The genetic characterisation of coat colour and patterning traits in cattle

A thesis presented in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in

Animal Science

at Massey University, Al Rae Centre for Genetics and Animal Breeding, Hamilton, New Zealand

Swati Jivanji

2022

Abstract

Coat colour and patterning traits have been of interest to cattle breeders for centuries and have undergone intense selection due to their ability to provide easy means of breed identification. The striking coat colour and patterning traits observed in modern-day cattle breeds have reached fixation in many breed populations, and causal mutations for almost all major coat patterning traits in cattle have been resolved, with the exception of the white spotting trait characteristic of Holstein cattle. Despite this, the molecular mechanisms through which characterised mutations modulate pigmentation, and how they might interact with each other to produce different patterning traits, remain poorly understood. The aims of this thesis were to attempt to discover the causal variants responsible for the proportion of white spotting on the coat in *Bos taurus*, identify epistatic interactions between coat colour loci, and explore the implementation of alternative breeding solutions to introgress favourable coat-colour variants with minimal genetic drag. A combination of molecular, quantitative, and bioinformatic tools were utilised to discover a mutation in the protein-coding sequence of the *PAX3* gene, a non-coding mutation within a highly conserved region of the *MITF* gene, and a novel structural variant upstream of the *KIT* gene, that all likely contribute towards the proportion of white spotting on the coat. Using genome-wide association analyses we also described an epistatic interaction between the *MITF* and *KIT* loci that causes a splotchy face in cattle with Hereford parentage, and an epistatic interaction between two mutations at the *KIT* locus that causes pigmentation around the eyes in white-faced Hereford cattle. Finally, unbiased whole-genome sequence analysis and long-molecule sequencing demonstrated that CRISPR-Cas9 gene-editing could be used to introgress a *PMEL* coat colour dilution mutation into a Holstein-Friesian background with no detectable off-target mutagenesis or genetic drag. This thesis reports several novel

candidate causal mutations and epistatic interactions previously unknown to the literature, and also the first study in cattle to rigorously investigate off-target mutations associated with the application of CRISPR-Cas9 gene-editing technologies. The results presented here provide insight into biological and physiological aspects of pigment biology, enhance our understanding of these processes with relevance across mammals, and provide a context for future implementation strategies to transfer these variants across breeds and species.

Key words: *White spotting, coat patterning traits, PAX3, MITF, KIT, PMEL, white-face, ACOP, GWAS, long-molecule sequencing, gene-editing, CRISPR-Cas9*

Acknowledgements

The discoveries made during the course of my PhD journey would not have been possible without the collaborations fostered over the years, and the guidance and support of the people around me. I would like to extend my thanks to Mathew Littlejohn and Dorian Garrick for being my mentors, pushing me to be the best I can be, and being patient and supportive over the years.

I would like to acknowledge the Barry Foundation and Al Rae Centre for Genetics and Breeding for funding my doctoral fellowship, and the New Zealand eScience Infrastructure for providing the computational resources required for the analyses presented throughout this thesis. I would like to thank everyone at the Al Rae Centre for their friendship, support, and advice over the years, and everyone in the Research & Development team at Livestock Improvement Corporation for welcoming me and offering their support and encouragement. My special thanks are extended to Russell Snell from the University of Auckland, and everyone in his lab group, who welcomed me, and offered technical and moral support during my time with them. Thanks to Goetz Laible and his team at AgResearch who welcomed me into their group, and taught me about gene-editing and reproductive technologies, and to our collaborators Richard Mort and Emma Wilkinson from the University of Lancaster, who have provided valuable insight into melanocyte biology.

Last, but most certainly not least, I would like to thank my parents Sandhya and Devyash Patel, my brother Pratik Jivanji, and my husband Zeelesh Govind, for their unconditional support and encouragement throughout this journey.

Table of Contents

List of Tables

Chapter Three: Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle

Chapter Four: Novel structural and epistatic mutations at the *KIT* locus are associated with white patterning traits in bovine breeds

List of Figures

Chapter Two: Review of the Literature

Chapter Three: Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle

Chapter Four: Novel structural and epistatic mutations at the *KIT* locus are associated with white patterning traits in bovine breeds

List of Abbreviations

ACOP	Ambilateral circumocular pigmentation
AI	Artificial insemination
ASIP	Agouti signalling protein
BIK	Bovine infectious keratoconjunctivitis
BLAST	Basic local alignment search tool
BLAT	BLAST-like alignment tool
bp	Base-pair
BTA	Bos taurus autosome
cAMP	Cyclic adenosine monophosphate
Chr	Chromosome
CNV	Copy number variant
CREB	Cyclic adenosine monophosphate-responsive element binding protein
CRISPR	Clustered, regularly interspaced, palindromic repeats
crRNA	Crispr ribonucleic acid
DCT	Deopachrome tautomerase
ddPCR	Digital droplet polymerase chain reaction
DHICA	5,6-dihydroxyindol-2-carboxylic acid
DNA	Deoxyribonucleic acid
E9	Embryonic day 9
EDN3	Endothelin 3
EGFP	Enhanced green fluorescent protein
F_1	First filial
F_2	Second filial
Fig	Figure
FISH	Fluorescence in situ hybridisation
gDNA	Genomic deoxyribonucleic acid
GERP	Genomic evolutionary rate profiling
GRM	Genomic relationship matrix
gRNA	Guide ribonucleic acid
GWAS	Genome-wide association study
HD	High definition
HF	Holstein-Friesian
HF×J	Holstein-Friesian, Jersey cross
HS	Hereford, Shorthorn cross
IGV	Integrative genomics viewer
IVF	In vitro fertilisation
J	Jersey
kb	Kilobase
kg	Kilogram
KIT	KIT proto-oncogene, receptor tyrosine kinase
KITL	KIT proto-oncogene, receptor tyrosine kinase ligand

LD	Linkage disequilibrium
LINE	Long interspaced nuclear element
LOCO	Leave-one-chromosome-out
LYST	Lysosomal trafficking regulatory
M. bovis	Moraxella bovis
MAF	Minor allele frequency
Mb	Megabase
MC1R	Melanocortin 1 receptor
MITF	Microphthalmia-associated transcription factor
MLMA	Mixed linear model association
MOET	Multiple ovulation embryo transfer
N	Number
NAHR	Non-allelic homologous recombination
NZ	New Zealand
OSCC	Ocular squamous cell carcinoma
PAM	Protospacer adjacent motif
PAX3	Paired box gene 3
PCR	Polymerase chain reaction
PKA	Protein kinase A
PMEL	Premelanosome protein
QTL	Quantitative trait locus
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequence
RTE-BovB	Bovine retrotransposable element
SCNT	Single cell nuclear transfer
sgRNA	Single guide ribonucleic acid
SIFT	Sorting intolerant from tolerant
SNP	Single nucleotide polymorphism
SOX10	SRY-box transcription factor 10
SRA	Sequence read archive
SV	Structural variant
tracrRNA	Trans-activating crispr ribonucleic acid
TSS	Transcription start site
TWIST2	Twist family transcription factor 2
TYR	Tyrosinase
TYRP1	Tyrosinase-related protein 1
TYRP2	Tyrosinase-related protein 2
UCOP	Unilateral circumocular pigmentation
UV	Ultra-violet
VCF	Variant call file
VEP	Variant effect predictor
WGS	Whole genome sequence
α-MSH	α-Melanocyte stimulating hormone

CHAPTER ONE

General Introduction

1.1 General introduction

Humans have selected animals for millennia and the domestication of livestock has been invaluable to the transition of human society from nomadic hunter-gatherers to sedentary farming communities throughout much of Europe, Asia, and Africa. The domestication of cattle can be traced back to over 10,000 years ago starting with the wild aurochs (1). Centuries of artificial selection, where animals with desirable traits were chosen to be the parents of the next generation, have created the multitude of modern cattle breeds observed today, suited to a range of purposes. Historically, artificial selection was accomplished based on outward appearances or subjective measures of production capacity, without understanding which genes influenced characteristics of interest (2). Cattle were selected based on morphological traits due to the conviction that these traits were associated with milk, or meat production levels, and/or enhanced adaptation to the environment. Such traits included the lack of horns (i.e., polled), a slick hair coat, and various coat colour and patterning traits which are now understood to be underpinned by major-effect mutations (2). By the late 1700s and early 1800s, breeders had formed breed societies where they kept pedigree records for their cattle with the intention to identify genetically superior families (3). Almost a century later, some breed societies had established production recording programs and were using daughter-dam comparisons to improve the productivity of their cattle, but it was not until contemporary comparisons were utilised that breeders saw better results (4). These methods more accurately accounted for environmental influences on production traits, and when coupled with progeny testing followed by artificial insemination, genetic progress began in earnest. Genetic gain was further facilitated by advances in computing that allowed the implementation of mixed linear models which incorporated pedigree and performance data to inform breeding decisions (5). With the

sequencing of the bovine genome, and the advent of cost-effective, high-throughput single nucleotide polymorphism (SNP) chip genotyping technologies, a new era of artificial selection was conceived: genomic selection. Genomic selection harnessed genome-wide markers to estimate breeding values (typically of bulls) for quantitative traits using Bayesian regression models and machine learning (5). As biotechnologies such as genomics and gene-editing evolve, it is likely that we will continue to see artificial breeding strategies evolve with them.

Throughout the development of modern cattle breeds, coat colour and coat patterning traits have been of particular interest. Indeed, early Lasaux cave drawings of cattle with white spotting suggest that variation in coat patterns of domesticated cattle have been of interest to breeders for many centuries (1). Coat colour and patterning traits were captured by early breeders due to their ability to provide easy means of breed identification. The genetic effects that underlie these traits are largely assumed to have originally presented as highly penetrant *de novo* mutations, being selected by historic breeders due to their visually striking nature (1). Intense selection of these now 'breed-defining' coat patterning traits, such as white spotting in Holstein cattle (6), a white face in Hereford cattle (7), and a 'belted' pattern in Belted Galloway cattle (8), have driven some of these traits to fixation in breed populations. While some coat colour and patterning traits are caused by a single mutation (i.e., monogenic), some are caused by several large effect mutations (i.e., oligogenic), and/or epistatic interactions between multiple genes and gene products (8,9). Owing to the easy visual characterisation of coat colour and patterning traits, the availability of sequencing and genotyping technologies, and our innate fascination with colours and patterns, the causal mutations

for almost all major coat patterning traits in cattle have been resolved (8,10,11). One conspicuous exception is the white spotting trait of Holstein cattle.

1.2 Aims

This thesis aims to discover mutations that contribute towards white spotting in cattle, and identify epistatic interactions between coat patterning loci. Achievement of these aims will contribute insight into new aspects of developmental biology and the molecular genetics of cattle traits, and given the fundamental nature of these phenotypes, this work has broad relevance for mammalian biology. Additionally, we aim to explore the implementation of alternative breeding solutions to introgress favourable coat-colour relevant genetic variants, and list considerations for the use of gene-editing technologies in the agricultural sector.

1.3 Thesis outline

Chapter Two provides an overview of the relevant concepts and literature to serve as context to the results chapters of this thesis. This chapter covers the physiology and biochemistry of coat pigmentation, the genetic determinants of coat colour and patterning in cattle, the relevance of coat colour to animal welfare, and finally the implementation of gene-editing technology to accelerate genetic gain in livestock.

Chapter Three presents a continuation of work conducted during my MBiomedSc thesis that investigated genetic contributions to the proportion of white spotting in New Zealand dairy cattle. This work includes a larger cohort of cattle (with the addition of

894 animals), the application of more sophisticated methods for genome-wide association analyses, and a deeper exploration of candidate causal loci and mutations. This work was published in *Genetics Selection Evolution* (12), where the study also identifies a structural variant region downstream of the *KIT* gene which is the subject of Chapter Four.

In Chapter Four, we characterise the ancestral allele for the wild-type solid coat colour trait, and the structural variant that likely underlies the white spotted trait that is a hallmark of several cattle breeds, including Holsteins. This chapter highlights the inadequacies of the current bovine reference genome (based on a Hereford cow named 'Dominette') for investigating the genetic basis of coat colour traits in other breeds. Chapter Four also investigates epistatic interactions at the *KIT* locus that cause adulteration of the Hereford white-face trait. This chapter is presented as a draft manuscript that will be submitted to an international journal. At the time of writing, submission was awaiting the generation of lab-based functional work to be produced by our collaborators at the University of Lancaster, with the hope that the outcome of those experiments will enable publication of the study in a high impact journal.

Chapter Five further explores coat trait epistatic interactions, with an investigation into the manifestation of ambilateral circumocular pigmentation, or pigmentation around the eyes, in white-faced American Hereford cattle. Although this trait is not considered a 'true-breeding' characteristic, it has welfare implications and has been shown to provide some protection against ocular diseases such as ocular squamous cell carcinoma and

bovine infectious keratoconjunctivitis. This chapter has been accepted for publication in the *New Zealand Journal of Animal Science and Production* (13).

Although it is possible to identify major-effect causal mutations for coat colour and patterning traits, sometimes it is not practical or viable to introgress favourable genetic variation into other breeds. This is due to the losses in genetic gain associated with traditional cross breeding methods, a phenomenon known as linkage drag. As an exploration of gene-editing methods to potentially remedy this problem, we introduced a coat colour dilution mutation into a cloned Holstein-Friesian line via CRISPR-based methods, creating two gene-edited calves and three non-edited control cloned calves. This project was a proof of concept, where my role in the project was to genotype the cloned (edited and non-edited) calves for the white spotting tag variants identified in Chapter Three. This paper has not been presented in this thesis, but has been published in *BMC Genomics* (14). Although this work is not presented here, in Chapter Six I present a *de novo* mutagenesis analysis on the calves and the cell-lines they were derived from. This chapter presents important considerations that should be incorporated into future applications of gene-editing for breeding purposes, and has relevance for policy making when considering gene-editing in these contexts. The work presented in this chapter was also published in *BMC Genomics* (15).

Finally, in Chapter Seven I discuss the broader importance of the work described in this thesis. Discussion topics include genome reference considerations when investigating causal mutations, and the contribution of structural variation and epistasis to coat patterning traits. I also review the implications and selection utility of coat patterning

traits in modern cattle breeds, and theorise on the role of evolving molecular methods in the future of artificial selection.

1.4 List of Publications

Jivanji S, Worth G, Lopdell TJ, Yeates A, Couldrey C, Reynolds E, et al. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. Genetics Selection Evolution. 2019;51(1):1–18.

Jivanji S, Kosch T, Littlejohn M, Garrick D. Genome-wide association study of UV-protectant ambilateral circumocular pigmentation in Hereford cattle. New Zeal Journal of Animal Science and Production. 2021; *In Press*.

Laible G, Cole SA, Brophy B, Wei, Leath S, Jivanji S, et al. Holstein Friesian dairy cattle edited for diluted coat color as a potential adaptation to climate change. BMC Genomics. 2021;22(865):1-12.

Jivanji S, Harland C, Cole S, Brophy B, Garrick D, Snell R, et al. The genomes of precision edited cloned calves show no evidence for off-target events or increased de novo mutagenesis. BMC Genomics. 2021;22(1):1–14.

1.5 Publications – *In preparation*

Jivanji S, Mears E, Yeates A, Harland C, Gray C, Couldrey C, et al. Novel structural and epistatic mutations at the *KIT* locus associated with white patterning traits in bovine breeds. *In prep.*

1.6 References

1. Olson TA. Genetics of colour variation. In: The genetics of cattle. 1999. p. 33–53.

2. Littlejohn M, Harland C. Finding causal variants for monogenic traits in dairy cattle breeding. In: Advances in breeding of dairy cattle. Burleigh Dodds Science Publishing; 2019. p. 409–40.

3. Armitage PL. Developments in British cattle husbandry from the Romano-British period to early modern times. Ark. 1982.

4. VanRaden PM, Miller RH. The USDA animal improvement programs laboratory: A century old and just getting started. AIPL Research Reports. 2008.

5. Weigel KA, VanRaden PM, Norman HD, Grosu H. A 100-Year Review: Methods and impact of genetic selection in dairy cattle—From daughter–dam comparisons to deep learning algorithms. J Dairy Sci. 2017;100(12):10234–50.

6. Liu L, Harris B, Keehan M, Zhang Y. Genome scan for the degree of white spotting in dairy cattle. Anim Genet. 2009;40(6):975–7.

7. Whitacre L. Structural variation at the KIT locus is responsible for the piebald phenotype in Hereford and Simmental cattle. University of Missouri, Columbia; 2014.

8. Awasthi Mishra N, Drögemüller C, Jagannathan V, Keller I, Wüthrich D, Bruggmann R, et al. A structural variant in the 5'-flanking region of the TWIST2 gene affects melanocyte development in belted cattle. PLoS One. 2017;12(6):e0180170.

9. Schmutz SM, Dreger DL. Interaction of *MC1R* and *PMEL* alleles on solid coat colors in Highland cattle. Anim Genet. 2013;44(1):9–13.

10. Durkin K, Coppieters W, Drögüller C, Ahariz N, Cambisano N, Druet T, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature. 2012;482(7383):81–4.

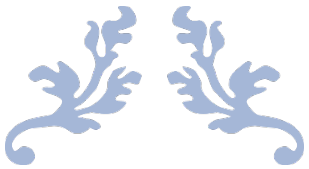
11. Brenig B, Beck J, Floren C, Bornemann-Kolatzki K, Wiedemann I, Hennecke S, et al. Molecular genetics of coat colour variations in White Galloway and White Park cattle. Anim Genet. 2013;44(4):450–3.

12. Jivanji S, Worth G, Lopdell TJ, Yeates A, Couldrey C, Reynolds E, et al. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. Genet Sel Evol. 2019;51(1):1–18.

