., Giuliana Gianino, Sudeep Perumbakkam, Zoë J. Williams, Matthew H. Bordbari, Keri L. Gardner, Erin Burns, Sichong Peng, Sian A. Durward-Akhurst, and Stephanie J. Valberg. 2018. "A Missense Mutation in MYH1 Is Associated with Susceptibility to Immune-Mediated Myositis in Quarter Horses." *Skeletal Muscle* 8 (1): 1–13. https://doi.org/10.1186/S13395-018-0155-0.

Fisher, Ronald Aylmer. 1930. *The Genetical Theory of Natural Selection*. Oxford University Press. https://books.google.co.nz/books?hl=en&lr=&id=sT4lIDk5no4C&oi=fnd&pg=PR6&dq=fisher+1930+genetical+theory&ots=oEKb2E2Z4j&sig=F5jqUIMDgh8kWnOrJmcYaatJE34#v=onepage&q=fisher 1930 genetical theory&f=false.

Foote, R H. 2002. "The History of Artificial Insemination: Selected Notes and Notables."

Freking, Brad A., Susan K. Murphy, Andrew A. Wylie, Simon J. Rhodes, John W. Keele, Kreg A. Leymaster, Randy L. Jirtle, and Timothy P.L. Smith. 2002. "Identification of the Single Base Change Causing the Callipyge Muscle Hypertrophy Phenotype, the Only Known Example of Polar Overdominance in Mammals." *Genome Research* 12 (10): 1496–1506. https://doi.org/10.1101/GR.571002.

Fu, Haiqing, Melvenia M. Martin, Marie Regairaz, Liang Huang, Yang You, Chi-Mei Lin, Michael Ryan, et al. 2015. "The DNA Repair Endonuclease Mus81 Facilitates Fast DNA Replication in the Absence of Exogenous Damage." *Nature Communications* 6. https://doi.org/10.1038/ncomms7746.

Fukami, Kiyoko, Kazuki Nakao, Takafumi Inoue, Yuki Kataoka, Manabu Kurokawa, Rafael A. Fissore, Kenji Nakamura, et al. 2001. “Requirement of Phospholipase Cδ4 for the Zona Pellucida-Induced Acrosome Reaction.” *Science* 292 (5518): 920–23. https://doi.org/10.1126/SCIENCE.1059042.

Gao, Ziyue, Darrel Waggoner, Matthew Stephens, Carole Ober, and Molly Przeworski. 2015. “An Estimate of the Average Number of Recessive Lethal Mutations Carried by Humans.” *Genetics* 199 (4): 1243–54. https://doi.org/10.1534/GENETICS.114.173351.

Garrick, Dorian J, Jeremy F Taylor, and Rohan L Fernando. 2009. “Deregressing Estimated Breeding Values and Weighting Information for Genomic Regression Analyses.” *Genetics Selection Evolution* 41. https://doi.org/10.1186/1297-9686-41-55.

Gasparini, P, G Novelli, X Estivill, D Olivieri, A Savoia, A Ruzzo, V Nunes, G Borgo, M Antonelli, and R Williamson. 1990. “The Genotype of a New Linked DNA Marker, MP6d-9, Is Related to the Clinical Course of Cystic Fibrosis.” *Journal of Medical Genetics* 27 (1): 17–20. https://doi.org/10.1136/JMG.27.1.17.

Ge, Xuan, Yu Wang, Karen SL Lam, and Aimin Xu. 2012. “Metabolic Actions of FGF21: Molecular Mechanisms and Therapeutic Implications.” *Acta Pharmaceutica Sinica B* 2 (4): 350–57. https://doi.org/10.1016/J.APSB.2012.06.011.

Geman, Stuart, and Donald Geman. 1984. “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6 (6): 721–41. https://doi.org/10.1109/TPAMI.1984.4767596.

Georges, Michel, Carole Charlier, and Ben Hayes. 2018. “Harnessing Genomic Information for Livestock Improvement.” *Nature Reviews Genetics* 20 (3): 135–56. https://doi.org/10.1038/s41576-018-0082-2.

Georges, Michel, Roger Drinkwater, Tracey King, Anuradha Mishra, Stephen S Moore, Dahlia Nielsen, Leslie S Sargeant, et al. 1993. “Microsatellite Mapping of a Gene Affecting Horn Development in Bos taurus.” *Nature Genetics* 4:206-10. https://doi.org/10.1038/ng0693-206.

Geweke, John. 1991. “Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments.” *Bayesian Statistics*. Vol. 4. https://www.researchgate.net/publication/2352607.

Göring, Harald H.H., Joesph D. Terwilliger, and John Blangero. 2001. “Large Upward Bias in Estimation of Locus-Specific Effects from Genomewide Scans.” *The American Journal of Human Genetics* 69 (6): 1357–69. https://doi.org/10.1086/324471.

Graham, Brett H., Katrina G. Waymire, Barbara Cottrell, Ian A. Trounce, Grant R. MacGregor, and Douglas C. Wallace. 1997. “A Mouse Model for Mitochondrial Myopathy and Cardiomyopathy Resulting from a Deficiency in the Heart/Muscle Isoform of the Adenine Nucleotide Translocator.” *Nature Genetics* 16 (3): 226–34. https://doi.org/10.1038/ng0797-226.

Gratten, Jacob, Naomi R Wray, Matthew C Keller, Peter M Visscher, and Nat Neurosci Author. 2014. “Large-Scale Genomics Unveils the Genetic Architecture of Psychiatric Disorders NIH Public Access Author Manuscript.” *Nat Neurosci* 17 (6): 782–90. https://doi.org/10.1038/nn.3708.

Grisart, Bernard, Wouter Coppieters, Frédéric Farnir, Latifa Karim, Christine Ford, Paulette Berzi, Nadine Cambisano, et al. 2002. “Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition.” *Genome Research* 12 (2): 222–31. https://doi.org/10.1101/gr.224202.

Grobet, Luc, Luis José Royo Martin, Dominique Poncelet, Dimitri Pirottin, Benoit Brouwers, Juliette Riquet, Andreina Schoeberlein, et al. 1997. “A Deletion in the Bovine Myostatin Gene Causes the Double–Muscled Phenotype in Cattle.” *Nature Genetics* 17 (1): 71–74. https://doi.org/10.1038/ng0997-71.

Guindo-Martínez, Marta, Ramon Amela, Silvia Bonàs-Guarch, Montserrat Puiggròs, Cecilia Salvoro, Irene Miguel-Escalada, Caitlin E. Carey, et al. 2021. “The Impact of Non-Additive Genetic Associations on Age-Related Complex Diseases.” *Nature Communications* 12 (1): 1–14. https://doi.org/10.1038/s41467-021-21952-4.

Habier, David, Rohan L. Fernando, Kadir Kizilkaya, and Dorian J. Garrick. 2011. “Extension of the Bayesian Alphabet for Genomic Selection.” *BMC Bioinformatics* 12 (1): 186. https://doi.org/10.1186/1471-2105-12-186.

Hakem, Razqallah. 2008. “DNA-Damage Repair; the Good, the Bad, and the Ugly.” *The EMBO Journal* 27 (4): 589–605. https://doi.org/10.1038/EMBOJ.2008.15.

Hanada, Katsuhiro, Magda Budzowska, Sally L Davies, Ellen van Drunen, Hideo Onizawa, H Berna Beverloo, Alex Maas, Jeroen Essers, Ian D Hickson, and Roland Kanaar. 2007. “The Structure-Specific Endonuclease Mus81 Contributes to Replication Restart by Generating Double-Strand DNA Breaks.” *Nature Structural & Molecular Biology* 14 (11): 1096–1104. https://doi.org/10.1038/nsmb1313.

Harris, BL, JM Clark, and RG Jackson. 1996. “Across Breed Evaluation of Dairy Cattle.” *New Zealand Society of Animal Production*. www.nzsap.org.nz.

Hastings, W. K. 1970. “Monte Carlo Sampling Methods Using Markov Chains and Their Applications.” *Biometrika* 57 (1): 97–109. https://doi.org/10.1093/BIOMET/57.1.97.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009. “Invited Review: Genomic Selection in Dairy Cattle: Progress and Challenges.” *Journal of Dairy Science* 92 (2): 433–43. https://doi.org/10.3168/JDS.2008-1646.

Hazel, L. N. 1943. “The Genetic Basis for Constructing Selection Indexes.” *Genetics* 28 (6).

Henderson, C. R. 1953. “Estimation of Variance and Covariance Components.” *Biometrics* 9 (2): 226. https://doi.org/10.2307/3001853.

Henderson, C. R. 1976. “A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values.” *Biometrics* 32 (1): 69. https://doi.org/10.2307/2529339.

Henderson, C R. 1985. “Equivalent Linear Models to Reduce Computations.” *Journal of Dairy Science* 68: 2267–77. https://doi.org/10.3168/jds.S0022-0302(85)81099-7.

Hill, William G., Michael E. Goddard, and Peter M. Visscher. 2008. “Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits.” PLOS Genetics 4 (2): e1000008. https://doi.org/10.1371/journal.pgen.1000008.

Hiltpold, Maya, Naveen Kumar Kadri, Fredi Janett, Ulrich Witschi, Fritz Schmitz-Hsu, and Hubert Pausch. 2021. “Autosomal Recessive Loci Contribute Significantly to Quantitative Variation of Male Fertility in a Dairy Cattle Population.” *BMC Genomics* 22. https://doi.org/10.1186/S12864-021-07523-3.

Horn, Michael, Reto Baumann, Jorge A. Pereira, Páris N. M. Sidiropoulos, Christian Somandin, Hans Welzl, Claudia Stendel, et al. 2012. “Myelin Is Dependent on the Charcot–Marie–Tooth Type 4H Disease Culprit Protein FRABIN/FGD4 in Schwann Cells.” *Brain* 135 (12): 3567–83. https://doi.org/10.1093/BRAIN/AWS275.

Hotelling, H. 1933. “Analysis of a Complex of Statistical Variables into Principal Components.” *Journal of Educational Psychology* 24 (6): 417–41. https://doi.org/10.1037/H0071325.

Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis. 2012. “Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-Phasing.” *Nature Genetics* 44 (8): 955–59. https://doi.org/10.1038/ng.2354.

Hu, Yaodong, Guilherme J M Rosa, and Daniel Gianola. 2015. “A GWAS Assessment of the Contribution of Genomic Imprinting to the Variation of Body Mass Index in Mice.” *BMC Genomics* 113. https://doi.org/10.1186/s12864-015-1721-z.

Huang, Wen, and Trudy F. C. Mackay. 2016. “The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis.” *PLOS Genetics* 12 (11): e1006421. https://doi.org/10.1371/JOURNAL.PGEN.1006421.

Huang, Wen, Stephen Richards, Mary Anna Carbone, Dianhui Zhu, Robert R H Anholt, Julien F Ayroles, Laura Duncan, et al. 2012. “Epistasis Dominates the Genetic Architecture of Drosophila Quantitative Traits.” *Proceedings of the National Academy of Sciences of the United States of America* 109 (39): 15553–59. https://doi.org/10.1073/pnas.1213423109.

Janssen, Erin, Erdyni Tsitsikov, Waleed Al-Herz, Gerard Lefranc, Andre Megarbane, Majed Dasouki, Francisco A. Bonilla, Talal Chatila, Lynda Schneider, and Raif S. Geha. 2014. “Flow Cytometry

Biomarkers Distinguish DOCK8 Deficiency from Severe Atopic Dermatitis." *Clinical Immunology* 150 (2): 220–24. https://doi.org/10.1016/j.clim.2013.12.006.

Jenko, Janez, Matthew C. McClure, Daragh Matthews, Jennifer McClure, Martin Johnsson, Gregor Gorjanc, and John M. Hickey. 2019. "Analysis of a Large Dataset Reveals Haplotypes Carrying Putatively Recessive Lethal and Semi-Lethal Alleles with Pleiotropic Effects on Economically Important Traits in Beef Cattle." *Genetics Selection Evolution* 51. https://doi.org/10.1186/S12711-019-0452-Z.

Jiang, Jicai, Li Ma, Dzianis Prakapenka, Paul M. VanRaden, John B. Cole, and Yang Da. 2019. "A Large-Scale Genome-Wide Association Study in U.S. Holstein Cattle." *Frontiers in Genetics* 10. https://doi.org/10.3389/fgene.2019.00412.

Jiang, Jicai, Botong Shen, Jeffrey R. O'Connell, Paul M. VanRaden, John B. Cole, and Li Ma. 2017. "Dissection of Additive, Dominance, and Imprinting Effects for Production and Reproduction Traits in Holstein Cattle." *BMC Genomics* 18. https://doi.org/10.1186/s12864-017-3821-4.

Jivanji, Swati, Gemma Worth, Thomas J. Lopdell, Anna Yeates, Christine Couldrey, Edwardo Reynolds, Kathryn Tiplady, et al. 2019. "Genome-Wide Association Analysis Reveals QTL and Candidate Mutations Involved in White Spotting in Cattle." *Genetics Selection Evolution* 51. https://doi.org/10.1186/s12711-019-0506-2.

Kadri, Naveen Kumar, Chad Harland, Pierre Faux, Nadine Cambisano, Latifa Karim, Wouter Coppieters, Sébastien Fritz, et al. 2016. "Coding and Noncoding Variants in HFM1, MLH3, MSH4, MSH5, RNF212, and RNF212B Affect Recombination Rate in Cattle." *Genome Research* 26 (10): 1323. http://www.genome.org/cgi/doi/10.1101/gr.204214.116.

Kang, Hyun Min, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin. 2010. "Variance Component Model to Account for Sample Structure in Genome-Wide Association Studies." *Nature Genetics* 42 (4): 348–54. https://doi.org/10.1038/ng.548.

Kang, Hyun Min, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. 2008. "Efficient Control of Population Structure in Model Organism Association Mapping." *Genetics* 178 (3): 1709–23. https://doi.org/10.1534/genetics.107.080101.

Karim, Latifa, Haruko Takeda, Li Lin, Tom Druet, Juan a C Arias, Denis Baurain, Nadine Cambisano, et al. 2011. "Variants Modulating the Expression of a Chromosome Domain Encompassing PLAG1 Influence Bovine Stature." *Nature Genetics* 43 (5): 405–13. https://doi.org/10.1038/ng.814.

Kathiresan, Sekar, Olle Melander, Candace Guiducci, Aarti Surti, Noël P Burtt, Mark J Rieder, Gregory M Cooper, et al. 2008. "Six New Loci Associated with Blood Low-Density Lipoprotein Cholesterol, High-Density Lipoprotein Cholesterol or Triglycerides in Humans." *Nature Genetics* 40 (2): 189–97. https://doi.org/10.1038/NG.75.