13. Jivanji S, Kosch T, Littlejohn M, Garrick D. Genome-wide association study of UV-protectant ambilateral circumocular pigmentation in Hereford cattle. New Zeal J Anim Sci Prod. 2021;81. [In Press]

14. Laible G, Cole S-A, Brophy B, Wei, Leath S, Jivanji S, et al. Holstein Friesian dairy cattle edited for diluted coat color as a potential adaptation to climate change. BMC Genomics. 2021;22(865):1-12.

15. Jivanji S, Harland C, Cole S, Brophy B, Garrick D, Snell R, et al. The genomes of precision edited cloned calves show no evidence for off-target events or increased de novo mutagenesis. BMC Genomics 2021 221. 2021;22(1):1–14.



CHAPTER TWO

Review of the Literature

2.1 Cellular and molecular mechanisms of coat colour determination

2.1.1 The origin of melanocytes

The coat colour and coat patterning traits we observe in vertebrates are the result of complex processes that are initiated early during embryonic development. Pigment cell (melanocyte) precursors (melanoblasts) arise from early embryonic stem cells known as neural crest cells. Neural crest cells have their origin in the neural tube, a hollow structure destined to give rise to the brain and spinal cord (1,2). Neural crest cells are highly migratory multi-potent stem cells that contribute to many specialized structures and tissues in the developing embryo. The neural crest can be subdivided into five regions: cranial, cardiac, vagal, trunk, and sacral (Fig 2.1a). Neural crest cells leave, or delaminate from, the dorsal region of the neural tube by transitioning from epithelial to mesenchymal cells (Fig 2.1b & c), and the antero-posterior position at which this occurs largely determines the fate of these cells (3). Experimental evidence suggests that cell fate is most likely specified before cells leave the neural crest. Cells that delaminate from the trunk region are destined to give rise to neurons, melanocytes, and glial cells (4,5).

Due to the challenges in studying temporal aspects of melanocyte biology, mouse studies comprise much of what is known. In mice, cell specification and migration out of the neural tube occurs around embryonic day 9 (E9)(6). Melanoblasts are specified around E9 by downregulation of the transcription factors *Foxd3* and *Sox2*, and upregulation of microphthalmia-associated transcription factor (*Mitf*). Around E10.5, melanoblasts migrate out of the neural crest and gene-expression shifts to upregulate

melanoblast-specific gene premelanosome protein (*Pmel*) and dopachrome tautomerase (*Dct*). After melanoblasts leave the neural crest, they undergo a rapid expansion. Luciani et al. (6) found that there were less than 100 melanoblasts present in the trunk region at E10.5 in mice, but this increased to over 20,000 melanoblasts by E15.5. Migration of proliferated melanoblasts to the dermis is dependent on the endothelin 3 (EDN3) pathway, and interactions between the melanoblast Kit proto-oncogene receptor kinase (KIT) receptor and its ligand (KITL), which is expressed by dermal fibroblasts and keratinocytes (7,8). Signalling through the KIT receptor facilitates melanoblast migration and survival, and mutations in the *Kit* gene or its ligand have been associated with the absence of melanocytes, manifesting as partial or complete lack of pigmentation in the hair and skin, and/or other developmental defects (9,10). Seminal work in mice and chicks (11–13) formed the long-standing hypothesis that these expanding populations of melanoblasts migrate across a dorsolateral pathway through the developing dermis after leaving the neural crest. Mort et al. (14), however used tracing of single melanoblast clones to study the migratory pathway of melanocytes post-delamination from the neural crest, and found that melanoblasts do not appear to have a preferred directionality. Upon delaminating from the neural crest, melanoblasts proliferate and diffuse through the dermis in a density-dependent manner (14).

After melanoblasts delaminate from the neural crest, there is a shift in gene expression to facilitate migration and proliferation. There is an upregulation in E-cadherin and loss of dependence on EDN3 by E12.5 as melanoblasts migrate from the dermis to the epidermis. Approximately three days later, melanoblasts downregulate E-cadherin, begin to localize the basal layer of skin epidermis and developing hair follicles, where they differentiate into melanocytes. A subset of melanocytes dedifferentiate and

colonise the hair follicle bulge where they act as melanocyte stem cells to replenish the differentiated melanocyte population during subsequent hair cycles (1,2). Failure for melanocytes to proliferate or migrate efficiently at any stage during development results in regions lacking pigmentation, manifesting as regions of white hair.

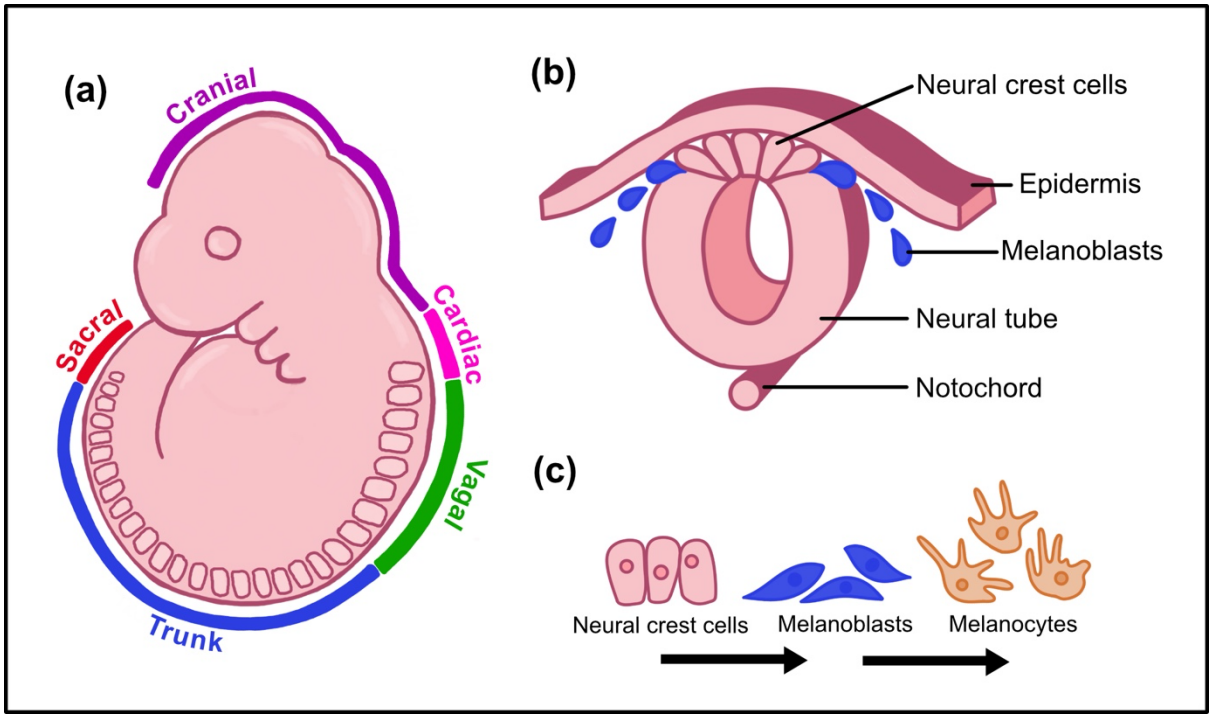


Figure 2.1 Overview of the neural crest and origin of melanoblasts. (a) Cranial, cardiac, vagal, trunk, and sacral regions of the neural crest highlighted on a 9.5 day old mouse embryo. The fate of neural crest cells is determined by which region they delaminate from. (b) Cross-section view of the neural crest in the trunk region. Some of the cells that delaminate from the trunk region differentiate into melanoblasts (blue). (c) When melanoblast precursors delaminate from the neural crest they transition from epithelial-like neural crest cells, to mesenchymal-like migratory melanoblasts, which eventually differentiate into mature dendritic melanocytes.

2.1.2 Pigment production

The complex epistatic interactions that occur within melanocyte cells occupying the hair follicle dictate pigment production and therefore the colour we observe on the hair coat

of mammals. The macromolecule melanin is a pigment granule produced within the melanosome, a melanocyte-specific organelle. Reactive intermediates, hydrogen peroxide and quinone, are generated during the production of melanin (melanogenesis), and so melanogenesis is restricted to the melanosome (15,16). Leakage of quinones into the cytoplasm is potentially toxic as they react with critical cellular macromolecules causing cytotoxicity and damage to the melanocyte (17,18). The melanosome matrix is predominantly composed of proteolytic fragments of the PMEL protein. The matrix sequesters reactive melanin intermediates, preventing damage to the melanosome (19). Mutations in the *Pmel* gene have been reported to cause dilution of the coat colour in *silver* mice (20), Highland, Galloway, and Charolais cattle (21–23), and complete hypopigmentation in Dominant White chickens (24).

There are two types of melanin that can be produced within the mammalian melanosome, eumelanin, responsible for brown/black pigment, and pheomelanin, responsible for red/yellow pigmentation. The ratio between these types of melanin is largely responsible for the coat colour variation we observe in mammals. Melanin granules are formed from an amino acid precursor L-tyrosine, through a series of enzymatic reactions largely reliant on tyrosinase (TYR), DCT (also referred to as tyrosinase-related protein 2), and tyrosinase-related protein 1 (TYRP1; Fig 2.2). TYR catalyses hydroxylation of L-tyrosine to L-DOPA, and the oxidation of DOPA to DOPAquinone. The hydroxylation of L-tyrosine to L-DOPA (l-3,4-dihydroxyphenylalanine) is the rate-limiting step of melanogenesis (25). The production of eumelanin (eumelanogenesis) and pheomelanin (pheomelanogenesis) diverge after the formation of DOPAquinone. In the absence of the amino acid cysteine, eumelanogenesis is favoured as DOPAquinone is converted into dopachrome, but in the

presence of excess cysteine pheomelanogenesis is favoured as DOPAquinone is converted into cysteinylDOPA (Fig 2.2) (26).

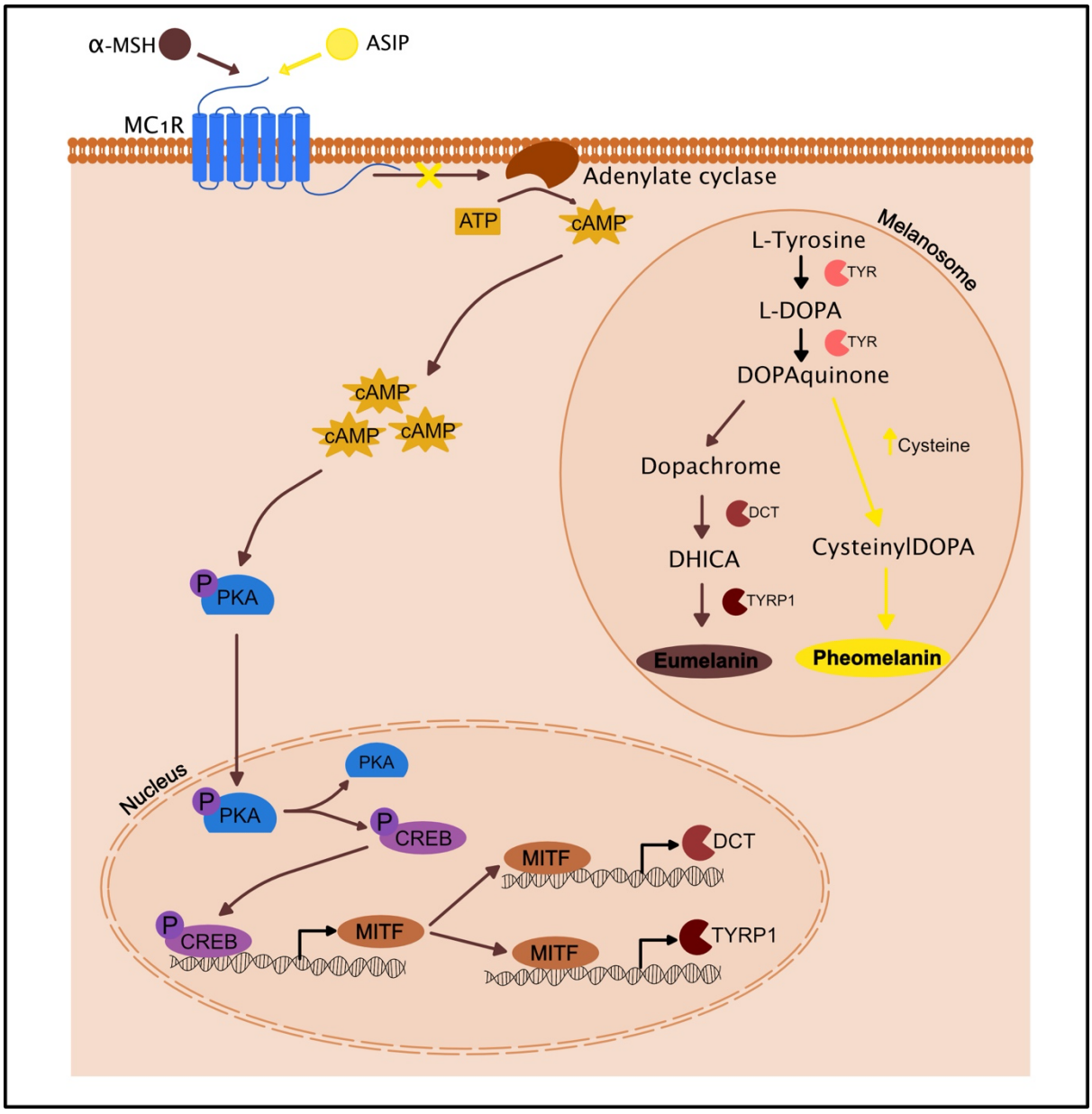


Figure 2.2 Pigment formation within the melanocyte. The signal cascade initiated by α-melanocyte stimulation hormone (α-MSH) binding to the melanocortin-1 receptor (MC1R) favours eumelanogenesis (pathway depicted in brown). If the agouti signalling protein (ASIP) binds to MC1R, adenylate cyclase is not activated, and pheomelanogenesis is favoured (pathway depicted in yellow).

Pigment type switching is largely reliant on the downstream effects of signalling through the melanocortin-1 receptor (MC1R) (Fig 2.2). Binding of the receptor agonist α-melanocyte stimulation hormone (α-MSH) to MC1R activates adenylate cyclase, and increases cytoplasmic concentrations of cyclic adenosine monophosphate (cAMP) (27). Elevated cAMP levels activate cAMP-dependent protein kinase A (PKA), which translocates to the nucleus and phosphorylates the *MITF* transcription factor cAMP-responsive element binding protein (CREB), increasing *MITF* expression. MITF upregulates the expression of *DCT* and *TYRP1* by binding to the M-box of the gene promoter regions (27). DCT catalyses conversion of dopachrome, to 5,6-dihydroxyindol-2-carboxylic acid (DHICA), and TYRP1 catalyses oxidation of DHICA into eumelanin (26). The agouti signalling protein (ASIP) competes with α-MSH in binding MC1R. ASIP is a MC1R antagonist, and if successful in binding, supresses adenylate cyclase. Suppression of adenylate cyclase and its associated signalling cascade, reduces the speed of pigment precursor formation and increases cellular cysteine levels, favouring pheomelanogenesis (26,28).

Mature follicular melanocytes are found in the upper matrix of the hair bulb below pre-cortical keratinocytes, with a small proportion of melanoblasts occupying the upper outer root sheath of the follicle bulge (Fig 2.3a). Hair bulb melanocytes are larger and more dendritic than epidermal melanocytes, and unlike epidermal pigmentation, follicular pigmentation is cyclic (29). Melanogenesis in hair bulb melanocytes is tightly coupled to the hair growth cycle, where melanogenesis is switched on during the hair growth phase (anagen), mature melanocytes undergo apoptosis during the regression phase (catagen), and melanogenesis remains absent during the quiescent phase (telogen) (30). During anagen and after melanin formation, melanosomes are transferred to

cortical and medullary keratinocytes in the growing hair shaft, pigmenting the growing hair. There are two hypotheses as to how this occurs: 1) the Shedding-Phagocytosis model, where melanocytes shed segments of their membrane enclosed around melanosome, that are then absorbed or phagocytosed by keratinocytes (Fig 2.3b) (31,32), and 2) the Exocytosis-Endocytosis model, where melanocytes release melanosomes into the extracellular space via exocytosis, and the keratinocytes internalise the melanosome via endocytosis (Fig 2.3c) (33). It remains unclear as to which of these mechanisms is responsible for pigmentation of the growing hair shaft. Melanoblasts from the upper outer root sheath replenish the population of hair bulb melanocytes during the early stages of anagen, for this process to begin again when a new hair forms (34).

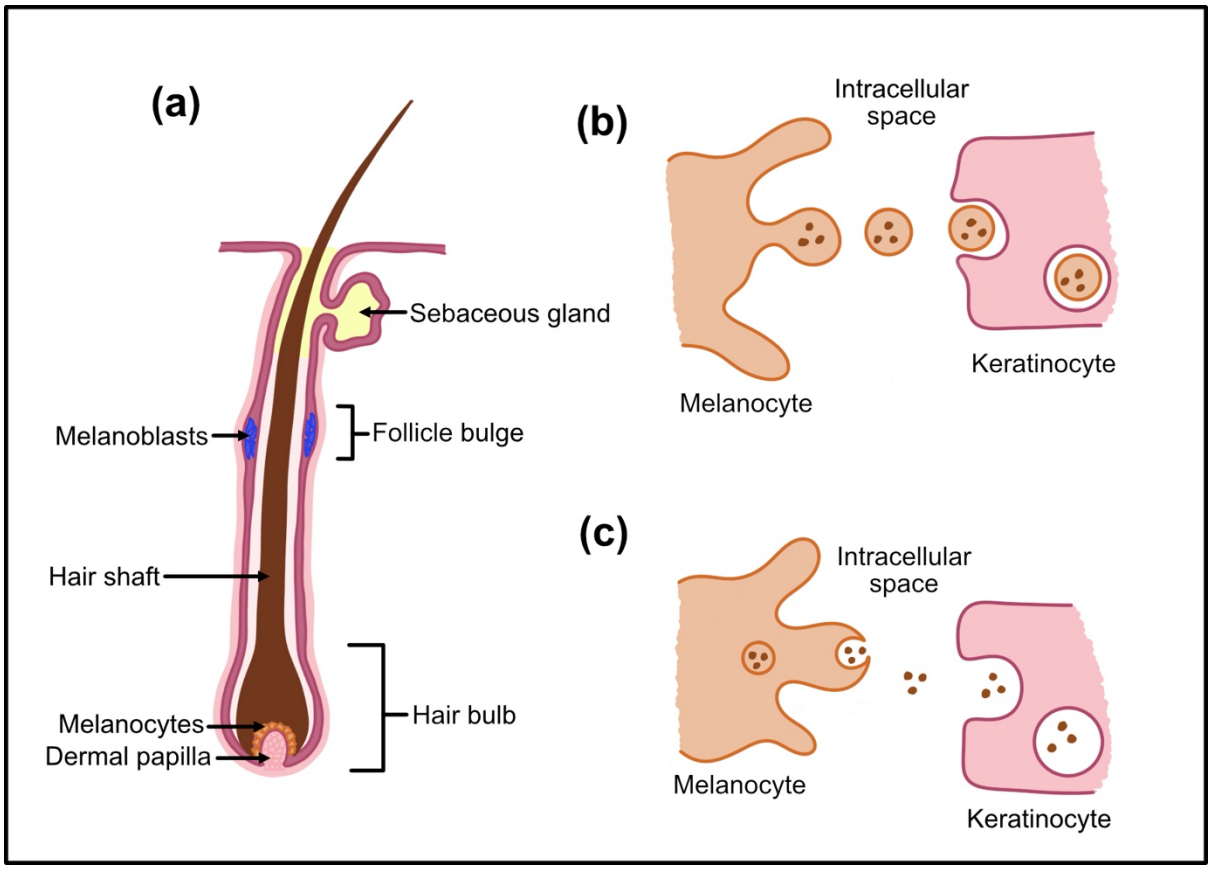


Figure 2.3 Structure and pigmentation of the hair shaft. (a) Melanoblasts occupy the follicle bulge of the hair where they act as a reservoir to replenish melanocytes during the growth phase of the hair cycle, and melanocytes occupy the upper matrix of the hair bulb and pigment the hair shaft as it grows. There are two popular hypotheses as to how the shaft is pigmented by melanocytes: (b) the Shedding-Phagocytosis model where the melanocyte blebs vesicles containing melanosomes, and keratinocytes take up the vesicles via phagocytosis, and (c) the Exocytosis-Endocytosis model where melanocytes release melanosomes into the intracellular space via exocytosis, which are then taken up by the keratinocytes via endocytosis.

2.2 Genetics of coat colour and coat patterning

2.2.1 Coat colour

As with many other traits of interest in cattle, coat colour and patterning traits are heritable and have been the focus of genetic selection efforts for hundreds, if not thousands of years. Selection for these instantly recognisable traits has generated a

myriad of different coat patterns, some of which are shown in Figure 2.4, and are discussed in the following section. Many of these traits have become breed-defining and are largely under-pinned by major-effect mutations, nearing fixation in some breed populations.

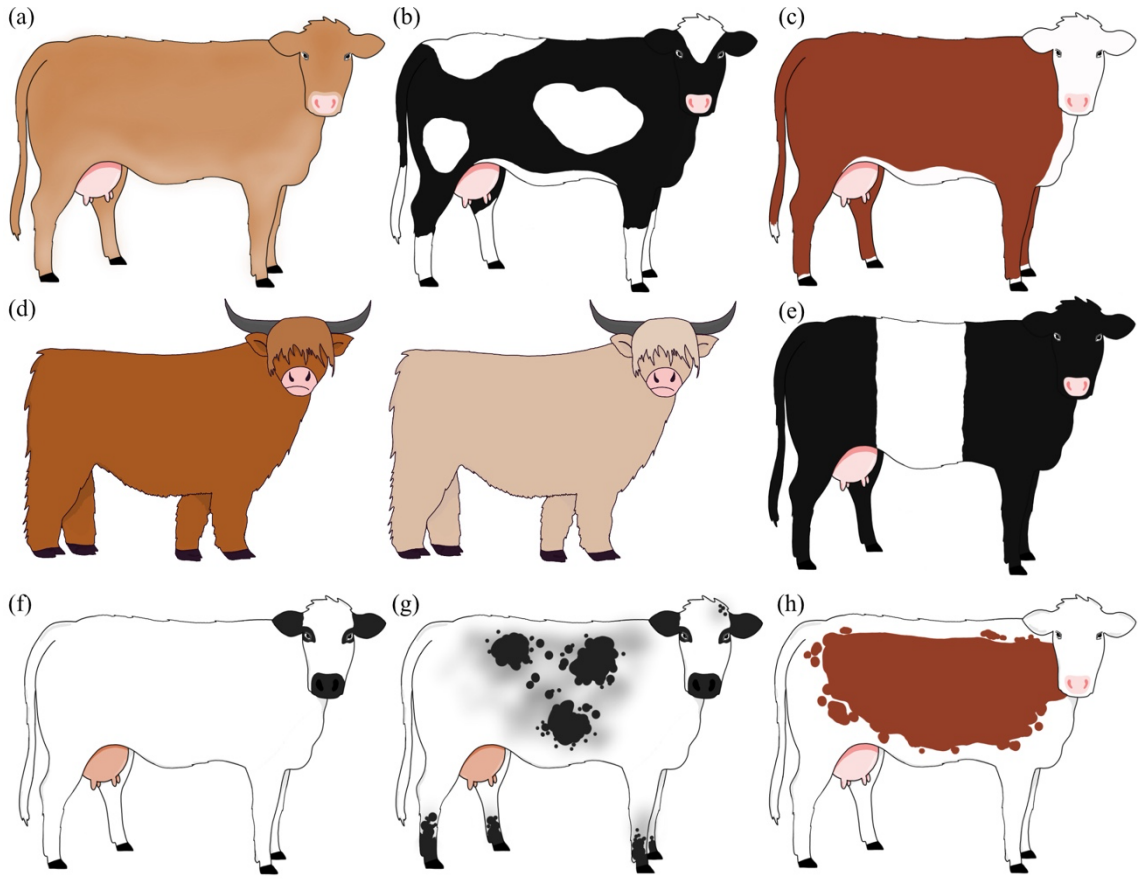


Figure 2.4 An illustration of some coat colour and patterning traits observed in modern day cattle breeds including: (a) a solid coloured cow, (b) a black and white spotted cow, (c) a red cow with a white face, (d) two cattle that represent dilution of the red coat pigmentation, (e) a belted cow, (f) a white park cow, (g) a speckled cow, and (h) a cow with the colour-sided trait.

Many coat characteristics are determined by single gene effects (i.e., Mendelian), perhaps more so than any other trait. The first pigmentation locus to be studied in cattle,

the *MC1R* gene, also referred to as the *extension locus*, is an example of this. The MC1R receptor is the major determinant of pigment-type switching and largely controls the ratio of pheomelanin (red) to eumelanin (black) pigment formation (see "Pigment production"; Fig 2.2). Klungland et al. (35), investigated the *MC1R* gene on chromosome 18 in five Norwegian cattle half-sib families and uncovered three functional alleles of this gene: E^+, E^D and *e*. The E^+ allele represents the most common form which is referred to as the wild-type and is sensitive to both receptor substrates, the receptor agonist α-MSH and the receptor antagonist ASIP. The dominantly inherited E^D allele is the product of a single base substitution, causing a change within the protein coding sequence of the *MC1R* gene. This change is hypothesised to allow binding of α-MSH to the MC1R receptor, but prevent ASIP binding, thus biasing pigment synthesis toward eumelanin and manifesting as black colouration of the coat. The recessively inherited *e* allele contains a premature termination signal, causing a non-functional MC1R receptor. Cattle that carry two copies of the *e* allele (homozygotes), appear red (35). Sequence analysis of the *MC1R* gene from 201 black-and-white, and 21 red-and-white Chinese Holstein cattle, suggested that the *e* allele described above, is likely causal for red coat colouration in this population (36). Animals that were homozygous for this mutation were all red-and-white, however those that only had one copy of the mutation (heterozygotes) displayed either phenotype, with six black-and-white and two red-and-white cattle observed (36). Animals carrying E^+/E^+ or E^+/e genotypes at the *extension locus* can have black, brown or red coat colouration, suggesting complex interactions that perhaps involve other genes and/or epistatic mechanisms that were not resolved in these studies (35).

Dilution of the base coat colours dictated by pheomelanin (red), or eumelanin (black) production have been observed in Charolais, Dexter, Highland and Galloway cattle. Khun & Weikard (23) investigated the dilution effect on coat colouration in 133 black-and-white German Holstein × Charolais crossbreed calves. Khun & Weikard (23) conducted a linkage analysis using 244 markers from across the genome, to investigate the relationship between transmission of these markers and the coat colour dilution phenotype within families. This study implicated a region on chromosome 5 in range of the *PMEL* gene as associated with coat colour dilution. Positional and functional analyses identified a mutation within the protein coding sequence of the *PMEL* gene as a candidate causal mutation for this trait (23). Gutierrez-Gil et al. (22) performed a linkage analysis using the same population reported by Khun & Weikard (23), incorporating additional cattle with red coat colour backgrounds and reciprocal backcrosses from the second filial (F_2) population (N=436 animals). The dilution effect was found to act on both black and red coat colour backgrounds (22). When the candidate mutation was fitted as a fixed effect in their regression model, the genome-wide significant effects identified for coat colour dilution were no longer significant, suggesting that coat colour dilution is either due to, or is in strong linkage disequilibrium (LD) with, the mutation. The identified mutation could not account for all phenotypic variation in coat colour dilution, which the authors attributed to either a double recombination event during meiosis, mis-scoring of the phenotype, or contributing effect from other loci influencing coat colour. Gutierrez-Gil et al. (22) suggest that the latter explanation is most likely, and implicated the *lysosomal trafficking regulator* (*LYST*) gene on chromosome 28 as a potential candidate for this effect. Schmutz & Dreger (21) investigated the *PMEL* gene as a candidate for coat colour dilution in 103 Highland cattle, with Charolais, Limousin, and Simmental cattle

included as controls. A three base pair deletion within the protein coding sequence of the *PMEL* gene was found to act in a codominant manner, where Highland cattle with one copy of the deletion had less dilution of the base coat colour than those that had two copies of the deletion. The deletion mutation was also found to segregate in Galloway cattle, but not in white Charolais cattle (21).

Berryere et al. (37) investigated the *TYRP1* gene sequence for genetic variation that may influence coat colour dilution in a variety of cattle breeds, including crossbred Charolais half-sibs, a Simmental full-sib family, Dexter cattle, Belted Galloway cattle, and Highland cattle. This analysis revealed a mutation within the protein coding sequence of the *TYRP1* gene that, in its homozygous state, segregated completely with the dun coat colour (dilution of the black base coat colour) in the Dexter cattle included in this study. Dexter cattle with a black coat colour were either heterozygous or homozygous wild-type at this position. This mutation was found only in Dexter cattle and could be ruled out as causal for coat colour dilution in all other breeds investigated (37).

2.2.2 Coat patterning traits

White spotting is a hallmark of Holstein-Friesian cattle, where they usually have a white tail, legs, and belly. Since these are true breeding characteristics, the proportion of white spotting was hypothesised to be highly heritable. A small number of previous studies have implicated major effect loci for white spotting in cattle with Holstein-Friesian ancestry. Liu et al. (38) used linkage analysis to investigate genetic variants influencing the proportion of white spotting in 737 F_2 Holstein-Friesian × Jersey crossbreed cows. They identified significant association signals on chromosomes 6, 18 and 22 and

suggested that the chromosomes 6 and 22 loci were likely underpinned by the *KIT* and *MITF* genes, respectively (38). Fontanesi et al. (39) compared the *MITF* gene sequence in spotted Italian Holstein and Simmental cattle, to that of solid coloured Italian Brown and Reggiana cattle, and found a region of sequence that is inherited together from a single parent (i.e., a haplotype) to be associated with the white spotting trait. The associated haplotype accounted for some, but not all variation observed in the white spotting phenotype, suggesting that the white spotting trait is either in strong LD with the causal mutation or influenced by additional loci (39). Hayes et al. (40) also implicated the *MITF* and *KIT* genes in a genome-wide association study (GWAS), an approach that involved scanning genetic markers from across the genome for an association with the proportion of black on the coat in black and white Holstein cows. An additional signal was present on chromosome 8, previously unobserved for white spotting. This signal implicated the *PAX5* gene as a potential candidate gene for the proportion of black spotting on the coat (40). Together, these studies report consensus for the candidacy of *KIT* and *MITF* in the manifestation and proportion of white spotting in Holstein-Friesian cattle, however the causal variants driving these effects have yet to be definitively identified, contrasting with the many mutations now catalogued for other coat characters (as discussed in the previous section).

Large (>1,000 base pair) deletions, insertions, duplications, inversions, and translocations of sequence, referred to as structural variants (SV), have been observed to influence coat patterning traits in cattle (Fig 2.5). Deletions and duplications are also referred to as copy number variants (CNVs), as cattle are expected to have two copies of each chromosome (diploid), but deletions result in less than two copies of the expected sequence, and duplications result in more than two copies of the expected sequence. The

belted trait, characterised by a circular belt of unpigmented hair and skin around the midsection in cattle, is an example of a trait caused by a CNV. Drogemuller et al. (41) conducted a genome scan using 186 microsatellite markers in 88 Brown Swiss cattle from six half-sib-families to investigate the genetic cause of this trait. Using an additional 19 genetic markers, Drogemuller et al. (41) identified a haplotype on chromosome 3 that was highly associated with the belted trait in this breed. Subsequent studies found that the same haplotype is associated with the belted trait in Belted Galloway and Lakenvelder, also known as Dutch Belted cattle (42). Whole genome sequence (WGS) comparison of a belted Brown Swiss cow and a Belted Galloway bull to 130 non-belted control cattle revealed a 6 kb duplication 16 kb upstream of the *TWIST2* gene (43) as a candidate mutation for the effect. Digital droplet polymerase chain reaction (ddPCR) of this region in 333 belted and 1,322 non-belted cattle suggested that the copy number ranged between 2 and 12 copies in belted cattle. The belted trait was highly associated with copy number at the duplicated site, and it was proposed that increased copy number was correlated with increased expression of the *TWIST2* gene. Indeed, transgenic overexpression of the bovine *TWIST2* gene hybridized to an enhanced green fluorescent protein (EGFP) cassette in zebrafish embryos demonstrated a significant decrease in the number of melanocytes 35 hours post-fertilisation compared to control samples (43).

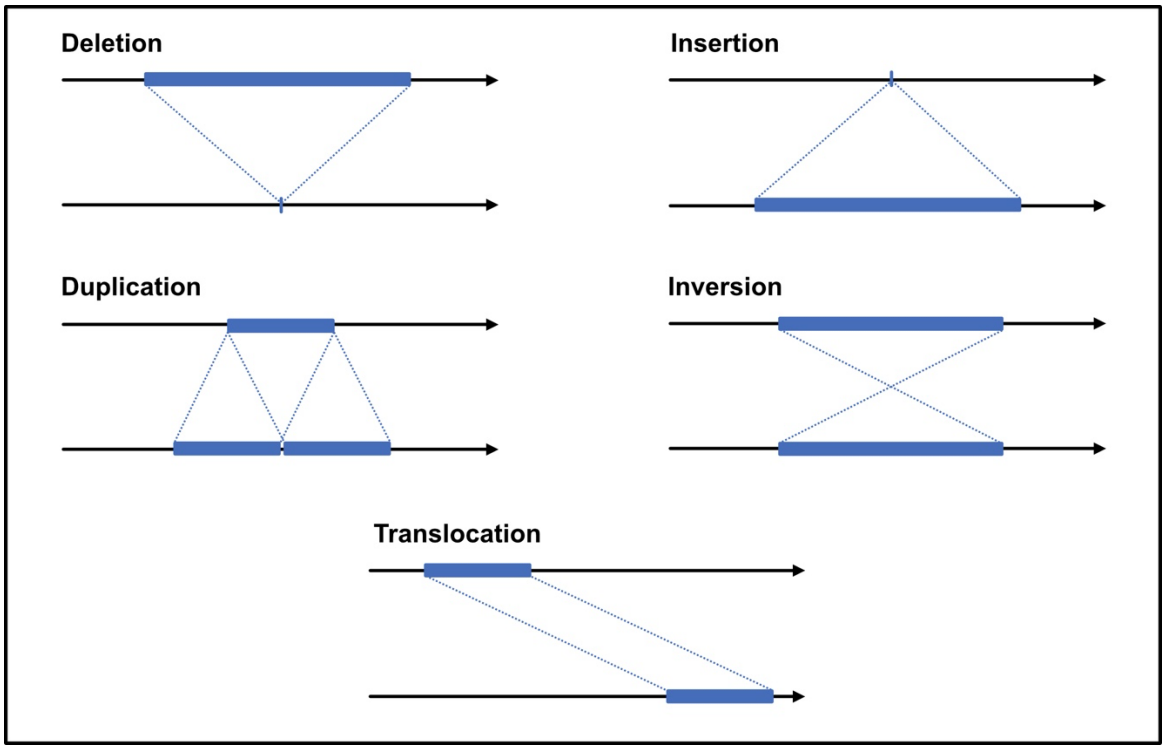


Figure 2.5 Different types of structural variants including deletion, insertion, duplication, inversion, and translocation events. Each structural variant depicted represents the reference form of the variant (top) and the mutant form (bottom) with the blue box representing the sequence affected, and the dotted line indicating the transformation of the affected sequence.