Kaukonen, Jyrki, Jukka K. Juselius, Valeria Tiranti, Aija Kyttälä, Massimo Zeviani, Giacomo P. Comi, Sirkka Keränen, Leena Peltonen, and Anu Suomalainen. 2000. "Role of Adenine Nucleotide Translocator 1 in MtDNA Maintenance." *Science.* 289 (5480): 782–85. https://www.science.org/doi/abs/10.1126/science.289.5480.782.

Kearney, Conor J., Katrina L. Randall, and Jane Oliaro. 2017. "DOCK8 Regulates Signal Transduction Events to Control Immunity." *Cellular and Molecular Immunology* 14 (5): 406–11. https://doi.org/10.1038/cmi.2017.9.

Khetarpal, Sumeet A., Katrine T. Schjoldager, Christina Christoffersen, Avanthi Raghavan, Andrew C. Edmondson, Heiko M. Reutter, Bouhouche Ahmed, et al. 2016. "Loss of Function of GALNT2 Lowers High-Density Lipoproteins in Humans, Nonhuman Primates, and Rodents." *Cell Metabolism* 24 (2): 234–45. https://doi.org/10.1016/J.CMET.2016.07.012.

Kirino, Yohei, George Bertsias, Yoshiaki Ishigatsubo, Nobuhisa Mizuki, Ilknur Tugal-Tutkun, Emire Seyahi, Yilmaz Ozyazgan, et al. 2013. "Genome-Wide Association Analysis Identifies New Susceptibility Loci for Behçet's Disease and Epistasis between HLA-B*51 and ERAP1." *Nature Genetics* 45 (2): 202–7. https://doi.org/10.1038/ng.2520.

Klungland, H., D. I. Vage, L. Gomez-Raya, S. Adalsteinsson, and S. Lien. 1995. "The Role of Melanocyte-Stimulating Hormone (MSH) Receptor in Bovine Coat Color Determination." *Mammalian Genome* 6 (9): 636–39. https://doi.org/10.1007/BF00352371.

Knapp, Karen M, Gemma Poke, Danielle Jenkins, Werner Truter, and Louise S. Bicknell. 2019. "Expanding the Phenotypic Spectrum Associated with DPF2: A New Case Report." *American Journal of Medical Genetics Part A* 179 (8): 1637–41. https://doi.org/10.1002/AJMG.A.61262.

Kosho, Tomoki, Noriko Miyake, and John C. Carey. 2014. "Coffin–Siris Syndrome and Related Disorders Involving Components of the BAF (MSWI/SNF) Complex: Historical Review and Recent Advances Using next Generation Sequencing." *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* 166 (3): 241–51. https://doi.org/10.1002/AJMG.C.31415.

Lambe, Teresa, Greg Crawford, Andy L. Johnson, Tanya L. Crockford, Tiphaine Bouriez-Jones, Aisling M. Smyth, Trung H.M. Pham, et al. 2011. "DOCK8 Is Essential for T-Cell Survival and the Maintenance of CD8 + T-Cell Memory." *European Journal of Immunology* 41 (12): 3423–35. https://doi.org/10.1002/eji.201141759.

Lander, ES, and NJ Schork. 1994. "Genetic Dissection of Complex Traits." *Science* 265 (5181): 2037–48. https://doi.org/10.1126/SCIENCE.8091226.

Larson, Greger, and Dorian Q. Fuller. 2014. "The Evolution of Animal Domestication." *Annual Review of Ecology, Evolution and Systematics* 45:115–36. https://doi.org/10.1146/annurev-ecolsys-110512-135813.

Lehnert, Klaus, Hamish Ward, Sarah D. Berry, Alex Ankersmit-Udy, Alayna Burrett, Elizabeth M. Beattie, Natalie L. Thomas, et al. 2015. "Phenotypic Population Screen Identifies a New

Mutation in Bovine DGAT1 Responsible for Unsaturated Milk Fat." *Scientific Reports* 5. https://doi.org/10.1038/srep08484.

Lek, Monkol, Konrad J. Karczewski, Eric v. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. "Analysis of Protein-Coding Genetic Variation in 60,706 Humans." *Nature* 536 (7616): 285–91. https://doi.org/10.1038/nature19057.

Lenffer, Johann, Frank W. Nicholas, Kao Castle, Arjun Rao, Stefan Gregory, Michael Poidinger, Matthew D. Mailman, and Shoba Ranganathan. 2006. "OMIA (Online Mendelian Inheritance in Animals): An Enhanced Platform and Integration into the Entrez Search Interface at NCBI." *Nucleic Acids Research* 34. https://doi.org/10.1093/NAR/GKJ152.

Li, Heng. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM." *ArXiv*, March. http://arxiv.org/abs/1303.3997.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. "FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/BIOINFORMATICS/BTT656.

LIC. 2020. "Investment in Genetic Research and Technology to Further Improve Cow Production." 2020. https://www.lic.co.nz/news/investment-genetic-research-and-technology-further-improve-cow-production/.

LIC. 2021. "New Zealand Scientists Receive Global Recognition for Cow Genetic Discoveries." 2021. https://www.lic.co.nz/news/new-zealand-scientists-receive-global-recognition-cow-genetic-discoveries/.

Lippert, Christoph, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. 2011. "FaST Linear Mixed Models for Genome-Wide Association Studies." *Nature Methods* 8 (10): 833–37. https://doi.org/10.1038/nmeth.1681.

Listgarten, Jennifer, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. 2012. "Improved Linear Mixed Models for Genome-Wide Association Studies." *Nature Methods* 9 (6): 525–26. https://doi.org/10.1038/nmeth.2037.

Littlejohn, Mathew D., Kristen M. Henty, Kathryn Tiplady, Thomas Johnson, Chad Harland, Thomas Lopdell, Richard G. Sherlock, et al. 2014. "Functionally Reciprocal Mutations of the Prolactin Signalling Pathway Define Hairy and Slick Cattle." *Nature Communications* 5. https://doi.org/10.1038/ncomms6861.

Littlejohn, Mathew D., Kathryn Tiplady, Tania A. Fink, Klaus Lehnert, Thomas Lopdell, Thomas Johnson, Christine Couldrey, et al. 2016. "Sequence-Based Association Analysis Reveals an MGST1 EQTL with Pleiotropic Effects on Bovine Milk Composition." *Scientific Reports* 6. https://doi.org/10.1038/srep25376.

Littlejohn, Mathew D., Kathryn Tiplady, Thomas Lopdell, Tania A. Law, Andrew Scott, Chad Harland, Ric Sherlock, et al. 2014. “Expression Variants of the Lipogenic AGPAT6 Gene Affect Diverse Milk Composition Phenotypes in Bos Taurus.” *PLoS ONE* 9 (1): 85757. https://doi.org/10.1371/journal.pone.0085757.

Livestock Improvement Corporation. 2020. *Dairy Statistics 2019-2020*. Hamilton, NZ: Livestock Improvement Corporation.

Loh, Po-Ru, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, et al. 2015. “Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts.” *Nature Genetics* 47 (3): 284–90. https://doi.org/10.1038/ng.3190.