Structural variation at the *KIT* locus has been shown to underlie the 'colour-sidedness' phenotype in cattle, a trait characterised by a white streak down the spine, pigmented regions along the torso, and a white belly (44). The colour-sidedness trait in Belgian Blue cattle is controlled by a translocation of 492 kb of sequence encompassing the *KIT* gene to chromosome 29, and then a reciprocal translocation of 575 kb encompassing the fused chromosome 6 and 29 sequence back to the *KIT* locus (44). Using mate-pair cluster analysis and fluorescence *in situ* hybridisation (FISH), Durkin et al. (44) demonstrated that the translocation took place with circular intermediates, a CNV-generating mechanism that at time of publication was a novel, previously unreported mechanism. Brenig et al. (45) suggested that the initial translocation from chromosome

6 to chromosome 29 described by Durkin et al. (44) was also present in White Galloway and White Park cattle. White Galloway and White Park cattle that are homozygous for the translocation appear to be solid white in colour, cattle heterozygous for the translocation exhibit irregular black speckling on the coat, a black muzzle, black ears and black on the feet, and cattle wild-type for the translocation (i.e., have the reference form) have a solid black coat (45).

Austrian Pinzgaur cattle and some Tux-Zillertaler cattle exhibit a colour-sided coat pattern, similar to that seen in Belgian Blue and Brown Swiss cattle, caused by a different structural variant at the same locus. A GWAS of 27 spotted and 16 non-spotted Tux-Zillertaler cattle using 629,657 genetic markers, and subsequent haplotype analysis in 43 Tux-Zillertaler and 141 Pinzgauer cattle, identified an approximately 122 kb region of interest downstream from the *KIT* gene (46). Whole genome sequence comparison revealed fusion of a duplicated region from chromosome 4 into a deleted region on chromosome 6. Short read sequence data was able to resolve one breakpoint, but due to the presence of repetitive elements at the other breakpoints, these were not resolved (46). A study by Artesi et al. (47) observed the same mutation in colour-sided Gloucester cattle. Using long molecule sequencing Artesi et al. (47) described the structural variant as a 9.4 kb deletion at Chr6:70,417,067-70,417,114bp, a 310 kb duplication of Chr4:85,174,890-85,174,937bp, and subsequent insertion of this duplicated sequence into the site of the deletion on chromosome 6.

Hereford cattle have white on the belly and legs, sometimes have a white streak down the back, and are fixed for a characteristically white face that is inherited in a dominant

manner. To identify the causal mutation for the 'white face' trait, Whitacre (48) looked for regions at the *KIT* locus that were devoid of variation (i.e., showed evidence of fixation), and were exclusive to Hereford cattle. Four candidate regions were identified, however the discrepancies between expected and observed allele frequencies of the markers within each of these regions suggested that the reference genome used for sequence mapping (UMD3) may be misassembled at the *KIT* locus, perhaps due to one or multiple collapsed repeat region(s). Examination of a bovine optical map generated using DNA from the Hereford cow (named 'Dominette') used to generate the UMD3 reference genome, showed that there was a tandem duplication within the gene sequence, likely collapsed during genome assembly, and approximately 110 kb of consecutive sequence missing from the reference genome at the *KIT* locus. These observations highlighted the inadequacies of the UMD3 bovine reference genome for structural variant analysis at the *KIT* locus. Using *de novo* Hereford and Angus assemblies, Whitacre (48) discovered two duplications that could be causal for the white face trait, a ~4.5 kb duplication 50 kb upstream of the *KIT* gene, and a ~15 kb duplication located within intron one (i.e., a non-protein coding region) of the *KIT* gene, both segregating with the white face phenotype (48).

2.3 Coat colour and animal welfare

2.3.1 MITF pleiotropy and auditory-pigmentary disorders

The *MITF* gene is a well described coat patterning gene, where many mutations affecting the expression of *MITF* have been identified to influence coat patterning in a variety of different mammals including cattle, dogs, horses, and mice (39,49–52). As highlighted in the previous section, the *MITF* gene has been associated with white

spotting in cattle (38–40,53), and has also been associated with bilateral deafness in German Fleckvieh cattle (54). Philipp et al. (54) identified a family of German Fleckvieh cattle that had incomplete albinism, heterochromatic irises and bilateral hearing loss. The inheritance of these traits appeared to be autosomal dominant and completely penetrant. A GWAS using 774,660 genetic markers across 88 cattle (seven affected and 79 non-affected) found a single significant signal on chromosome 22 near the *MITF* gene. Sequencing of the *MITF* gene found that all affected cattle had one copy of a mutation in exon seven. The group of phenotypes observed in affected Fleckvieh cattle were similar to those observed in human auditory-pigmentary syndromes, Waardenburg Syndrome Type 2A, and Tietz Syndrome (55), consisting of sensorineural hearing loss, very pale skin, light-coloured eyes, and sometimes abnormal pigmentation of the hair. Complete silencing of *MITF* expression in pigs (56) and mutations in the coding region of *MITF* in horses (57) have also been associated with auditory-pigmentary syndromes.

2.3.2 KIT pleiotropy and gonadal hypoplasia

Like *MITF*, *KIT* is another gene that plays many roles beyond its impact on coat colour and patterning. The *KIT* gene plays a vital role during embryonic development where knock-outs of the gene are embryonic lethal (58), and some protein coding mutations have been reported to cause several diseases in humans including mast cell disease and a variety of tumours (59–61). Northern Finncattle and Swedish Mountain cattle, also known as Swedish Highland cattle, carry a *KIT* mutation that causes both coat patterning and disease. Northern Finncattle and Swedish Mountain cattle vary in colour and patterns, where some appear to have white coats with pigmented ears and pigmented muzzles, and others with spotted sides and pigmented legs, similar to other

breeds such as Belgian Blue and Brown Swiss cattle (44,62). Some cattle with largely white coats also manifest gonadal hypoplasia (62). Gonadal hypoplasia is characterised by small and underdeveloped gonads, caused by failed migration and proliferation of primordial germ cells. In Northern Finncattle and Swedish Mountain cattle, gonadal hypoplasia affects both sexes equally and typically manifests on the left side, causing impaired fertility. Venhoranta et al. (62) conducted a GWAS and CNV analysis for gonadal hypoplasia in 21 affected cattle and 73 non-affected cattle, using 777,962 genetic markers. Two highly significant signals on chromosomes 6 and 29 were identified in these analyses and further cytogenetic and PCR analyses revealed that an approximately 500 kb region harbouring the entire *KIT* gene was duplicated and translocated to a region on chromosome 29. All affected cattle in this study were homozygous for the structural variant, and Venhoranta et al. (62) suggest a recessive mode of inheritance with incomplete penetrance. This mutation has been previously described to cause 'colour-sidedness' in Belgian Blue and Brown Swiss cattle (44) among other breeds, however to our knowledge, gonadal hypoplasia has not been reported in these breeds.

2.3.3 Rat-tail syndrome

Rat-tail syndrome is a form of hereditary hypotrichosis hypothesised to be caused by an interaction between coat colour loci. This disorder is observed in crosses between black cattle breeds (Angus and Holstein cattle) and some European breeds that carry the coat colour dilution trait (Simmental and Charolais). Rat-tail syndrome is characterised by missing hairs at the tail switch, and malformed, short, curly and sometimes sparse hair that only manifests within regions of pigmentation on the hair coat. Schalles & Cundiff (63) visually scored 660 Simmental × Angus crossbred calves for rat-tail syndrome to

investigate the health implications of these conformational changes to the hair coat. They identified 64 calves with rat-tail syndrome, and found that there was no difference in birth weight, weaning weight or weight gain from birth to weaning between calves affected by rat-tail syndrome and the other calves, but affected calves had significantly reduced rates of weight gain during winter months. Schalles & Cundiff (63) proposed that reduced growth rate during winter months may be indicative of impaired thermoregulation due to abnormal hair structure in pigmented regions of the hair coat.

Knaust et al. (64) investigated the incidence of rat-tail syndrome in a Charolais × German Holstein population and found that the disease only manifested in cattle that had a least one copy of the *MC1R* dominant black allele, and that were heterozygous for a *PMEL* coat colour dilution mutation. However, only 50% of cattle that carried this combination of genotypes manifested with rat-tail syndrome. Linkage analysis on 388 F_2 cattle across 6,802 genetic markers, revealed significant loci co-locating to the expected regions in proximity to the *MC1R* gene on BTA18, and the *PMEL* gene on BTA5. An additional signal was seen on BTA5 at 18 Mb, far from the *PMEL* gene that mapped to 57.6 Mb. To refine the localization of the additional signal observed at 18 Mb on BTA5, a GWAS was performed on a subset of the population that had at least one dominant black allele, and one copy of the *PMEL* dilution allele. This GWAS was conducted on 384 cattle across 37,218 genetic markers, and suggested the *KITL* gene as a positional candidate. Sequencing of the *KITL* gene coding region, and transcriptome analysis across BTA5 10-25 Mb, did not reveal any candidate causal mutations, perhaps suggesting that the causal mutation is regulatory in nature. Knaust et al. (64) concluded that the incidence of rat-tail syndrome is likely the result of an epistatic interaction

between the dominant black allele at the *MC1R* locus, the dilution mutation at the *PMEL* locus, and a mutation at the *KITL* locus.

2.3.4 Ocular disease

The lack of pigmentation around the eyes and prolonged exposure to ultraviolet (UV) radiation are two major predisposing factors for at least two ocular diseases in cattle: ocular squamous cell carcinoma, and bovine infectious keratoconjunctivitis (65–67). Ocular squamous cell carcinoma (OSCC) is the most prevalent form of malignant tumour affecting cattle, characterised by a chronically invasive tumour that metastasizes via draining lymphatic vessels of the head and neck (66). The occurrence of OSCC causes pain and discomfort to the affected animal and poses economic strain on farmers as treatment is intensive and costly, and if caught too late, carcases from affected cattle are condemned at meat processing plants (66,68). Bovine infectious keratoconjunctivitis (BIK), commonly referred to as pinkeye, also causes pain and discomfort to the affected animal, and in severe cases causes eye disfigurement and blindness. BIK is characterised by conjunctivitis and ulcerative keratitis, and in some (but not all) cases is caused by infection from the highly contagious *Moraxella bovis* (*M. bovis*) (69). Increased risk of developing OSCC or BIK has been associated with the lack of pigmentation around the eyes, often seen in white-faced cattle such as Simmental, Hereford, and Fleckvieh cattle (65–67).

A longitudinal study that followed 396 purebred Hereford cattle over five years observed that 51% of the study population were identified to have a lesion on their eye at least once during their lifetime (70). Nishimura & Frisch (71) studied the association

between OSCC and pigmentation around the eyes, or circumocular pigmentation, in Hereford × Shorthorn (HS), Brahman × HS, and Africander × HS crossbred cattle, and purebred Brahman and Africander cattle. Circumocular pigmentation was measured in 1,750 eight- to ten-month-old calves, and the incidence of OSCC was measured in 1,495 of these cattle. Nishimura & Frisch (71) found that the incidence of OSCC was significantly higher in cattle with lower proportions, or no pigmentation around their eyes. Ward & Nielson (72), reported a similar observation in cattle with BIK. Circumocular pigmentation and the occurrence of BIK infection were recorded in 963 cows and 679 calves across Hereford, Hereford × Angus, and Charolais × Hereford × Angus breeds over four years. Ward & Nielson (72) found that cattle with a lower circumocular pigmentation score, particularly Hereford cattle, had significantly higher incidence of BIK, and more severe disease outcomes. It is not completely understood why the lack of pigmentation around the eyes increases the incidence and severity of BIK.

Genetic variation associated with the manifestation of circumocular pigmentation has been investigated in Fleckvieh cattle using breeding values from 320,186 cows and genotypes from their collective 3,579 sires (73). This study reported a heritability of 0.79 for ambilateral circumocular pigmentation (ACOP), and identified twelve major effect quantitative trait loci (QTL), presenting eight candidate genes: *MCM6*, *PAX3*, *KITLG*, *ERBB3*, *KIT*, *ATRN*, *MITF* and *NBEAL2*. The causal mutations that underpin these effects remain unresolved (73).

2.3.5 Heat stress

A disruption in thermo-homeostasis collectively resulting from interactions between coat colour, coat morphology, and the environment, results in heat stress. Heat stress has been associated with reduced milk production (74), fertility (75), and immune function (76) in dairy cattle and is becoming an increasingly relevant issue facing dairy farmers as global temperatures continue to rise. Acute heat stress is sufficient to invoke a cascade of intra- and extra-cellular responses that alter physiological responses within an animal, where cattle display a variety of short-term and prolonged responses during heat stress, including increased breathing rates, increased sweating, increased water intake, decreased feed intake, decreased milk production, altered milk composition, and altered blood hormone concentrations (77). Cattle also adopt shade seeking behaviour, display crowding, change orientation relative to the sun, stand in water, stand next to water troughs and may demonstrate refusal to lie down (78). Heat stress has severe animal welfare implications and the potential for economic losses for farmers.

Hansen (79) used nine predominantly white Holstein cows and 11 predominantly black Holstein cows to test the hypothesis that cattle with a higher proportion of white spotting on their coat have reduced heat stress and consequently increased production compared to their darker counterparts. Cows with predominantly white coats were found to have lower rectal temperatures, skin surface temperatures, breaths per minute, and open mouth panting when exposed to UV radiation with no shade, in comparison to cows with predominantly black coats. Cows with predominantly white coats, kept in non-shaded areas, exhibited an average decrease in milk production of 1.5 kg per day compared to those kept in shaded areas, and predominantly black cows exhibited a daily average decrease of 3.3 kg per day (79). These results demonstrate a difference in

response to solar radiation exposure and its effect on production, however interpretation of these results should be tempered by the small sample size investigated. Bercerril et al. (80) conducted an observational study on 4,293 lactating Holstein cows across eight dairy farms in Florida to investigate the effect of proportion of white spotting on reproductive performance and milk production. Proportion of white spotting was visually scored based on a side-on view of each cow and milk production was measured over 100 days. Milk production and the proportion of white spotting on the coat showed a positive linear relationship, with a regression coefficient of 1.91 kg per percentage of white over a single lactation, and reproductive performance was comparatively worse in Holstein cows with a greater proportion of black on their coat. Differences in fat and protein production with proportion of white spotting were in the expected direction, but were not statistically significant (80). Despite these observations, not all studies have concluded that an increased proportion of white spotting on the coat positively impacts production in Holstein cows. Goodwin et al. (81) and Hansen (79), demonstrated that predominantly black Holstein-Frisian cows produced more milk than those that were predominantly white, although both of these studies exposed the cows to only short periods of solar radiation.

Although a greater proportion of white on the coat appears to have a positive impact on thermo-homeostasis when exposed to direct sunlight, serious and extensive cutaneous tissue damage in areas of white coat colour has been observed in Holstein cattle, likely attributed to the greater transmittance coefficients of the white hairs compared to darker hairs (82). These factors should be taken into consideration before making breeding decisions. It is also worth noting that other major effect coat character genes have been implicated in thermal tolerance in cows. Autosomal dominant mutations in the *prolactin*

receptor (*PRLR*) gene and its receptor ligand *prolactin hormone* (*PRL*) have antagonistic, pleiotropic effects on hair development and thermoregulation. A frameshift mutation in the *PRLR* gene causes a short 'slick' coat in Senepol cattle, while a missense mutation in the *PRL* causes excessive hairiness in Holstein-Friesian cattle (83). Interestingly, the differences in response to heat stress between hairy and slick cattle is not likely due solely to hair coat length, with differences in sweating ability, irrespective of coat length, shown by Littlejohn et al. (83) and others (84).

2.4 CRISPR-Cas9 as a tool for introgressing favourable variation

2.4.1 Gene-editing as a tool for accelerated genetic gain

Livestock improvement programs select bulls that carry desirable genetic variants associated with enhanced welfare, production, efficiency and/or sustainability, to sire the next generation of cattle and increase the expression of these favourable traits within herds. This information can be leveraged via genomic prediction, or marker assisted selection for major effect Mendelian traits. In the latter context, if these variants exist within the breed of interest, a GWAS could be used to discover the relevant mutation(s), which could then be used in breeding schemes. However, breeding challenges arise when desirable traits do not segregate within the breed of interest. These traits can be introgressed via crossbreeding, although this results in genetic drag where other variation that might not be desirable in all farming contexts is also inherited (for example beef versus dairy characters). Many generations of backcrossing are thus required to eliminate genetic drag, stalling genetic gain. Stabilising the introgressed trait within the larger breed population presents another issue, where increased selection

intensity and overuse of a small handful of the relevant sires to propagate the introgressed trait can lead to inbreeding, inbreeding depression, and recessive disorders.

As an alternative, though as yet a mostly theoretical alternative, gene-editing could be used to introgress desirable genetic variation from other breeds into populations without costs to genetic gain. One example of this approach is introduction of the polled allele that causes hornlessness in cattle. Disbudding and dehorning of dairy cattle to improve the safety of farm workers, and decrease injury to cattle from aggressive behaviour is a common practice on farm, but is increasingly seen as an animal welfare issue (85). The polled allele segregates in some (generally low genetic merit) cattle with a low allele frequency, and some beef breeds, but selection of dairy cattle with the polled allele, or introgression of the polled allele from beef cattle, would cause losses in genetic gain that would take many generations of breeding to regain (86). Carlson et al. (87) used transcription activator-like effector nucleases (TALENs) to introduce a 212bp duplication in place of a 10bp deletion on chromosome 1, known as the Celtic polled allele, into four different bovine embryonic fibroblast lines. Four cloned calves were born of three gene-edited lines, three of which were homozygous for the gene-edit, and the other heterozygous. No evidence of horn buds were observed in these calves (87). Two of the three homozygous polled calves were kept to be founder animals for a breeding program, however partial integration of the gene-editing vector was found to linger in the genomes of these calves, rendering them unfit for purpose (88). Nevertheless, the use of TALENs to introduce a welfare-relevant trait into dairy cattle within a single generation, without losses in genetic gain associated with crossbreeding, was proof of concept. The clustered, regularly interspaced, palindromic repeats (CRISPR)-associated (Cas) system has since been popularised and refined for more

flexible, precise, and accurate gene-editing. The major limiting factor of using gene-editing technologies for genetic improvement, is that the causal mutation with a large effect on the trait of interest must be known to pursue implementation. Given the identity of several coat colour mutations are now known, and that these variants may benefit breeding objectives through welfare (i.e., heat tolerance) or animal parentage identification, such mutations make attractive examples of traits that could be introgressed via editing. These targets, and other beneficial Mendelian characters are discussed further below.

2.4.2 An overview of CRISPR-Cas9

The type II bacterial CRISPR-Cas system is a ribonucleic acid (RNA)-guided adaptive immune system used by bacteria and archaea to respond to viral and plasmid challenges (89). This system has been manipulated to enable versatile, efficient, and targeted modification of the genome, i.e., gene-editing. To facilitate gene-editing a Cas9 endonuclease, typically isolated from *Streptococcus pyogenes* bacteria, is complexed with a synthetic single guide RNA (sgRNA). The sgRNA is composed of a crispr RNA (crRNA), a 17-20-nucleotide sequence designed to have simple base-pair complementarity with a target site, and a trans-activating crRNA (tracrRNA), which serves as a scaffold for the Cas nuclease to bind to (90,91). The sgRNA recognises a 5'-NGG-3' protospacer adjacent motif (PAM) sequence on the complement DNA strand at the targeted site, and allows complementation with up to 20 nucleotides of target DNA sequence. Upon recognition of the target site, the Cas9 endonuclease makes a blunt cut between the 17th and 18th base in the target sequence. This blunt cut can either be repaired via non-homologous end joining (NHEJ), or a repair template can be introduced to induce homology-directed repair (HDR; Fig 2.6) (90,91). Several early

studies demonstrated that it is not uncommon for shorter target recognition sites to result in unwanted off-target mutations, largely due to the potential for non-unique matching and sequence mismatches distal from the PAM sequences at the 5' end of the sgRNA (92–94). In terms of widespread implementation of the technology for medicine and breeding applications, one of the major concerns of using CRISPR-Cas9 for gene-editing is the potential for unwanted off-target events. Accurately detecting such effects is therefore of key interest, and will ultimately be required to introgress new mutations into agricultural populations.

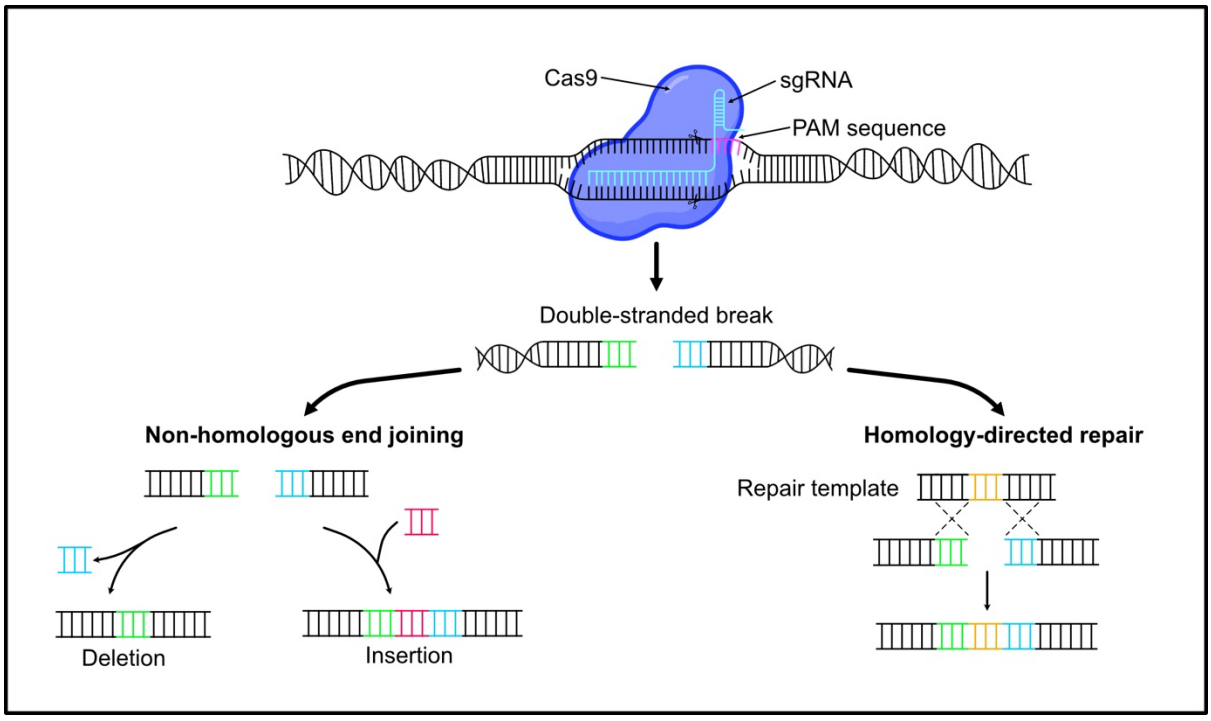


Figure 2.6 CRISPR-Cas9 induced non-homologous end joining (NHEJ) and homology-directed repair (HDR). The Cas9 enzyme creates a double-stranded break at a site targeted by the sgRNA, which is then either repaired via NHEJ or HDR. NHEJ results in insertions/deletions at the double-stranded break site, whereas HDR can introduce precise variation by recombination with a repair template.

2.4.3 CRISPR-Cas9 induced off-target mutations

Mutations at sites other than the targeted edit site, also referred to as 'off-target' mutations or 'secondary' mutations, occur when the sgRNA binds to sites other than the intended target site. Bioinformatic tools such as Cas-OFFinder (95), CasOT (96), and CT-Finder (97) can be used to identify regions of the genome that have sequence homology with the sgRNA, where non-specific binding may occur and induce off-target events. Some studies have used these tools to preselect sites for amplification and off-target mutation investigation, but a major limitation of these studies is that they neglect to consider potential off-target events at sites with low on-target similarity, and thus cannot detect these events should they occur (95–97).

Schaefer et al. (98) used an unbiased WGS approach to detect off-target mutations in mice edited for a blindness correcting mutation in the *Pde6b* gene. These mice were edited via injection of a plasmid carrying the sgRNA, a single stranded repair template, and the Cas9 protein at the one-cell embryo stage (99). Two gene-edited mice and one unrelated non-edited control mouse were sequenced, and variant calls at sites with at least 23× sequence coverage were retained for analysis (98). Sequence comparison between the gene-edited and control mice revealed hundreds of off-target mutations in the gene-edited samples. However, this result was later found to be flawed when the authors reported no excess mutations upon conducting WGS analysis with additional mouse lines (100). The mutations Schaefer et al. (98) reported as 'off-target mutations' were suggested to be the result of non-editing-related spontaneous *de novo* mutagenesis, i.e., mutations that arise due to asymmetric cell division during development, or mutations that were inherited from the parents, but not present in the other mice. This study was the first to highlight the importance of considering inherited and spontaneous

de novo mutations in such investigations, and the importance of using appropriate controls that facilitate the ability to distinguish between spontaneous *de novo* mutation and inherited variation from legitimate CRISPR-Cas9 induced off-target mutations.

2.4.4 CRISPR-Cas9 gene-editing in large animals

A limited number of studies have demonstrated the application of CRISPR-Cas9 mediated gene-editing in livestock and attempted to quantify off-target mutagenesis using WGS data. Li et al. (101) used a trio-based experimental design to assess *de novo* mutation rates in off-spring of gene-edited goats. Li et al. (101) generated two male and three female gene edited goats by injection of four sgRNAs targeting the *MSTN* and *FGF5* genes in one-cell stage embryos. Disruption of the *MSTN* gene resulted in increased muscle mass, and disruption of the *FGF5* gene increased the number and length of hair fibres, both economically valuable traits. The five gene edited goats, one non-edited female, and five off-spring were sequenced and whole genome sequence data were interrogated for single nucleotide variants, small insertions or deletions, and structural variants that might be due to CRISPR-Cas9 off-target effects. The incidence of *de novo* mutations was low in the gene-edited founder population and their progeny, where the estimated average *de novo* mutation rate was equivalent to that observed in human and cattle populations. Analysis of transcription data (RNA sequence data) from the progeny provided valuable biological insight into the transcriptional consequences of *MSTN* and *FGF5* knockout mutations, highlighting the use of CRISPR-Cas9 in understanding the cellular and molecular consequences of potentially favourable mutations prior to large scale selection (101). Wang et al. (102) used CRISPR-Cas9 to genetically modify sheep through injection of six sgRNA targeting three genes, *MSTN, ASIP* and *BCO2*, into one cell stage embryos. Trio-based whole genome sequence data

from the gene edited sheep and their parents showed that the mutational load in gene edited off-spring is equivalent to the mutation rate in human studies. One of the three gene edited sheep harboured a large 2.4 kb inversion at the *MSTN* locus, thought to be due to a simultaneous double stranded break induced by Cas9 at two sgRNA target sites, however this sheep did not show visible or clinical evidence of any abnormalities. Although the inversion was found to be rare (one in 54 sheep), it highlights the importance of a robust post-editing pipeline for the detection of off-target mutations, and the power of whole genome sequence data in detecting large structural variants induced by CRISPR-Cas9 gene editing. As an aside, in addition to detailing off-target detection approaches, both Li et al. (101) and Wang et al. (102) demonstrated the usefulness and success of multiplexed gene editing via CRISPR-Cas9 in large animals in these studies.

2.5 References

1. Mort RL, Jackson IJ, Patton EE. The melanocyte lineage in development and disease. Development. 2015;142(7):1387–1387.
2. Kelsh RN, Harris ML, Colanesi S, Erickson CA. Stripes and belly-spots-A review of pigment cell morphogenesis in vertebrates. Semin Cell Dev Biol. 2009;20(1):90–104.
3. Dupin E, Creuzet S, Le Douarin NM. The contribution of the neural crest to the vertebrate body. Adv Exp Med Biol. 2006;589:96–119.
4. Henion PD, Weston JA. Timing and pattern of cell fate restrictions in the neural crest lineage. Development. 1997;124(21):4351–9.
5. Krispin S, Nitzan E, Kassem Y, Kalcheim C. Evidence for a dynamic spatiotemporal fate map and early fate restrictions of premigratory avian neural

crest. Development. 2010;137(4):585–95.

6. Luciani F, Champeval D, Herbette A, Denat L, Aylaj B, Martinozzi S, et al. Biological and mathematical modeling of melanocyte development. Development. 2011;138(18):3943–54.

7. Jordan SA, Jackson IJ. MGF (KIT Ligand) Is a Chemokinetic Factor for Melanoblast Migration into Hair Follicles. Dev Biol. 2000;225(2):424–36.

8. Lee HO, Levorse JM, Shin MK. The endothelin receptor-B is required for the migration of neural crest-derived melanocyte and enteric neuron precursors. Dev Biol. 2003;259(1):162–75.

9. Reith AD, Rottapel R, Giddens E, Brady C, Forrester L, Bernstein A. W mutant mice with mild or severe developmental defects contain distinct point mutations in the kinase domain of the c-k t receptor. Genes Dev. 1990;4(3):390–400.

10. Richards KA, Fukai K, Oiso N, Paller AS. Case report: A novel KIT mutation results in piebaldism with progressive depigmentation. J Am Acad Dermatol. 2001;44(2):288–92.

11. Rawles ME. The development of melanophores from embryonic mouse tissues grown in the coelom of chick embryos. Proc Natl Acad Sci. 1940;26(12):673–80.

12. Rawles ME. Origin of pigment cells from the neural crest in the mouse embryo. Physiol Zool. 1947;20(3):248–66.

13. Mintz B. Gene control of mammalian pigmentary differentiation, I. Clonal origin of melanocytes. Proc Natl Acad Sci USA. 1967;58(1):344–51.

14. Mort RL, Ross RJH, Hainey KJ, Harrison OJ, Keighren MA, Landini G, et al. Reconciling diverse mammalian pigmentation patterns with a fundamental mathematical model. Nat Commun. 2016;7(1):1–13.

15. Meyskens FL, McNulty SE, Buckmeier JA, Tohidian NB, Spillane TJ, Kahlon

RS, et al. Aberrant redox regulation in human metastatic melanoma cells compared to normal melanocytes. Free Radic Biol Med. 2001;31(6):799–808.

16. Nappi AJ, Vass E. Hydrogen peroxide generation associated with the oxidations of the eumelanin precursors 5,6-dihydroxyindole and 5,6-dihydroxyindole-2-carboxylic acid. Melanoma Res. 1996;6(5):341–9.

17. Borovansky J, Mirejovsky P, Riley PA. Possible relationship between abnormal melanosome structure and cytotoxic phenomena in malignant melanoma. Neoplasma. 1991;4(38):393–400.

18. Pawelek JM, Lerner AB. 5,6-Dihydroxyindole is a melanin precursor showing potent cytotoxicity. Nature. 1978;276(5688):627–8.

19. Simon JD, Peles D, Wakamatsu K, Ito S. Current challenges in understanding melanogenesis: Bridging chemistry, biological control, morphology, and function. Pigment Cell Melanoma Res. 2009;22(5):563–79.

20. Kwon BS, Halaban R, Ponnazhagan S, Kim K, Chintamaneni C, Bennett D, et al. Mouse silver mutation is caused by a single base insertion in the putative cytoplasmic domain of Pmel 17. Nucleic Acids Res. 1995;23(1):154–8.

21. Schmutz SM, Dreger DL. Interaction of *MC1R* and *PMEL* alleles on solid coat colors in Highland cattle. Anim Genet. 2013;44(1):9–13.

22. Gutiérrez-Gil B, Wiener P, Williams JL. Genetic effects on coat colour in cattle: dilution of eumelanin and phaeomelanin pigments in an F2-Backcross Charolais × Holstein population. BMC Genet. 2007;8(1):56.

23. Kühn C, Weikard R. An investigation into the genetic background of coat colour dilution in a Charolais × German Holstein F2 resource population. Anim Genet. 2007;38(2):109–13.

24. Kerje S, Sharma P, Gunnarsson U, Kim H, Bagchi S, Fredriksson R, et al. The

Dominant white, Dun and Smoky color variants in chicken are associated with insertion/deletion polymorphisms in the PMEL17 gene. Genetics. 2004;168(3):1507–18.

25. Hearing VJ, Tsukamoto K. Enzymatic control of pigmentation in mammals. FASEB J. 1991;5(14):2902–9.

26. Ito S, Wakamatsu K, Ozeki H. Chemical analysis of melanins and its application to the study of the regulation of melanogenesis. Pigment Cell Res. 2000;13(8):103–9.

27. Bertolotto C, Busca R, Ballotti R, Ortonne JP. Cyclic AMP is a key messenger in the regulation of skin pigmentation. Pigment Cell Res. 2001;17(2):177–85.

28. Morgan AM, Lo J, Fisher DE. How does pheomelanin synthesis contribute to melanomagenesis? Two distinct mechanisms could explain the carcinogenicity of pheomelanin synthesis. BioEssays. 2013;35(8):672–6.

29. Tobin DJ, Paus R. Graying: Gerontobiology of the hair follicle pigmentary unit. Exp Gerontol. 2001;36(1):29–54.

30. Slominski A, Paus R. Melanogenesis is coupled to murine anagen: Toward new concepts for the role of melanocytes and the regulation of melanogenesis in hair growth. J Invest Dermatol. 1993;101(s1):90S-97S.

31. Ando H, Niki Y, Ito M, Akiyama K, Matsui MS, Yarosh DB, et al. Melanosomes are transferred from melanocytes to keratinocytes through the processes of packaging, release, uptake, and dispersion. J Invest Dermatol. 2012;132(4):1222–9.

32. Wu XS, Masedunskas A, Weigert R, Copeland NG, Jenkins NA, Hammer JA. Melanoregulin regulates a shedding mechanism that drives melanosome transfer from melanocytes to keratinocytes. Proc Natl Acad Sci USA.

2012;109(31):E2101–9.

33. Tarafder AK, Bolasco G, Correia MS, Pereira FJC, Iannone L, Hume AN, et al. Rab11b mediates melanin transfer between donor melanocytes and acceptor keratinocytes via coupled exo/endocytosis. J Invest Dermatol. 2014;134(4):1056–66.

34. Nishimura EK, Jordan SA, Oshima H, Yoshida H, Osawa M, Moriyama M, et al. Dominant role of the niche in melanocyte stem-cell fate determination. Nature. 2002;416(6883):854–60.

35. Klungland H, Vfige DI, Gomez-Raya L, Adalsteinsson S, Lien S. The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. Mamm Genome. 1995;6:636–9.