Lohmueller, Kirk E., Celeste L. Pearce, Malcolm Pike, Eric S. Lander, and Joel N. Hirschhorn. 2003. “Meta-Analysis of Genetic Association Studies Supports a Contribution of Common Variants to Susceptibility to Common Disease.” *Nature Genetics* 33 (2): 177–82. https://doi.org/10.1038/NG1071.

Lopdell, Thomas J., Kathryn Tiplady, Christine Couldrey, Thomas J.J. Johnson, Michael Keehan, Stephen R. Davis, Bevin L. Harris, Richard J. Spelman, Russell G. Snell, and Mathew D. Littlejohn. 2019. “Multiple QTL Underlie Milk Phenotypes at the CSF2RB Locus.” *Genetics Selection Evolution* 51. https://doi.org/10.1186/s12711-019-0446-x.

Lopdell, Thomas J., Kathryn Tiplady, Maksim Struchalin, Thomas J. J. Johnson, Michael Keehan, Ric Sherlock, Christine Couldrey, et al. 2017. “DNA and RNA-Sequence Based GWAS Highlights Membrane-Transport Genes as Key Modulators of Milk Lactose Content.” *BMC Genomics* 18. https://doi.org/10.1186/s12864-017-4320-3.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15. https://doi.org/10.1186/S13059-014-0550-8.

Lucy, M. C., G. A. Verkerk, B. E. Whyte, K. A. Macdonald, L. Burton, R. T. Cursons, J. R. Roche, and C. W. Holmes. 2009. “Somatotropic Axis Components and Nutrient Partitioning in Genetically Diverse Dairy Cows Managed under Different Feed Allowances in a Pasture System.” *Journal of Dairy Science* 92 (2): 526–39. https://doi.org/10.3168/JDS.2008-1421.

Lush, Jay Laurence. 1940. “Intra-Sire Correlations or Regressions of  Offspring on Dam as a Method of Estimating  Heritability of Characteristics.” *Proceedings of the American Society of Animal Production* 33: 293–301. https://doi.org/10.2527/jas1940.19401293x.

Lynch, Michael, and Bruce Walsh. 1998. *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates.

Ma, Li, Jeffrey R. O’Connell, Paul M. VanRaden, Botong Shen, Abinash Padhi, Chuanyu Sun, Derek M. Bickhart, et al. 2015. “Cattle Sex-Specific Recombination and Genetic Control from a Large

Pedigree Analysis." *PLOS Genetics* 11 (11): e1005387. https://doi.org/10.1371/JOURNAL.PGEN.1005387.

Mackay, Trudy F. C. 2001. "The Genetic Architecture of Quantitative Traits." *Annual Review of Genetics* 35: 303–39. https://doi.org/10.1146/ANNUREV.GENET.35.102401.090633.

Mackay, Trudy F. C. 2013. "Epistasis and Quantitative Traits: Using Model Organisms to Study Gene–Gene Interactions." *Nature Reviews Genetics* 15 (1): 22–33. https://doi.org/10.1038/nrg3627.

MacKinnon, Michael. 2010. "Cattle 'Breed' Variation and Improvement in Roman Italy: Connecting the Zooarchaeological and Ancient Textual Evidence." *World Archaeology* 42 (1): 55–73. https://doi.org/10.1080/00438240903429730.

Markan, Kathleen R., Meghan C. Naber, Magdalene K. Ameka, Maxwell D. Anderegg, David J. Mangelsdorf, Steven A. Kliewer, Moosa Mohammadi, and Matthew J. Potthoff. 2014. "Circulating FGF21 Is Liver Derived and Enhances Glucose Uptake During Refeeding and Overfeeding." *Diabetes* 63 (12): 4057–63. https://doi.org/10.2337/DB14-0595.

Marouli, Eirini, Mariaelisa Graff, Carolina Medina-Gomez, Ken Sin Lo, Andrew R. Wood, Troels R. Kjaer, Rebecca S. Fine, et al. 2017. "Rare and Low-Frequency Coding Variants Alter Human Adult Height" *Nature* 542: 186–90. https://doi.org/10.1038/nature21039.

Matukumalli, Lakshmi K., Cynthia T. Lawley, Robert D. Schnabel, Jeremy F. Taylor, Mark F. Allan, Michael P. Heaton, Jeff O'Connell, et al. 2009. "Development and Characterization of a High Density SNP Genotyping Assay for Cattle." *PLOS ONE* 4 (4): e5350. https://doi.org/10.1371/JOURNAL.PONE.0005350.

McKusick, Victor A. 2007. "Mendelian Inheritance in Man and Its Online Version, OMIM." *American Journal of Human Genetics* 80 (4): 588. https://doi.org/10.1086/514346.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R.S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-0974-4.

McPherson, John Peter, Bénédicte Lemmers, Richard Chahwan, Ashwin Pamidi, Eva Migon, Elzbieta Matysiak-Zablocki, Mary Ellen Moynahan, et al. 2004. "Involvement of Mammalian Mus81 in Genome Integrity and Tumor Suppression." *Science* 304 (5678): 1822–26. https://doi.org/10.1126/SCIENCE.1094557.

Mendel, Gregor. 1865. "Experiments in Plant Hybridization". Translated by William Bateson. 1901. *Journal of the Royal Horticultural Society* 26:1-32.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087-92. https://www.doi.org/10.1063/1.1699114.

Metzker, Michael L. 2010. "Sequencing Technologies — the next Generation." *Nature Reviews Genetics* 11 (1): 31–46. https://doi.org/10.1038/nrg2626.

Meuwissen, T. H E, B. J. Hayes, and M. E. Goddard. 2001. "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." *Genetics* 157 (4): 1819–29. https://doi.org/11290733.

Michot, Pauline, Sabine Chahory, Andrew Marete, Cécile Grohs, Dimitri Dagios, Elise Donzel, Abdelhak Aboukadiri, et al. 2016. "A Reverse Genetic Approach Identifies an Ancestral Frameshift Mutation in RP1 Causing Recessive Progressive Retinal Degeneration in European Cattle Breeds." *Genetics Selection Evolution* 48 (1). https://doi.org/10.1186/s12711-016-0232-y.

Mishra, Nivedita Awasthi, Cord Drögemüller, Vidhya Jagannathan, Irene Keller, Daniel Wüthrich, Rémy Bruggmann, Julia Beck, et al. 2017. "A Structural Variant in the 5'-Flanking Region of the TWIST2 Gene Affects Melanocyte Development in Belted Cattle." *PLOS ONE* 12 (6): e0180170. https://doi.org/10.1371/journal.pone.0180170.

Moser, Gerhard, Sang Hong Lee, Ben J. Hayes, Michael E. Goddard, Naomi R. Wray, and Peter M. Visscher. 2015. "Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model." *PLOS Genetics* 11 (4): e1004969. https://doi.org/10.1371/journal.pgen.1004969.

Ng, Pauline C., and Steven Henikoff. 2003. "SIFT: Predicting Amino Acid Changes That Affect Protein Function." *Nucleic Acids Research* 31 (13): 3812–14. https://doi.org/10.1093/nar/gkg509.

North, B. v., D. Curtis, and P. C. Sham. 2002. "A Note on the Calculation of Empirical P Values from Monte Carlo Procedures." *American Journal of Human Genetics* 71 (2): 439. https://doi.org/10.1086/341527.

Ono, Yuichi, Hiroyuki Nakanishi, Miyuki Nishimura, Mayumi Kakizaki, Kenichi Takahashi, Masako Miyahara, Keiko Satoh-Horikawa, Kenji Mandai, and Yoshimi Takai. 2000. "Two Actions of Frabin: Direct Activation of Cdc42 and Indirect Activation of Rac." *Oncogene* 19 (27): 3050–58. https://doi.org/10.1038/sj.onc.1203631.

Palmieri, Luigi, Simona Alberio, Isabella Pisano, Tiziana Lodi, Mija Meznaric-Petrusa, Janez Zidar, Antonella Santoro, et al. 2005. "Complete Loss-of-Function of the Heart/Muscle-Specific Adenine Nucleotide Translocator Is Associated with Mitochondrial Myopathy and Cardiomyopathy." *Human Molecular Genetics* 14 (20): 3079–88. https://doi.org/10.1093/hmg/ddi341.

Pausch, Hubert, Reiner Emmerling, Hermann Schwarzenbacher, and Ruedi Fries. 2016. "A Multi-Trait Meta-Analysis with Imputed Sequence Variants Reveals Twelve QTL for Mammary Gland Morphology in Fleckvieh Cattle." *Genetics Selection Evolution* 48. https://doi.org/10.1186/S12711-016-0190-4.

Pausch, Hubert, Iona M. MacLeod, Ruedi Fries, Reiner Emmerling, Phil J. Bowman, Hans D. Daetwyler, and Michael E. Goddard. 2017. "Evaluation of the Accuracy of Imputed Sequence Variant Genotypes and Their Utility for Causal Variant Detection in Cattle." *Genetics Selection Evolution* 49. https://doi.org/10.1186/S12711-017-0301-X.

Pérez, Paulino, and Gustavo de Los Campos. 2014. "Genome-Wide Regression and Prediction with the BGLR Statistical Package." *Genetics* 198: 483. https://doi.org/10.1534/genetics.114.164442.

Phillips, Patrick C. 2008. "Epistasis — the Essential Role of Gene Interactions in the Structure and Evolution of Genetic Systems." *Nature Reviews Genetics* 9 (11): 855–67. https://doi.org/10.1038/nrg2452.

Powell, Joseph E., Anjali K. Henders, Allan F. McRae, Jinhee Kim, Gibran Hemani, Nicholas G. Martin, Emmanouil T. Dermitzakis, Greg Gibson, Grant W. Montgomery, and Peter M. Visscher. 2013. "Congruence of Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP Data." *PLOS Genetics* 9 (5): e1003502. https://doi.org/10.1371/JOURNAL.PGEN.1003502.

Pryce, Jennie E, Mekonnen Haile-Mariam, Michael E Goddard, and Ben J Hayes. 2014. "Identification of Genomic Regions Associated with Inbreeding Depression in Holstein and Jersey Dairy Cattle." *Genetics Selection Evolution* 46. https://doi.org/10.1186/S12711-014-0071-7.

Randall, Katrina L, Teresa Lambe, Andy L Johnson, Bebhinn Treanor, Edyta Kucharska, Heather Domaschenz, Belinda Whittle, et al. 2009. "Dock8 Mutations Cripple B Cell Immunological Synapses, Germinal Centers and Long-Lived Antibody Production." *Nature Immunology* 10: 1283–1291. https://doi.org/10.1038/ni.1820.

Raven, Lesley Ann, Benjamin G. Cocks, Kathryn E. Kemper, Amanda J. Chamberlain, Christy J. vander Jagt, Michael E. Goddard, and Ben J. Hayes. 2016. "Targeted Imputation of Sequence Variants and Gene Expression Profiling Identifies Twelve Candidate Genes Associated with Lactation Volume, Composition and Calving Interval in Dairy Cattle." *Mammalian Genome* 27 (1–2): 81–97. https://doi.org/10.1007/s00335-015-9613-8.

Reed, Danielle R., Maureen P. Lawler, and Michael G. Tordoff. 2008. "Reduced Body Weight Is a Common Effect of Gene Knockout in Mice." *BMC Genetics* 9. https://doi.org/10.1186/1471-2156-9-4.

Reynolds, Edwardo G. M., Catherine Neeley, Thomas J. Lopdell, Michael Keehan, Keren Dittmer, Chad S. Harland, Christine Couldrey, et al. 2021. "Non-Additive Association Analysis Using Proxy Phenotypes Identifies Novel Cattle Syndromes." *Nature Genetics* 53 (7): 949–54. https://doi.org/10.1038/s41588-021-00872-5.

Reynolds, Edwardo G.M., Thomas Lopdell, Yu Wang, Kathryn M. Tiplady, Chad S. Harland, Thomas J. J. Johnson, Catherine Neeley, et al. 2022. "Non-Additive QTL Mapping of Lactation Traits in

124,000 Cattle Reveals Novel Recessive Loci". *Genetics Selection Evolution* 54. https://doi.org/10.1186/s12711-021-00694-3

Rheenen, Wouter van, Aleksey Shatunov, Annelot M Dekker, Russell L McLaughlin, Frank P Diekstra, Sara L Pulit, Rick A A van der Spek, et al. 2016. "Genome-Wide Association Analyses Identify New Risk Variants and the Genetic Architecture of Amyotrophic Lateral Sclerosis." *Nature Genetics* 48 (9): 1043–48. https://doi.org/10.1038/ng.3622.

Risch, Neil, and Kathleen Merikangas. 1996. "The Future of Genetic Studies of Complex Human Diseases." *Science* 273 (5281): 1516–17. https://doi.org/10.1126/SCIENCE.273.5281.1516.

Robinson, James T, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz, and Jill P Mesirov. 2011. "Integrative Genomics Viewer." *Nature Biotechnology* 29 (1): 24–26. https://doi.org/10.1038/nbt.1754.

Roos, A P W de, B J Hayes, R J Spelman, and M E Goddard. 2008. "Linkage Disequilibrium and Persistence of Phase in Holstein–Friesian, Jersey and Angus Cattle." *Genetics* 179 (3): 1503–12. https://doi.org/10.1534/GENETICS.107.084301.

Rosen, Benjamin D, Derek M Bickhart, Robert D Schnabel, Sergey Koren, Christine G Elsik, Elizabeth Tseng, Troy N Rowan, et al. 2020. "De Novo Assembly of the Cattle Reference Genome with Single-Molecule Sequencing." *GigaScience* 9 (3): 1–9. https://doi.org/10.1093/GIGASCIENCE/GIAA021.

Rudnik-Schöneborn, Sabine, Dorothee Röhrig, Garth Nicholson, and Klaus Zerres. 1993. "Pregnancy and Delivery in Charcot-Marie-Tooth Disease Type 1." *Neurology* 43 (10): 2011–2011. https://doi.org/10.1212/WNL.43.10.2011.

Sanders, Anna A.W.M., Erik de Vrieze, Anas M. Alazami, Fatema Alzahrani, Erik B. Malarkey, Nasrin Sorusch, Lars Tebbe, et al. 2015. "KIAA0556 Is a Novel Ciliary Basal Body Component Mutated in Joubert Syndrome." *Genome Biology* 16. https://doi.org/10.1186/s13059-015-0858-z.

Sandre-Giovannoli, A de, V Delague, T Hamadouche, M Chaouch, M Krahn, I Boccaccio, T Maisonobe, et al. 2005. "Homozygosity Mapping of Autosomal Recessive Demyelinating Charcot-Marie-Tooth Neuropathy (CMT4H) to a Novel Locus on Chromosome 12p11.21-Q13.11." *Journal of Medical Genetics* 42 (3): 260–65. https://doi.org/10.1136/JMG.2004.024364.