36. Li Q-L, Li J-B, Zhang Z-F, Wang H-M, Wang C-F, Gao Y-D, et al. Study on red ocat color gene and prediction of the secondary structure in Chinese Holstein. Agric Sci China. 2008;7(8):1016–21.

37. Berryere TG, Schmutz SM, Schimpf RJ, Cowan CM, Potter J. TYRP1 is associated with dun coat colour in Dexter cattle or how now brown cow? Anim Genet. 2003;34(3):169–75.

38. Liu L, Harris B, Keehan M, Zhang Y. Genome scan for the degree of white spotting in dairy cattle. Anim Genet. 2009;40(6):975–7.

39. Fontanesi L, Scotti E, Russo V. Haplotype variability in the bovine MITF gene and association with piebaldism in Holstein and Simmental cattle breeds. Anim Genet. 2012;43(3):250–6.

40. Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME. Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS

Genet. 2010;6(9):e1001139.

41. Drögemüller C, Engensteiner M, Moser S, Rieder S, Leeb T. Genetic mapping of the belt pattern in Brown Swiss cattle to BTA3. Anim Genet. 2009;40(2):225–9.

42. Drögemüller C, Demmel S, Engensteiner M, Rieder S, Leeb T. A shared 336 kb haplotype associated with the belt pattern in three divergent cattle breeds. Anim Genet. 2010;41(3):304–7.

43. Awasthi Mishra N, Drögemüller C, Jagannathan V, Keller I, Wüthrich D, Bruggmann R, et al. A structural variant in the 5'-flanking region of the TWIST2 gene affects melanocyte development in belted cattle. PLoS One. 2017;12(6):e0180170.

44. Durkin K, Coppieters W, Drögüller C, Ahariz N, Cambisano N, Druet T, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature. 2012;482(7383):81–4.

45. Brenig B, Beck J, Floren C, Bornemann-Kolatzki K, Wiedemann I, Hennecke S, et al. Molecular genetics of coat colour variations in White Galloway and White Park cattle. Anim Genet. 2013;44(4):450–3.

46. Küttel L, Letko A, Häfliger IM, Signer-Hasler H, Joller S, Hirsbrunner G, et al. A complex structural variant at the KIT locus in cattle with the Pinzgauer spotting pattern. Anim Genet. 2019;50(5):423–9.

47. Artesi M, Tamma N, Deckers M, Karim L, Coppieters W, Van den Broeke A, et al. Colour-sidedness in Gloucester cattle is associated with a complex structural variant impacting regulatory elements downstream of *KIT*. Anim Genet. 2020;51(3):461–5.

48. Whitacre L. Structural variation at the KIT locus is responsible for the piebald phenotype in Hereford and Simmental cattle. University of Missouri, Columbia;

2014.

49. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, et al. Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet. 2007;39(11):1321–8.

50. Baranowska Körberg I, Sundström E, Meadows JRS, Rosengren Pielberg G, Gustafson U, Hedhammar Å, et al. A Simple Repeat Polymorphism in the MITF-M Promoter Is a Key Regulator of White Spotting in Dogs. PLoS One. 2014;9(8):e104363.

51. Hauswirth R, Haase B, Blatter M, Brooks SA, Burger D, Drögemüller C, et al. Mutations in MITF and PAX3 cause "splashed white" and other white spotting phenotypes in horses. PLoS Genet. 2012;8(4):e1002653.

52. Baxter LL, Hou L, Loftus SK, Pavan WJ. Spotlight on spotted mice: A review of white spotting mouse mutants and associated human pigmentation disorders. Pigment Cell Res. 2004;17(3):215–24.

53. Hofstetter S, Seefried F, Häfliger IM, Jagannathan V, Leeb T, Drögemüller C. A non-coding regulatory variant in the 5′-region of the MITF gene is associated with white-spotted coat in Brown Swiss cattle. Anim Genet. 2019;50(1):27–32.

54. Philipp U, Lupp B, Mömke S, Stein V, Tipold A, Eule JC, et al. A MITF mutation associated with a dominant white phenotype and bilateral deafness in German Fleckvieh cattle. PLoS One. 2011;6(12):4–9.

55. Léger S, Balguerie X, Goldenberg A, Drouin-Garraud V, Cabot A, Amstutz-Montadert I, et al. Novel and recurrent non-truncating mutations of the MITF basic domain: genotypic and phenotypic variations in Waardenburg and Tietz syndromes. Eur J Hum Genet. 2012;20(5):584–7.

56. Chen L, Guo W, Ren L, Yang M, Zhao Y, Guo Z, et al. A de novo silencer

causes elimination of MITF-M expression and profound hearing loss in pigs. BMC Biol. 2016;14(1):1–15.

57. Henkel J, Lafayette C, Brooks SA, Martin K, Patterson-Rosa L, Cook D, et al. Whole-genome sequencing reveals a large deletion in the MITF gene in horses with white spotted coat colour and increased risk of deafness. Anim Genet. 2019;50(2):172–4.

58. Sun G, Liang X, Qin K, Qin Y, Shi X, Cong P, et al. Functional analysis of KIT gene structural mutations causing the porcine dominant white phenotype using genome edited mouse models. Front Genet. 2020;11:138.

59. Yavuz AS, Lipsky PE, Yavuz S, Metcalfe DD, Akin C. Evidence for the involvement of a hematopoietic progenitor cell in systemic mastocytosis from single-cell analysis of mutations in the c-kit gene. Blood. 2002;100(2):661–5.

60. Vila L, Liu H, Al-Quran SZ, Coco DP, Dong HJ, Liu C. Identification of c-kit gene mutations in primary adenoid cystic carcinoma of the salivary gland. Mod Pathol. 2009;22(10):1296–302.

61. Andea AA, Patel R, Ponnazhagan S, Kumar S, Devilliers P, Jhala D, et al. Merkel cell carcinoma: Correlation of KIT expression with survival and evaluation of KIT gene mutational status. Hum Pathol. 2010 ;41(10):1405–12.

62. Venhoranta H, Pausch H, Wysocki M, Szczerbal I, Hänninen R, Taponen J, et al. Ectopic KIT copy number variation underlies impaired migration of primordial germ cells associated with gonadal hypoplasia in cattle (Bos taurus). PLoS One. 2013;8(9):e75659.

63. Schalles RR, Cundiff L V. Inheritance of the “rat-tail” syndrome and its effect on calf performance. J Anim Sci. 1999;77(5):1144–7.

64. Knaust J, Hadlich F, Weikard R, Kuehn C. Epistatic interactions between at least

three loci determine the “rat-tail” phenotype in cattle. Genet Sel Evol. 2016;48(1):26.

65. D’Mello S, Finlay G, Baguley B, Askarian-Amiri M. Signaling pathways in melanogenesis. Int J Mol Sci. 2016;17(7):1144.

66. Heeney JL, Valli VEO. Bovine ocular squamous cell carcinoma: An epidemiological perspective. Can J Comp Med. 1985;49(1):21–6.

67. Seid A. Review on infectious bovine keratoconjunctivitis and its economic impacts in cattle. J Dairy Vet Sci. 2019;9(5):555774.

68. Williams DL. Welfare Issues in Farm Animal Ophthalmology. Vol. 26, Veterinary Clinics of North America - Food Animal Practice. 2010. p. 427–35.

69. Snowder GD, Van Vleck LD, Cundiff L V., Bennett GL. Genetic and environmental factors associated with incidence of infectious bovine keratoconjunctivitis in preweaned beef calves. J Anim Sci. 2005;83(3):507–18.

70. Russell WC, Brinks JS, Kainer RA. Incidence and heritability of ocular squamous cell tumors in Hereford cattle. J Anim Sci. 1976 Dec 1;43(6):1156–62.

71. Nishimura H, Frisch JE. Eye cancer and circumocular pigmentation in bos taurus, bos indicus and crossbred cattle. Aust J Exp Agric. 1977;17(88):709–11.

72. Ward JK, Nielson MK. Pinkeye (bovine infectious keratoconjunctivitis) in beef cattle. J Anim Sci. 1976;49(2):361–6.

73. Pausch H, Wang X, Jung S, Krogmeier D, Edel C, Emmerling R, et al. Identification of QTL for UV-protective eye area pigmentation in cattle by progeny phenotyping and genome-wide association analysis. PLoS One. 2012;7(5):e36346.

74. Keister ZO, Moss KD, Zhang HM, Teegerstrom T, Edling RA, Collier RJ, et al. Physiological responses in thermal stressed Jersey cows subjected to different

management strategies. J Dairy Sci. 2010;85(12):3217–24.

75. Flamenbaum I, Galon N. Management of heat stress to improve fertility in dairy cows in Israel. J Reprod Dev. 2010;56(S):S36–41.

76. Elvinger F, Natzke RP, Hansen PJ. Interactions of heat stress and bovine somatotropin affecting physiology and immunology of lactating cows. J Dairy Sci. 2010;75(2):449–62.

77. Kadokawa H, Sakatani M, Hansen PJ. Perspectives on improvement of reproduction in cattle during heat stress in a future Japan. Anim Sci J. 2012;83(6):439–45.

78. Mitlohner FM, Galyean ML, McGlone JJ. Shade effects on performance, carcass traits, physiology, and behavior of heat-stressed feedlot heifers. J Anim Sci. 2002;80(8):2043–50.

79. Hansen PJ. Effects of coat colour on physiological responses to solar radiation in Holsteins. Vet Rec. 1990;127(13):333–4.

80. Bercerril CM, Wilcox CJ. Determination of percentage of white coat color from registry certificates in Holsteins. J Dairy Sci. 1992;75(12):3582–6.

81. Goodwin PJ, Josey M, Cowan JM. Coat color and its effect on production in Holstein-Friesians in Southeast Queensland. Aust Assoc Anim Breed Genet Conf. 1995;11(1995):295–8.

82. Morais D. Variation in the coat characteristics, thyroid hormone levels and milk yield of dairy cows in a hot, dry environment. Universidade Estadual Paulista; 2002.

83. Littlejohn MD, Henty KM, Tiplady K, Johnson T, Harland C, Lopdell T, et al. Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. Nat Commun. 2014;5:1–8.

84. Dikmen SE, Khan FA, Huson HJ, Sonstegard TS, Moss JI, Dahl GE, Hansen PJ. The SLICK hair locus derived from Senepol cattle confers thermotolerance to intensively managed lactating Holstein cows. J Dairy Sci. 2014;97(9):5508–20.

85. Ventura BA, Keyserlingk MAG von, Weary DM. The welfare of dairy cattle: perspectives of industry stakeholders. In: The Ethics of Consumption: The Citizen, the Market and the Law. Wageningen Academic Publishers, Wageningen; 2013. p. 221–4.

86. Medugorac I, Seichter D, Graf A, Russ I, Blum H, Göpel KH, et al. Bovine polledness – An autosomal dominant trait with allelic heterogeneity. PLoS One. 2012;7(6):e39477.

87. Carlson DF, Lancto CA, Zang B, Kim E-S, Walton M, Oldeschulte D, et al. Production of hornless dairy cattle from genome-edited cell lines. Nat Biotechnol 2016 345. 2016;34(5):479–81.

88. Chakraborty S. Unreported off-target integration of beta-lactamase from plasmid in gene-edited hornless cows. OSF Preprints; 2019.

89. Wiedenheft B, Sternberg SH, Doudna JA. RNA-guided genetic silencing systems in bacteria and archaea. Nature. 2012;482(7385):331–8.

90. Cong L, Zhang F. Genome engineering using crispr-cas9 system. Chromosom Mutagen Second Ed. 2014;8(11):197–217.

91. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816–22.

92. Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat Biotechnol. 2013;31(9):839–43.

93. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. 2013;31(9):822–6.

94. Zhang X-H, Tee LY, Wang X-G, Huang Q-S, Yang S-H. Off-target effects in CRISPR/Cas9-mediated genome engineering. Mol Ther - Nucleic Acids. 2015;4:e264.

95. Bae S, Park J, Kim J-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics. 2014;30(10):1473–5.

96. Xiao A, Cheng Z, Kong L, Zhu Z, Lin S, Gao G, et al. CasOT: A genome-wide Cas9/gRNA off-target searching tool. Bioinformatics. 2014;30(8):1180–2.

97. Zhu H, Misel L, Graham M, Robinson ML, Liang C. CT-Finder: A Web Service for CRISPR Optimal Target Prediction and Visualization. Sci Rep. 2016;6(1):1–8.

98. Schaefer K, Wu W, Colgan D, Tsang S, Bassuk A, Mahaja V. Unexpected mutations after CRISPR–Cas9 editing in vivo. Nat Methods. 2017;14(6):547.

99. Wu WH, Tsai YT, Justus S, Lee TT, Zhang L, Lin CS, et al. CRISPR repair reveals causative mutation in a preclinical model of retinitis pigmentosa. Mol Ther. 2016;24(8):1388–94.

100. Schaefer KA, Darbro BW, Colgan DF, Tsang SH, Bassuk AG, Mahajan VB. Corrigendum and follow-up: Whole genome sequencing of multiple CRISPR-edited mouse lines suggests no excess mutations. bioRxiv. 2017. p. 154450.

101. Li C, Zhou S, Li Y, Li G, Ding Y, Li L, et al. Trio-based deep sequencing reveals a low incidence of off-target mutations in the offspring of genetically edited goats. Front Genet. 2018;9:449.

102. Wang X, Liu J, Niu Y, Li Y, Zhou S, Li C, et al. Low incidence of SNVs and indels in trio genomes of Cas9-mediated multiplex edited sheep. BMC Genomics. 2018;19(1):397.

CHAPTER THREE

Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle

PUBLISHED IN GENETICS SELECTION EVOLUTION (2019)

Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle

Swati Jivanji[1]*, Gemma Worth[2], Thomas J. Lopdell[2], Anna Yeates[2], Christine Couldrey[2], Edwardo Reynolds[1], Kathryn Tiplady[2], Lorna McNaughton[2], Thomas J.J. Johnson[2], Stephen R Davis[2], Bevin Harris[2], Richard Spelman[2], Russell G Snell[3], Dorian Garrick[1] and Mathew D. Littlejohn[2]

[1]Massey University Manawatu, Private Bag 11 222, Palmerston North 4442, New Zealand

[2]Livestock Improvement Corporation (LIC), 605 Ruakura Rd, Newstead 3286, New Zealand

[3]The University of Auckland, Private Bag 92019, Auckland 1142, New Zealand

*Corresponding author

Email: swati.jivanji.1@uni.massey.ac.nz

3.1 Abstract

Background: White spotting of the coat is a characteristic trait of various domestic species including cattle and other mammals. It is a hallmark of Holstein-Friesian cattle, and several previous studies have detected genetic loci with major effects for white spotting in animals with Holstein-Friesian ancestry. Here, our aim was to better understand the underlying genetic and molecular mechanisms of white spotting, by conducting the largest mapping study for this trait in cattle, to date.

Results: Using imputed whole-genome sequence data, we conducted a genome-wide association analysis in 2,973 mixed-breed cows and bulls. Highly significant quantitative trait loci (QTL) were found on chromosomes 6 and 22, highlighting the well-established coat colour genes *KIT* and *MITF* as likely responsible for these effects. These results are in broad agreement with previous studies, although we also report a third significant QTL on chromosome 2 that appears to be novel. This signal maps immediately adjacent to the *PAX3* gene, which encodes a known transcription factor that controls *MITF* expression and is the causal locus for white spotting in horses. More detailed examination of these loci revealed a candidate causal mutation in *PAX3* (p.Thr424Met), and another candidate mutation (rs209784468) within a conserved element in intron 2 of *MITF* transcripts expressed in the skin. These analyses also revealed a mechanistic ambiguity at the chromosome 6 locus, where highly dispersed association signals suggested multiple or multiallelic QTL involving *KIT* and/or other genes in this region.

Conclusions: Our findings extend those of previous studies that reported *KIT* as a likely causal gene for white spotting, and report novel associations between candidate causal

mutations in both the *MITF* and *PAX3* genes. The sizes of the effects of these QTL are substantial, and could be used to select animals with darker, or conversely whiter, coats depending on the desired characteristics.

3.2 Background

Coat patterning traits provide visual characteristics that allow differentiation between domesticated animal breeds and between strains within breeds. White spotting is one of these phenotypes, and is a feature of a variety of mammals including cattle, horses, dogs, cats and mice. White spotting is a complex quantitative trait, for which several genes with major effects have been described and are relevant across species, as well as many other loci with small effects that account for the remaining genetic variance (1). This oligogenic architecture derives from the multifaceted biology that contributes to white spotting of the coat, which is hypothesised to arise from abnormal melanocyte precursor migration and/or development. Mouse models have demonstrated that pigment cells originate from the neural crest cells via the SOX10 positive glial bipotent progenitor cells during embryogenesis, and migrate dorsally via the neural tube (2). These cells proceed to differentiate into melanoblasts by acquiring expression of the genes *micropthalmia-associated transcription factor* (*MITF*), *proto-oncogene receptor tyrosine kinase* (*KIT*) and *dopachrome tautomerase* (*DCT*), and migrate down the ventral axis of the body. When the cells reach their destination, they migrate into the epidermis where some melanoblasts localise to the hair follicle and differentiate into melanocytes. A subset of melanoblasts dedifferentiate, losing *MITF* and *KIT* gene expression, and colonise the hair follicle bulge where they act as melanocyte stem cells

and replenish differentiated melanocytes during subsequent hair cycles (2). Disruption of any of the above processes is expected to result in parts of the body lacking mature melanocytes, and thus regions of abnormal pigmentation in the hair coat.