Sargolzaei, Mehdi, and Flavio S. Schenkel. 2009. "QMSim: A Large-Scale Genome Simulator for Livestock." *Bioinformatics* 25 (5): 680–81. https://doi.org/10.1093/bioinformatics/btp045.

Schlein, Christian, Saswata Talukdar, Markus Heine, Alexander W. Fischer, Lucia M. Krott, Stefan K. Nilsson, Martin B. Brenner, Joerg Heeren, and Ludger Scheja. 2016. "FGF21 Lowers Plasma Triglycerides by Accelerating Lipoprotein Catabolism in White and Brown Adipose Tissues." *Cell Metabolism* 23 (3): 441–53. https://doi.org/10.1016/J.CMET.2016.01.006.

Schutte, JE, JC Longhurst, FA Gaffney, BC Bastian, and CG Blomqvist. 1981. "Total Plasma Creatinine: An Accurate Measure of Total Striated Muscle Mass." *Journal of Applied Physiology:*

*Respiratory, Environmental and Exercise Physiology* 51 (3): 762–66. https://doi.org/10.1152/JAPPL.1981.51.3.762.

Searle, Shayle R, George Casella, and Charles E McCulloch. 2009. *Variance Components*. Vol. 391. John Wiley & Sons.

Shen, Botong, Jicai Jiang, Eyal Seroussi, George E. Liu, and Li Ma. 2018. "Characterization of Recombination Features and the Genetic Basis in Multiple Cattle Breeds." *BMC Genomics* 19. https://doi.org/10.1186/S12864-018-4705-Y.

Shirakawa, Taro, Klaus A. Deichmann, Kenji Izuhara, Xiao Quan Mao, Chaker N. Adra, and Julian M. Hopkin. 2000. "Atopy and Asthma: Genetic Variants of IL-4 and IL-13 Signalling." *Immunology Today*. https://doi.org/10.1016/S0167-5699(99)01492-9.

Stapley, Jessica, Philine G. D. Feulner, Susan E. Johnston, Anna W. Santure, and Carole M. Smadja. 2017. "Variation in Recombination Frequency and Distribution across Eukaryotes: Patterns and Processes." *Philosophical Transactions of the Royal Society B: Biological Sciences* 372 (1736). https://doi.org/10.1098/RSTB.2016.0455.

Steinthorsdottir, Valgerdur, Gudmar Thorleifsson, Patrick Sulem, Hannes Helgason, Niels Grarup, Asgeir Sigurdsson, Hafdis T Helgadottir, et al. 2014. "Identification of Low-Frequency and Rare Sequence Variants Associated with Elevated or Reduced Risk of Type 2 Diabetes." *Nature Genetics* 46 (3): 294–98. https://doi.org/10.1038/ng.2882.

Stendel, Claudia, Andreas Roos, Tine Deconinck, Jorge Pereira, François Castagner, Axel Niemann, Janbernd Kirschner, et al. 2007. "Peripheral Nerve Demyelination Caused by a Mutant Rho GTPase Guanine Nucleotide Exchange Factor, Frabin/FGD4." *American Journal of Human Genetics* 81 (1): 158. https://doi.org/10.1086/518770.

Stephens, Matthew, and David J. Balding. 2009. "Bayesian Statistical Methods for Genetic Association Studies." *Nature Reviews Genetics* 10 (10): 681–90. https://doi.org/10.1038/nrg2615.

Storey, John D, Andrew J Bass, Alan Dabney, and David Robinson. 2020. "Qvalue: Q-Value Estimation for False Discovery Rate Control." http://github.com/jdstorey/qvalue.

Storey, John D., and Robert Tibshirani. 2003. "Statistical Significance for Genomewide Studies." *Proceedings of the National Academy of Sciences of the United States of America* 100 (16): 9440–45. https://doi.org/10.1073/pnas.1530509100.

Strange, Amy, Francesca Capon, Chris CA Spencer, Jo Knight, Michael E Weale, Michael H Allen, Anne Barton, et al. 2010. "Genome-Wide Association Study Identifies New Psoriasis Susceptibility Loci and an Interaction between HLA-C and ERAP1." *Nature Genetics* 42 (11): 985. https://doi.org/10.1038/NG.694.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, et al. 2015. "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age." *PLOS Medicine* 12 (3): e1001779. https://doi.org/10.1371/JOURNAL.PMED.1001779.

Sun, Chuanyu, Paul M. VanRaden, John B. Cole, and Jeffrey R. O'Connell. 2014. "Improvement of Prediction Ability for Genomic Selection of Dairy Cattle by Including Dominance Effects." *PLoS ONE* 9 (8): e103934. https://doi.org/10.1371/journal.pone.0103934.

Szigeti, Kinga, and James R Lupski. 2009. "Charcot–Marie–Tooth Disease." *European Journal of Human Genetics* 17 (6): 703–10. https://doi.org/10.1038/ejhg.2009.31.

Tacon, Terry. 2002. "Young Bull Sets Record for Breeding." *Taranaki Daily News*, 18.

The 1000 Genomes Project Consortium, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

The Bovine Genome Sequencing and Analysis Consortium, Christine G. Elsik, Ross L. Tellam, and Kim C. Worley. 2009. "The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution." *Science* 324 (5926): 522–28. https://doi.org/10.1126/SCIENCE.1169588.

Tiplady, Kathryn M., Thomas J. Lopdell, Edwardo Reynolds, Richard G. Sherlock, Michael Keehan, Thomas JJ. Johnson, Jennie E. Pryce, et al. 2021. "Sequence-Based Genome-Wide Association Study of Individual Milk Mid-Infrared Wavenumbers in Mixed-Breed Dairy Cattle." *Genetics Selection Evolution* 53. https://doi.org/10.1186/S12711-021-00648-9.

Toosi, Ali, Rohan L. Fernando, and Jack C. M. Dekkers. 2018. "Genome-Wide Mapping of Quantitative Trait Loci in Admixed Populations Using Mixed Linear Model and Bayesian Multiple Regression Analysis" *Genetics Selection Evolution* 50. https://doi.org/10.1186/s12711-018-0402-11.

Toro, Miguel A, and Luis Varona. 2010. "A Note on Mate Allocation for Dominance Handling in Genomic Selection" *Genetics Selection Evolution* 42. https://doi.org/10.1186/1297-9686-42-33.

Truswell, A S. 2005. "The A2 Milk Case: A Critical Review." *European Journal of Clinical Nutrition* 59 (5): 623–31. https://doi.org/10.1038/sj.ejcn.1602104.

Vanraden, P M. 2008. "Efficient Methods to Compute Genomic Predictions." *Journal of Dairy Science* 91 (11): 4414–23. https://doi.org/10.3168/jds.2007-0980.

VanRaden, P. M., and G. R. Wiggans. 1991. "Derivation, Calculation, and Use of National Animal Model Information." *Journal of Dairy Science* 74 (8): 2737–46. https://doi.org/10.3168/JDS.S0022-0302(91)78453-1.

VanRaden, P.M., K.M. Olson, D.J. Null, and J.L. Hutchison. 2011. “Harmful Recessive Effects on Fertility Detected by Absence of Homozygous Haplotypes.” *Journal of Dairy Science* 94 (12): 6153–61. https://doi.org/10.3168/JDS.2011-4624.

Varona, Luis, Andres Legarra, Miguel A. Toro, and Zulma G. Vitezica. 2018. “Non-Additive Effects in Genomic Selection.” *Frontiers in Genetics* 9. https://doi.org/10.3389/fgene.2018.00078.

Vasileiou, Georgia, Silvia Vergarajauregui, Sabine Endele, Bernt Popp, Christian Büttner, Arif B. Ekici, Marion Gerard, et al. 2018. “Mutations in the BAF-Complex Subunit DPF2 Are Associated with Coffin-Siris Syndrome.” *American Journal of Human Genetics* 102 (3): 468. https://doi.org/10.1016/J.AJHG.2018.01.014.

Véniant, Murielle M., Renee Komorowski, Ping Chen, Shanaka Stanislaus, Katherine Winters, Todd Hager, Lei Zhou, Russell Wada, Randy Hecht, and Jing Xu. 2012. “Long-Acting FGF21 Has Enhanced Efficacy in Diet-Induced Obese Mice and in Obese Rhesus Monkeys.” *Endocrinology* 153 (9): 4192–4203. https://doi.org/10.1210/EN.2012-1211.

Venter, J C, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, et al. 2001. “The Sequence of the Human Genome.” *Science* 291 (5507): 1304–51. https://doi.org/10.1126/science.1058040.

Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. “Heritability in the Genomics Era — Concepts and Misconceptions.” *Nature Reviews Genetics* 9 (4): 255–66. https://doi.org/10.1038/nrg2322.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. “10 Years of GWAS Discovery: Biology, Function, and Translation.” *The American Journal of Human Genetics* 101 (1): 5–22. http://dx.doi.org/10.1016/j.ajhg.2017.06.005. .

Vitezica, Zulma G., Luis Varona, and Andres Legarra. 2013. “On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope.” *Genetics* 195 (4): 1223–30. https://doi.org/10.1534/genetics.113.155176.

Wang, Kai, Xijian Hu, and Yingwei Peng. 2013. “An Analytical Comparison of the Principal Component Method and the Mixed Effects Model for Association Studies in the Presence of Cryptic Relatedness and Population Stratification.” *Human Heredity* 76. https://doi.org/10.1159/000353345.

Wang, Xue, Changyun Ping, Puwen Tan, Chenguang Sun, Guang Liu, Tao Liu, Shuchun Yang, et al. 2021. “HnRNPLL Controls Pluripotency Exit of Embryonic Stem Cells by Modulating Alternative Splicing of Tbx3 and Bptf.” *The EMBO Journal* 40 (4): e104729. https://doi.org/10.15252/embj.2020104729.

Wang, Yu, Kathryn Tiplady, Thomas J.J. Johnson, Chad Harland, Michael Keehan, Edwardo Reynolds, Ric G. Sherlock, et al. 2020. “Evaluating the Accuracy of Imputed Whole-Genome Sequence Data in Admixed Dairy Cattle.” In *International Conference of Quantitative Genetics 6*.

Watson, J. D., and F. H. Crick. 1953. “The Structure of DNA.” *Cold Spring Harbor Symposia on Quantitative Biology* 18: 123–31. https://doi.org/10.1101/SQB.1953.018.01.020.

Wei, Wen Hua, Gibran Hemani, and Chris S. Haley. 2014. “Detecting Epistasis in Human Complex Traits.” *Nature Reviews Genetics* 15 (11): 722–33. https://doi.org/10.1038/nrg3747.

Weller, J. I., Y. Kashi, and M. Soller. 1990. “Power of Daughter and Granddaughter Designs for Determining Linkage Between Marker Loci and Quantitative Trait Loci in Dairy Cattle.” *Journal of Dairy Science* 73 (9): 2525–37. https://doi.org/10.3168/JDS.S0022-0302(90)78938-2.

Weng, Zi-Qing, Mahdi Saatchi, Robert D Schnabel, Jeremy F Taylor, and Dorian J Garrick. 2014. “Recombination Locations and Rates in Beef Cattle Assessed from Parent-Offspring Pairs.” *Genetics Selection Evolution* 46. https://doi.org/10.1186/1297-9686-46-34.

Wiggans, George R., John B. Cole, Suzanne M. Hubbard, and Tad S. Sonstegard. 2017. “Genomic Selection in Dairy Cattle: The USDA Experience*.” *Annual Review of Animal Biosciences* 5 (1): 309–27. https://doi.org/10.1146/ANNUREV-ANIMAL-021815-111422.

Willer, Cristen J, Serena Sanna, Anne U Jackson, Angelo Scuteri, Lori L Bonnycastle, Robert Clarke, Simon C Heath, et al. 2008. “Newly Identified Loci That Influence Lipid Concentrations and Risk of Coronary Artery Disease.” *Nature Genetics* 40 (2): 161–69. https://doi.org/10.1038/NG.76.

Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, et al. 2014. “Defining the Role of Common Variation in the Genomic and Biological Architecture of Adult Human Height.” *Nature Genetics* 46 (11): 1173–86. https://doi.org/10.1038/ng.3097.

Wray, Naomi R. 2005. “Allele Frequencies and the R2 Measure of Linkage Disequilibrium: Impact on Design and Interpretation of Association Studies.” *Twin Research and Human Genetics* 8 (02): 87–94. https://doi.org/10.1375/twin.8.2.87.

Wu, Xiaoping, Ming Fang, Lin Liu, Sheng Wang, Jianfeng Liu, Xiangdong Ding, Shengli Zhang, et al. 2013. “Genome Wide Association Studies for Body Conformation Traits in the Chinese Holstein Cattle Population.” *BMC Genomics* 14. https://doi.org/10.1186/1471-2164-14-897.

Xiang, Ruidong, Irene van den Berg, Iona M. MacLeod, Benjamin J. Hayes, Claire P. Prowse-Wilkins, Min Wang, Sunduimijid Bolormaa, et al. 2019. “Quantifying the Contribution of Sequence Variants with Regulatory and Evolutionary Significance to 34 Bovine Complex Traits.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (39): 19398–408. https://doi.org/10.1073/pnas.1904159116.

Yang, Jian, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, et al. 2010. "Common SNPs Explain a Large Proportion of the Heritability for Human Height" *Nature Genetics* 42 (7): 565–69. https://doi.org/10.1038/ng.608.

Yang, Jian, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Pamela A F Madden, Andrew C Heath, Nicholas G Martin, et al. 2012. "Conditional and Joint Multiple-SNP Analysis of GWAS Summary Statistics Identifies Additional Variants Influencing Complex Traits." *Nature Genetics* 44 (4): 369–75. https://doi.org/10.1038/ng.2213.

Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. "GCTA: A Tool for Genome-Wide Complex Trait Analysis." *American Journal of Human Genetics* 88 (1): 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Yang, Jian, Noah A Zaitlen, Michael E Goddard, Peter M Visscher, and Alkes L Price. 2014. "Advantages and Pitfalls in the Application of Mixed-Model Association Methods." *Nature Genetics* 46 (2): 100–106. https://doi.org/10.1038/ng.2876.