Quantitative trait loci (QTL) and mutations that cause white spotting have been described for a variety of species. Genetic studies in the horse revealed an inversion in the *KIT* gene associated with the Tobiano white-spotting (3), and a mutation in the *PAX3* gene associated with a splashed white pattern (4,5). Several mutations in the *KIT* gene have also been associated with complete white (6) or roan coat phenotypes (7). Studies on white spotting in dogs have revealed associations with the *MITF* gene (8), and in mice more than 10 genes have been reported to be associated with white spotting traits, including the *KIT* and *MITF* genes (9). Comparatively few studies have investigated the genetics of white spotting in cattle. Liu et al. (10) found significant QTL on chromosomes 6, 18 and 22 using linkage analysis within Holstein-Friesian (HF) × Jersey (J) crossbred cows. It has been suggested that the QTL on chromosomes 6 and 22 might be underpinned by the *KIT* and *MITF* genes, respectively (10). Fontanesi et al. (11) compared the sequences of the *MITF* gene in white spotted Italian Holstein and Simmental cattle, and solid coloured Italian Brown and Reggiana cattle, and found a haplotype (carrying allele g.31831615T) that is associated with white spotting. This haplotype accounts for some, but not all of the variation observed in the white spotting phenotype (11). More recently, Hofstetter et al. (12) investigated atypical white spotting in Brown Swiss cattle. They identified two completely linked single nucleotide variants within the 5' regulatory region of the *MITF* gene associated with white spotting, and although these variants largely account for the manifestation of

white spotting, they do not account for the variability between individuals, which provides further evidence for a polygenic trait (12). Hayes et al. (1) detected the *MITF* and *KIT* genes in a genome-wide association study (GWAS) that investigated the proportion of black in black and white Holstein cows, and reported an additional signal on chromosome 8, which carries *PAX5* i.e., another potential candidate gene for this trait (1). Together these studies converge on the involvement of *KIT* and *MITF* gene expression in white spotting in dairy cattle, however the causal variants that drive these effects have yet to be definitively identified and may be breed-specific.

Here, our aim was to investigate white spotting in New Zealand dairy cattle, by using whole-genome sequence genotype data to conduct the largest GWAS of white spotting to date. We report three genome-wide significant QTL for white spotting. Effects on chromosomes 6 and 22 extend on previous associations at these loci, and further implicate the *KIT* and *MITF* genes as responsible for these effects. For the first time, we also report a QTL on chromosome 2 that implicates the *PAX3* gene in white spotting of dairy cattle and highlight an amino acid substitution that may underlie this effect.

3.3 Methods

3.3.1 Study population

White spotting data were derived from several cohorts of animals that included: 885 outbred dairy bulls (223 J, 327 HF, and 335 HF×J), 1,389 outbred dairy cows (51 J, 265 HF, and 1,073 HF×J), and 699 HF×J F_2 cross cows from an experimental pedigree.

Breed definitions, in these cases, define animals from a 4-generation pedigree that were $^{16}/_{16}$ J or HF as purebreds, with $^{15}/_{16}$ animals defined as crossbreeds. The F_2 animals were ½ HF×½ J, representing a study population that was previously described in several publications (10,13–15). Genotyping data were available for 2,973 animals, with genotype and phenotype information derived as described in the following sections.

3.3.2 Measurements of white spotting in our study population

For animals in the F_2 population, proportion of white spotting values that had been derived for a previous study (10) were used directly in the current study. Video footage was recorded on 1,389 cows walking single file either into or out of the milking shed using a GoPro HERO4 camera, at a 4,000-pixel horizontal resolution. Still images that provide a clear side-on view of each animal were captured from the video footage using VideoPad Video Editor (v5.3). Additional side-on images representing either the right or left profile of 885 bulls were made available by LIC and incorporated into the dataset. First, cows and bulls were scored for the presence or absence of white on their coat and, then, the proportion of white spotting was quantified. Quantification was carried out manually using the image processing software, GNU Image Manipulation Program (GIMP, v2.9.8), to generate an objective measurement of the proportion of white colour. The freehand tool was used to trace each animal and remove the background. The pixel count from the remaining image, and the pixel count after manually subtracting the white regions on the coat, were used to calculate the proportion of white spotting on the coat.

3.3.3 Genotypes, whole-genome sequencing, and sequence imputation

For 760 of the outbred cows included in the study, tissue samples were obtained from ear tissue biopsies and DNA extraction and genotyping were performed by GeneSeek (Lincoln, NE, USA) using the GeneSeek GGP50k SNP chip. For all the remaining individuals, we used available single nucleotide polymorphism (SNP) genotypes that were previously obtained by genotyping at Geneseek on a variety of platforms including the Geneseek GGPv1, GGPv2, GGPv3, GGP50k, Illumina BovineSNP50 or BovineHD 777k SNP chips. A full list of the genotyping platforms, the number of SNPs per panel and the number of animals genotyped per panel are in Table 3.S1. Subsets of the reference and target populations that are described in this paper have been published by Lopdell et al. (16), and Littlejohn et al. (14,17).

Whole-genome sequencing, read mapping, and variant calling were performed on a population of 116 HF, 95 J and 354 crossbred cattle as previously described (16,17). Briefly, DNA samples were sequenced based on 100bp paired-end reads on the Illumina Hiseq platform, read mapping was performed using the UMD3.1 genome build and the BWA MEM (v0.7.8) software (18) and resulted in mean and median mapped read depths of 15× and 8×, respectively. Variants were called using the GATK HaplotypeCaller (v3.2) software (19), which incorporates base quality score recalibration. Then, phasing of the variants was performed using Beagle 4 (20), and variants with phasing allelic R^2 metrics lower than 0.95 were excluded for quality filtering purposes. These criteria yielded the ~19.5 million whole-genome sequence variants that constituted the reference set for imputation into the 2,973 SNP chip genotyped samples used for GWAS.

A step-wise imputation was performed using the Beagle 4 software (20). Note that these procedures were conducted to create an imputed sequence resource that is much larger than that used in the current study and represented ~150,000 animals, which have been accumulated over time and imputed in three different batches. The overall pipeline was as follows: first, the animals that were typed on the GGP panels were imputed to a reference panel representing the BovineSNP50 SNP-chip. Then, BovineSNP50 data (now consisting of both imputed and physically genotyped data) were used to impute all the animals to the BovineHD platform. We also conducted a parallel step to impute all the samples to the GGPv3 platform, to recover non-overlapping content between that platform and the BovineSNP50 SNP-chip. These steps yielded two datasets that comprised an 'all animals imputed to BovineHD' set, and an 'all animals imputed to GGPv3' set. These datasets were then merged, creating a scaffold for genome sequence imputation that contained all the animals imputed to all content from all SNP-chips. Following sequence imputation (by using Beagle 4), data were then filtered to remove variants with extreme Hardy-Weinberg statistics (HW exact test; removal of 47,660 variants based on $p < 1\times10^{-30}$), and near-monomorphic positions (minor allele frequency (MAF) < 0.0001; removal of 911,633 variants). These criteria yielded 18,641,995 variants, which were extracted for the subset of 2,973 animals with colour phenotypes from the larger ~150,000 animal dataset. In terms of genetic representativeness between the sequence reference animals and the 2,973 GWAS animals, 1,282 cattle were directly represented by both a sequenced sire and maternal grandsire in the reference dataset, of which 1,122 were represented by a sire or maternal grandsire in this population.

3.3.4 Population structure adjustments, covariates, and GWAS

To address population stratification in the association models due to breed and relatedness, genomic relationship matrices (GRM) were generated using GCTA (v1.91.1 beta). These calculations involved the creation of 29 GRM, one for each bovine autosome, to enable a 'leave one chromosome out' GWAS approach where each GRM differs by the absence of a single autosome – thus avoiding double fitting when testing the effect of candidate variants. These GRM were calculated using a curated subset of variants from the Illumina BovineSNP50 platform, which comprised 34,963 variants that had been quality-filtered based on Mendelian concordance parameters, minor allele frequency (those with a MAF < 0.02 were removed), LD pruning (those with a $R^2 > 0.9$ were removed), and deviation from Hardy Weinberg equilibrium (those with a $p < 0.15$ were removed). The GCTA (v1.91.1 beta) software was used to conduct the mixed linear model-based association analysis (MLMA), which incorporates the GRM as outlined above, in addition to fixed effects for farm of origin and cohort (the latter relevant to the F_2 animals with the first cohort born in spring 2000 and the second cohort born in spring 2001 (13–15,17,21)). Whole-genome sequence variants were filtered to remove the variants with a MAF lower than 0.005 prior to MLMA, this filter being different to that applied previously based on the frequencies present in the subpopulation of 2,973 animals. To account for multiple hypothesis testing, a *p*-value threshold of 5×10^{-8} was deemed to be significant for variant associations.

3.3.5 Visualization and interpretation of association results and candidate variants

To assess candidacy of the associated variants, RNA-seq data representing black and white bovine skin were sourced from a data submission accompanying the Koufariotis et al. (22) paper, and uploaded into the Integrative Genomics Viewer (IGV) for visualization (23). Sequence variants in intervals of interest were functionally annotated by using SNPEff (v4.3) (24) and the Ensembl UMD3.1 gene annotation set, with custom scripts to visualize these effects in Manhattan plots. To assess conservation metrics for candidate causal variants, genome evolutionary rate profiling (GERP) scores were obtained for the 32-way amniota vertebrae alignments (v92.31) from the Ensembl portal, with both element and site-wise scores reported in the text (25,26). For multiple protein alignments that were used to investigate the conservation of the *PAX3* p.Thr424Met mutation, *PAX3* homologues were retrieved for other species using BLAST, and aligned using the Geneious software (27).

3.3.6 Structural variant analysis

Sequence alignments representing the three major QTL regions were manually inspected in animals that displayed segregating tag-SNP genotypes to detect gene-disrupting structural mutations that might explain these QTL. However, given the ambiguity of the association signals at the chromosome 6 locus, a more formal analysis was conducted. Here, CNVnator (v0.3.3) (28) was used to predict the presence of structural variants based on sequence read depth, using the same whole-genome sequence dataset as described in the ‘Genotypes, whole-genome sequencing, and

sequence imputation' section. This analysis used a sliding window size of 1,000bp with a 500bp overlap and focused on a 20 Mb region on chromosome 6 (60 to 80 Mb). Then, predicted structural variants were ranked based on their genotype correlation with the top two QTL tag variants at the chromosome 6 locus (Chr6 g.64210286A>G rs451683615 and Chr6 g.71722665C>T rs463810013). Sequence alignments of relevant variants were visually inspected in IGV (23) to assess evidence of a legitimate structural variant at each of these sites, weighted in the context of read mapping quality, gaps and/or other issues with the reference genome assembly, and whether the variant was polymorphic between samples. CNVnator-assigned genotypes were assessed in the same way for multimodality by visual inspection of copy number histograms.

3.4 Results

Since white spotting might be influenced by genes that operate via different mechanisms, we conducted two separate GWAS that differed in the definition of the phenotype. First, white spotting was scored as the presence or absence of white on the coat and encoded as a binary phenotype (N=2,973 animals). Second, white spotting was coded as a quantitative variable, where animals were scored based on the overall proportion of white (N=2,232 animals). Solid coloured animals were not included in the latter population, for which proportion of white was also log-transformed prior to association analysis to render data in a form approximating a normal distribution. All phenotypic measures were based on manual analysis of photographs (see Methods section), that included images representing 699 Holstein-Friesian × Jersey (HF×J) F_2 cows scored as part of a previous QTL study (10). The breed composition and sexes of

the remaining animals are described in the Methods section, which include a mixture of HF, J, and HF×J cows and bulls.

Genome-wide association analysis was conducted based on imputed whole-genome sequence genotypes using the GCTA (v1.91.6) software. Genotypes were imputed to sequence resolution using a reference population of 565 whole-genome-sequenced animals and methods that are similar to those described previously (see Methods section and Lopdell et al. (16)). The mixed linear models assumed additivity and incorporated adjustments for farm of origin, cohort (10), and a genomic relationship matrix (GRM) computed in GCTA (v1.91.6). Imputed data were also filtered to remove variants that had a MAF lower than 0.005 and met other quality filtering criteria described in more detail in the Methods section. Results of the association analysis for presence/absence of white on the coat revealed three signals that surpassed the genome-wide significance threshold of $p=5\times10^{-8}$ and were located on chromosomes 2, 6, and 22 (Fig 3.1a). The top variants for these QTL mapped to Chr22 g.31769747A>G (rs209784468, $p=1.51\times10^{-56}$), Chr6 g.64210286A>G (rs451683615, $p=3.73\times10^{-53}$), and Chr2 g.111576221A>C (rs109979909, $p=3.26\times10^{-15}$).

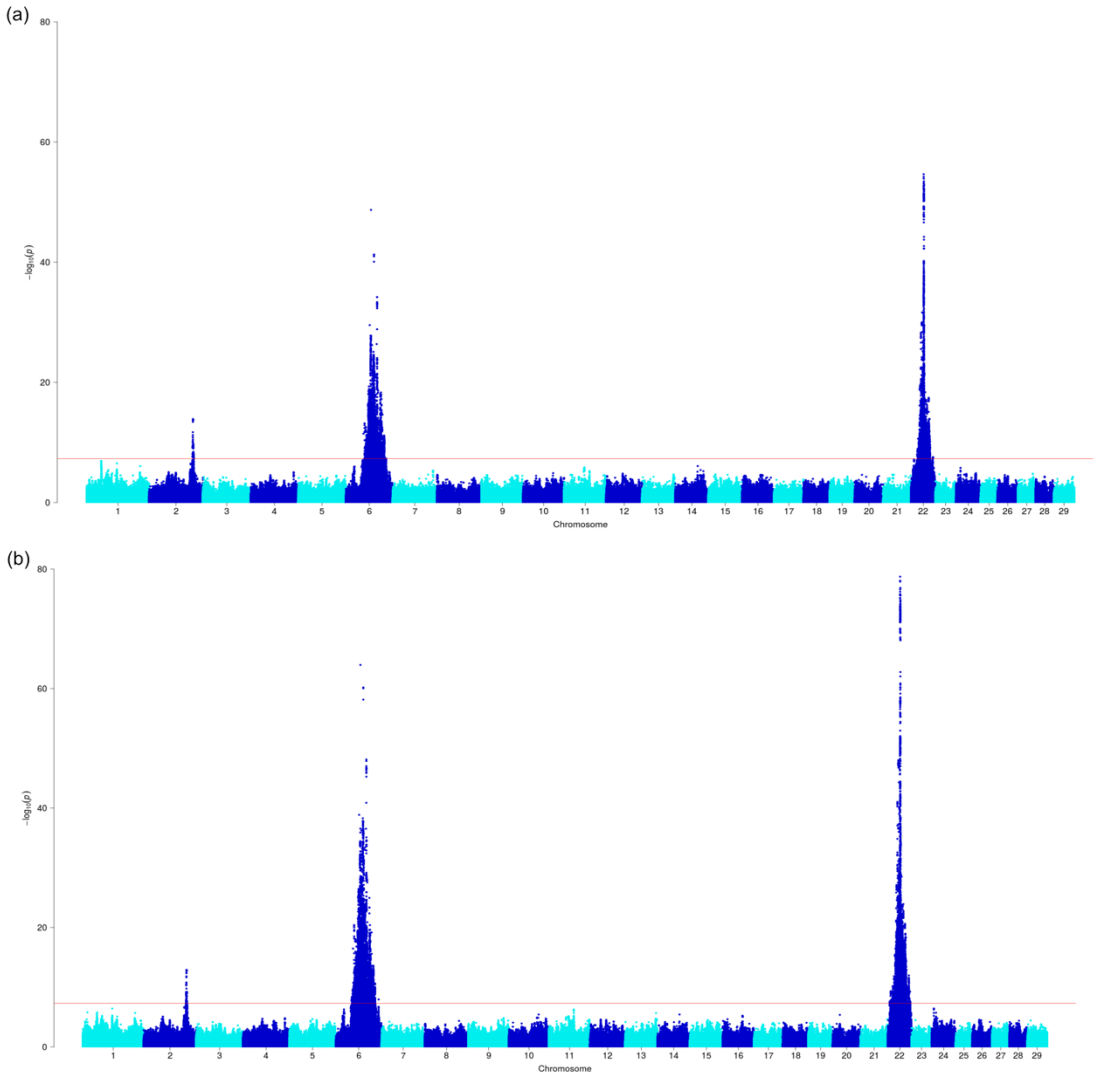


Figure 3.1 (a) Manhattan plot based on the GWAS results for the presence/absence of white colour on the coat. The top variants on chromosome 22, 6 and 2 have p-values of 1.51×10^{-56}, 3.73×10^{-53} and 3.26×10^{-15}, respectively. (b) Manhattan plot based on the GWAS results for the proportion of white spotting. The top variants on chromosome 22, 6 and 2 have p-values of 1.83×10^{-79}, 1.1×10^{-64} and 1.27×10^{-13}, respectively. The red line indicates the genome-wide significance threshold $p=5\times10^{-8}$.

For the analysis that treated white spotting as a quantitative variable (proportion of white spotting), GWAS revealed the same three regions as those described for the binary-encoded trait ($p < 5\times10^{-8}$; Fig 3.1b). Furthermore, this analysis presented the same three top-associated variants that were identified in the first GWAS, which suggested that these signals represented the same QTL. These results are in agreement with previous findings that described white spotting as a quantitative trait, i.e., under the

control of multiple QTL (1,11). Given that the signals derived from the quantitative phenotype were also more significant than for the binary trait, this phenotype became the focus of the analyses that are presented below. Table 3.1 lists the top 10 associated variants and the effects of these QTL. Notably, the sizes of the effects of all three QTL were very large, with allele substitution effects of 3.2, 12.9, and 11.5% for the top tag SNPs on chromosomes 2, 6, and 22, respectively.