Yengo, Loic, Julia Sidorenko, Kathryn E Kemper, Zhili Zheng, Andrew R Wood, Michael N Weedon, Timothy M Frayling, et al. 2018. "Meta-Analysis of Genome-Wide Association Studies for Height and Body Mass Index in ~700,000 Individuals of European Ancestry." *Human Molecular Genetics* 27 (20): 3641-49. https://doi.org/10.1093/hmg/ddy271.

Yu, Jianming, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, et al. 2006. "A Unified Mixed-Model Method for Association Mapping That Accounts for Multiple Levels of Relatedness." *Nature Genetics* 38 (2): 203–8. https://doi.org/10.1038/ng1702.

Zarate, Yuri A., Elizabeth Bhoj, Julie Kaylor, Dong Li, Yoshinori Tsurusaki, Noriko Miyake, Naomichi Matsumoto, et al. 2016. "SMARCE1, a Rare Cause of Coffin–Siris Syndrome: Clinical Description of Three Additional Cases." *American Journal of Medical Genetics Part A* 170 (8): 1967–73. https://doi.org/10.1002/AJMG.A.37722.

Zeng, Sicong, Tao Xiang, Tej K. Pandita, Ignacio Gonzalez-Suarez, Susana Gonzalo, Curtis C. Harris, and Qin Yang. 2009. "Telomere Recombination Requires the MUS81 Endonuclease." *Nature Cell Biology* 11 (5): 616–23. https://doi.org/10.1038/ncb1867.

Zhang, Jing, Megan Teh, Jamie Kim, Megan M Eva, Romain Cayrol, Rachel Meade, Anastasia Nijnik, et al. 2019. "A Loss-of-Function Mutation in the Integrin Alpha L (Itgal) Gene Contributes to Susceptibility to Salmonella Enterica Serovar Typhimurium Infection in Collaborative Cross Strain CC042." *Infection and Immunity* 88 (1): e00656-19. https://doi.org/10.1128/IAI.00656-19.

Zhang, Qian, Christopher G. Dove, Jyh Liang Hor, Heardley M. Murdock, Dara M. Strauss-Albee, Jordan A. Garcia, Judith N. Mandl, et al. 2014. "DOCK8 Regulates Lymphocyte Shape Integrity for Skin Antiviral Immunity." *Journal of Experimental Medicine* 211 (13): 2549–66. https://doi.org/10.1084/jem.20141307.

Zhang, Qianqian, Goutam Sahana, Guosheng Su, Bernt Guldbrandtsen, Mogens Sandø Lund, and Mario P. L. Calus. 2018. "Impact of Rare and Low-Frequency Sequence Variants on Reliability of Genomic Prediction in Dairy Cattle" *Genetics Selection Evolution* 50 (62). https://doi.org/10.1186/s12711-018-0432-8.

Zhang, Wei, Chao Xu, Chuanbing Bian, Wolfram Tempel, Lissete Crombet, Farrell MacKenzie, Jinrong Min, Zhonglai Liu, and Chao Qi. 2011. "Crystal Structure of the Cys2His2-Type Zinc Finger Domain of Human DPF2." *Biochemical and Biophysical Research Communications* 413 (1): 58–61. https://doi.org/10.1016/J.BBRC.2011.08.043.

Zhou, Xiang, and Matthew Stephens. 2012. "Genome-Wide Efficient Mixed-Model Analysis for Association Studies." *Nature Genetics* 44 (7): 821–24. https://doi.org/10.1038/ng.2310.

Zhu, Zhihong, Andrew Bakshi, Anna A.E. Vinkhuyzen, Gibran Hemani, Sang Hong Lee, Ilja M. Nolte, Jana v. van Vliet-Ostaptchouk, et al. 2015. "Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits." *American Journal of Human Genetics* 96 (3): 377–85. https://doi.org/10.1016/j.ajhg.2015.01.001.

Zilmer, Monica, Andrew C. Edmondson, Sumeet A. Khetarpal, Viola Alesi, Maha S. Zaki, Kevin Rostasy, Camilla G. Madsen, et al. 2020. "Novel Congenital Disorder of O-Linked Glycosylation Caused by GALNT2 Loss of Function." *Brain* 143 (4): 1114–26. https://doi.org/10.1093/brain/awaa063.

Zimin, Aleksey v, Arthur L Delcher, Liliana Florea, David R Kelley, Michael C Schatz, Daniela Puiu, Finnian Hanrahan, et al. 2009. "A Whole-Genome Assembly of the Domestic Cow, Bos Taurus." *Genome Biology 2009 10:4* 10 (4): 1–10. https://doi.org/10.1186/GB-2009-10-4-R42.

Zou, Yaqun, Daniela Zwolanek, Yayoi Izu, Shreya Gandhy, Gudrun Schreiber, Knut Brockmann, Marcella Devoto, et al. 2014. "Recessive and Dominant Mutations in COL12A1 Cause a Novel EDS/Myopathy Overlap Syndrome in Humans and Mice." *Human Molecular Genetics* 23 (9): 2339–52. https://doi.org/10.1093/HMG/DDT627.

# Appendix

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

# STATEMENT OF CONTRIBUTION
# DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

| | |
|---|---|
| Name of candidate: | Edwardo GM Reynolds |
| Name/title of Primary Supervisor: | Mathew Littlejohn |
| Name of Research Output and full reference: | |
| Reynolds, Edwardo G. M., Catherine Neeley, et al. 2021. "Non-Additive Association Analysis Using Proxy Phenotypes Identifies Novel Cattle Syndromes." Nature Genetics 53 (7): 949–54. https://doi.org/10.1038/s41588-021-00872-5. | |
| In which Chapter is the Manuscript /Published work: | 2 |
| Please indicate: | |
| • The percentage of the manuscript/Published Work that was contributed by the candidate: | 70% |
| and | |
| • Describe the contribution that the candidate has made to the Manuscript/Published Work: | |
| The candidate carried out many of the analyses under the guidance of supervisors. The candidate and the supervisors wrote and edited the manuscript. | |
| For manuscripts intended for publication please indicate target journal: | |
| Published in Nature Genetics | |
| Candidate's Signature: | EgmReynolds |
| Date: | 11/11/2021 |
| Primary Supervisor's Signature: | |
| Date: | 29/11/21 |

(This form should  appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)

DRC 16

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

## STATEMENT OF CONTRIBUTION
## DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

| | |
|---|---|
| Name of candidate: | Edwardo GM Reynolds |
| Name/title of Primary Supervisor: | Mathew Littlejohn |
| Name of Research Output and full reference: | |
| Reynolds, Edwardo G M, Thomas Lopdell et al. 2021. "Non-Additive QTL Mapping of Lactation Traits in 124,000 Sequence-Imputed Cattle Reveals Novel Recessive Loci." BioRxiv. https://doi.org/10.1101/2021.08.30.457863. | |
| In which Chapter is the Manuscript /Published work: | 5 |
| Please indicate: | |
| • The percentage of the manuscript/Published Work that was contributed by the candidate: | 90% |
| and | |
| • Describe the contribution that the candidate has made to the Manuscript/Published Work: | |
| The candidate carried out most of the experiments under the guidance of supervisors. The candidate wrote the manuscript, which was then edited by supervisors. | |
| For manuscripts intended for publication please indicate target journal: | |
| Published on BioRxiv, In review at Genetics Selection Evolution | |
| Candidate's Signature: | EgmReynolds |
| Date: | 11/11/2021 |
| Primary Supervisor's Signature: | |
| Date: | 29/11/21 |

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)