Table 3.1 Top 10 variants for each significant quantitative trait locus detected in the genome-wide association analysis for proportion of white spotting

	Variant Reference ID	Genomic position	Effect size (%)*	Standard error	*p*-value
		Chromosome 22			
1	rs209784468	Chr22 g.31769747A>G	11.52	0.129	$1.83x10^{-79}$
2	rs461193589	Chr22 g.31783093T>C	11.35	0.129	$8.67x10^{-79}$
3	rs456585934	Chr22 g.31888569A>G	11.43	0.13	$1.21x10^{-78}$
4	rs209274730	Chr22 g.32386542A>C	11.06	0.129	$1.38x10^{-77}$
5	rs480312583	Chr22 g.31958551G>A	11.2	0.13	$2.47x10^{-77}$
6	NA	Chr22 g.31768931A>T	10.88	0.129	$6.11x10^{-77}$
7	rs208958980	Chr22 g.31769772T>C	10.84	0.129	$1.59x10^{-76}$
8	rs433645096	Chr22 g.31768933A>T	10.84	0.129	$1.59x10^{-76}$
9	NA	Chr22 g.31768928TG>T	10.82	0.129	$2.29x10^{-76}$
10	rs209837244	Chr22 g.32369667G>A	10.94	0.129	$2.57x10^{-76}$
		Chromosome 6			
1	rs451683615	Chr6 g.64210286A>G	12.86	0.15	$1.10x10^{-64}$
2	rs463810013	Chr6 g.71722665C>T	12.27	0.152	$6.37x10^{-61}$
3	rs109512689	Chr6 g.71873479T>C	12.02	0.151	$8.08x10^{-61}$
4	rs385773341	Chr6 g.71873455A>C	12.02	0.151	$8.08x10^{-61}$
5	rs474403670	Chr6 g.71698814A>G	12.22	0.152	$8.99x10^{-61}$
6	rs208251862	Chr6 g.71692344C>A	10.62	0.146	$7.05x10^{-59}$
7	rs43469863	Chr6 g.79629052T>C	7.76	0.139	$7.47x10^{-49}$
8	rs43469866	Chr6 g.79631054T>C	7.69	0.139	$1.34x10^{-48}$
9	rs43764915	Chr6 g.79649488A>G	7.54	0.139	$9.51x10^{-48}$
10	rs208257925	Chr6 g.79640038G>A	7.48	0.139	$1.90x10^{-47}$
		Chromosome 2			
1	rs109979909	Chr2 g.111576221A>C	3.19	0.157	$1.27x10^{-13}$
2	NA	Chr2 g.111588505GA>G	3.19	0.157	$1.40x10^{-13}$

3	rs379031581	Chr2 g.111587292A>G	3.19	0.157	1.40×10^{-13}
4	rs385337886	Chr2 g.111573853A>G	3.19	0.157	1.40×10^{-13}
5	rs468881264	Chr2 g.111615661G>A	3.19	0.157	1.40×10^{-13}
6	NA	Chr2 g.111601410A>G	3.18	0.156	1.41×10^{-13}
7	rs381689348	Chr2 g.111604662A>C	3.18	0.156	1.41×10^{-13}
8	rs377769439	Chr2 g.111634835G>A	3.18	0.157	1.55×10^{-13}
9	rs385963805	Chr2 g.111570788G>A	3.18	0.157	1.55×10^{-13}
10	rs380782402	Chr2 g.111560710G>A	3.17	0.156	1.58×10^{-13}

*Effect size is expressed as the percentage of white on the coat attributed to each additional 'Q' allele.

3.4.1 Analysis of the significant loci on each detected chromosome

3.4.1.1 Chromosome 22

A SNP at Chr22 g.31769747A>G (rs209784468) was identified as the most significant variant in our association analysis ($p=1.83\times10^{-79}$), and mapped to a region 284bp upstream of the Ensembl-annotated transcription start site (TSS) of the *MITF* gene. The MITF transcription factor is involved in melanocyte survival, maintenance and differentiation (29), and is therefore the most obvious candidate at this locus. Based on the Ensembl v92.31 gene build (25,26), *MITF* is also one of the only two annotated protein-coding genes that are present within a 1 Mb window around rs209784468, which provides strong support for the causative status of this gene. Figure 3.2a shows a Manhattan plot of this interval, with the variants being colour-coded according to predicted functional impact using SNPEff (24). To assess whether the signal observed on chromosome 22 was likely representative of a single biallelic QTL, we ran an additional analysis, by fitting the top-associated SNP (rs209784468) as a fixed effect in the association model. This analysis removed significance at almost all the variants within a 1 Mb interval (Fig 3.2a and 3.2b), but a slight residual signal remained (smallest $p=8.53\times10^{-11}$ for Chr22 g.31730376 rs109549448; Fig 3.2b). Although

imputation error or unaddressed population stratification might explain the small residual signal revealed in this analysis, the well-described allelic heterogeneity for *MITF* supports the potential existence of multiple and/or multiallelic QTL. It should be noted that, in a recent analysis in Brown Swiss cattle, Hofstetter et al. (12) identified a SNP (rs722765315), located within the 5' region of the *MITF* gene as a candidate causal variant for white spotting (12). However, examination of this site in our whole-genome sequenced cohort shows that it is invariant in Holstein-Friesian and Jersey animals, which suggests the presence of one or more alternate causal variants in the New Zealand population.

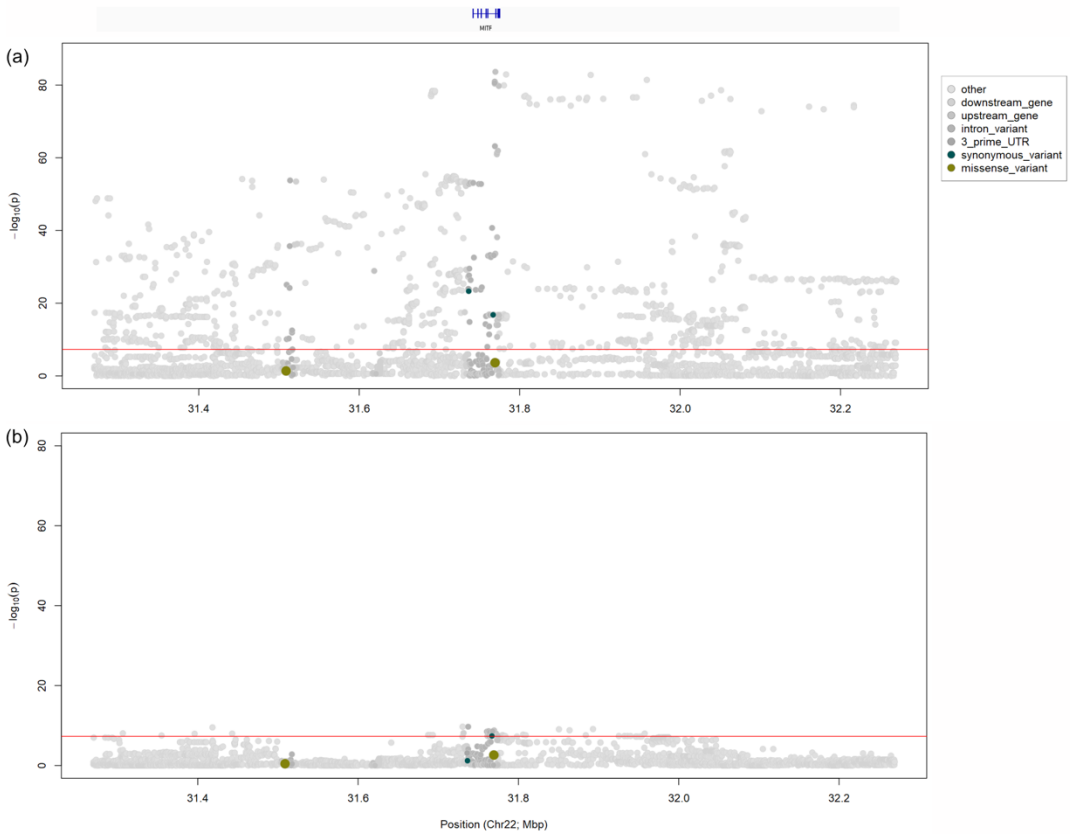


Figure 3.2 QTL analysis of chromosome 22 with variants colour-coded according to predicted functional impact using SNPEff. (a) 1 Mb window of imputed whole-genome sequence association data centred around the top variant Chr22 g.31769747A>G (rs209784468) with the corresponding annotated gene track above. (b) 1 Mb window of imputed sequence association data with rs209784468 fitted as a fixed effect in the association model. The red line indicates the genome-wide significance threshold $p=5\times10^{-8}$.

A novel, polymorphic MITF pseudogene as a candidate for the white spotting QTL. Notably, we observed a predicted missense mutation that affects *MITF* at Chr22 g.31769331C>T (rs110881545; Fig 3.2a). Although it could be a candidate mutation for the QTL, this variant was not significant, and was called at a very low frequency in the genome sequence reference population used for imputation (MAF < 0.01). Manual inspection of sequence alignments from animals heterozygous for this variant showed read depth anomalies around annotated intron-exon boundaries, which led us to analyse in more detail these features. Although we used DNA-based sequence data, at these boundaries we observed an increased sequencing depth for the exons, which are reminiscent of RNA-sequence alignments (Fig 3.S1). Analysis of soft-clipped reads from the exons showed that the mismatches corresponded to neighbouring exon structures, which suggest that they were derived from a mis-mapped, processed *MITF* pseudogene. Non-exonic read pairs from the apparent *MITF* pseudogene mapped to a single location on chromosome 12 at 58.7 Mb, indicating that this locus is the likely site of integration of the pseudogene. Notably, this pseudogene was polymorphic across animals, which raised the possibility that the QTL might be caused by this structural variant. String match searching for spliced *MITF* sequence reads from the whole-genome sequence alignments, allowed us to genotype the 565 whole-genome-sequenced animals in our reference population for the pseudogene, giving a MAF of 0.026 for the integrated allele. This MAF value contrasted markedly with that of the top tag variant from GWAS (MAF = 0.304); and when pairwise linkage disequilibrium statistics were examined between the pseudogene 'genotype' and variants from the broader chromosome 22 and chromosome 12 regions, the most highly correlated markers were also non-significant in the GWAS (chromosome 12, maximum R^2=0.72 for rs461882713 Chr12:6060748C>G, p=0.72; chromosome 22, maximum R^2=0.69 for

rs384283283 Chr22:31734120C>T, p=0.67). Although the processed *MITF* pseudogene was a good biological candidate for the modulation of coat colour or pattern, these observations led us to assume that it was not responsible for the white spotting QTL in our study.

Evolutionarily conserved, candidate causative regulatory variants at the MITF locus. Apart from the *MITF* pseudogene identified above, no other protein-coding changes were identified in *MITF* that could explain this QTL. Although two synonymous *MITF* variants exceeded genome-wide significance, their association was sufficiently weak to discard them as underpinning the QTL (Fig 3.2a). Together, these observations suggested an expression-based mechanism for a MITF-derived effect on white spotting. The top associated variant Chr22 g.31769747A>G (rs209784468) is a reasonable candidate in this regard, as it maps to a region immediately upstream of the annotated transcription start site (TSS). However, inspection of RNA-sequence (RNA-seq) data for black and white bovine skin samples published by Koufariotis et al. (22) showed alternative gene structures that include additional 5' exons to the Ensembl-derived annotation (MITF-201; Ensembl v92.31), in which the rs209784468 variant mapped to intron 2 of the two predominant RNA-seq derived structures (Fig 3.3). Similarly, examination of the transcripts annotated on the newest version of the bovine reference assembly at the time of the preparation of this paper (ARS-UCD1.2) showed alternative *MITF* structures, for which the skin-derived transcripts were best represented by the MITF-205 and MITF-206 transcripts (Ensembl v96.12). Notably, 18 additional variants that displayed association statistics that were broadly similar to those of rs209784468 ($p < 5\times10^{-70}$) also mapped within introns 1, 2, 3, and up to 100 kb upstream of the

alternate *MITF* isoforms. To further investigate these variants, we downloaded genome evolutionary rate profiling (GERP) scores from the Ensembl portal to assess conservation metrics of the sites (Table 3.2) (25,26). Although the location of this variant was less appealing than some of the others that map closer to the assumed 5' *MITF* promoter, the top-associated SNP is the only variant that mapped to a conserved element identified from the 32-way amniote vertebrate alignments (Fig 3.3). This SNP is also conserved on a site-wise basis (GERP score = 1.21), and based on its association ranking, it constitutes a plausible candidate regulatory mutation for this QTL.

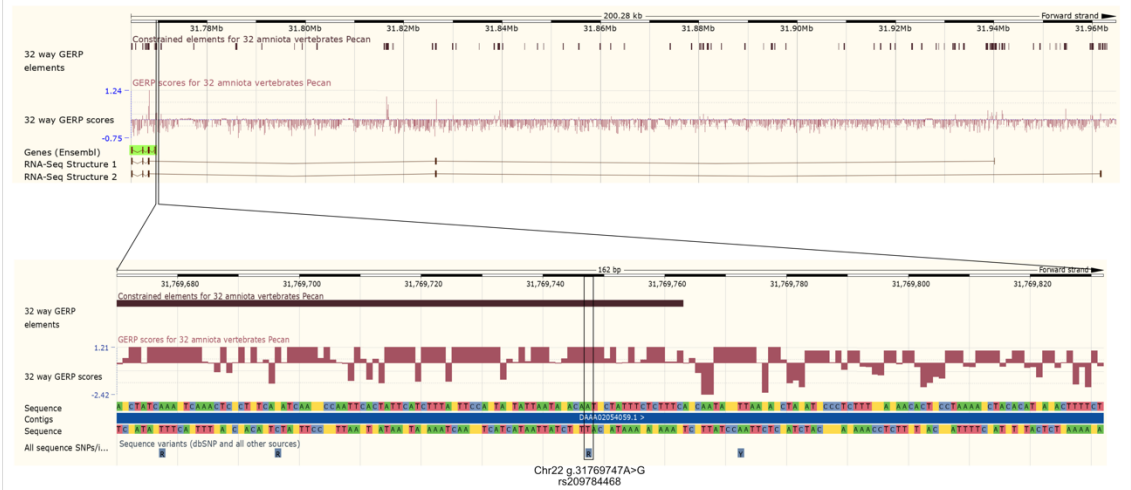


Figure 3.3 Detailed view of introns 1 to 3 of the Ensembl-derived MITF gene structure and introns 1 to 5 of the RNA-seq derived MITF structures, with constrained elements and GERP score for 32 amniota vertebrates from Ensembl (Bos taurus v92.31). g.31769747A>G (rs209784468) is highlighted and located to a highly conserved region within intron 2 of the RNA-seq derived MITF gene structures.

Table 3.2 Top variants mapping within introns 1, 2, 3 and up to 100 kb upstream of the annotated MITF TSS, with conservation (GERP) score for 32 amniota vertebrates (Ensembl Bos taurus v92.31 - UMD3.1)

Variant Reference ID	Genomic Position on Ch22	GERP Score	Constrained Element	P-value
rs209784468	g.31769747A>G	1.21	Yes	1.83×10^{-79}
rs461193589	g.31783093T>C	0.07	No	8.67×10^{-79}
NA	g.31768931A>T	-1.85	No	6.11×10^{-77}
rs433645096	g.31768933A>T	0.92	No	1.59×10^{-76}

rs208958980	g.31769772T>C	1.21	No	$1.59x10^{-76}$
NA	g.31768928TG>T	-0.22	No	$2.29x10^{-76}$
rs385179918	g.31780393C>A	0	No	$6.69x10^{-76}$
rs110372927	g.31774043C>T	-1.52	No	$7.97x10^{-76}$
rs384965533	g.31807384A>G	0.2	No	$8.11x10^{-73}$
rs109143893	g.31805754C>T	-1.25	No	$1.28x10^{-72}$
rs385825679	g.31811182C>T	0	No	$2.82x10^{-72}$
rs209226877	g.31873774A>C	0.65	No	$4.06x10^{-72}$
rs109756444	g.31853470A>G	-1.69	No	$5.25x10^{-72}$
rs378395938	g.31838217G>A	-0.08	No	$6.63x10^{-72}$
rs110467669	g.31849617A>G	-1.69	No	$6.63x10^{-72}$
rs110989002	g.31812468A>T	-0.09	No	$8.75x10-^{71}$
rs110743578	g.31821264C>G	-1.69	No	$1.54x10^{-70}$
rs110276495	g.31863698C>T	-0.4	No	$2.59x10^{-70}$

3.4.1.2 Chromosome 6

The top variant at the chromosome 6 locus (Chr6 g.64210286A>G rs451683615, p=1.1×10^{-64}), maps to an intergenic region approximately 280 kb downstream of the *KCTD8* gene, which represents quite a considerable distance from the *KIT* gene (~7.5 Mb). However, the third and fourth most strongly associated variants map within the fourth intron of *KIT* (Chr6 g.71873479T>C rs109512689, p=8.08×10^{-61} and Chr6 g.71873455A>C rs385773341, p=8.08×10^{-61}).

Given the dispersion of the chromosome 6 signal, and the association of variants that are located within and adjacent to the strong *a priori* candidate gene *KIT*, we considered a large interval (16 Mb) around the top variant rs451683615 for functional prediction of variant effects. The following genes map to this interval: *C6H4orf19, TBC1D1, KLF3, TMEM156, KLB, UBE2K, RHOH, RBM47, APBB2, UCHL1, BEND4, SHISA3, HTATSF1, KCTD8, YIPF7, GABRG1, GABRA2, MGC127695, GABRB1, ATPq0D, NFXL1, TXK, SLAIN2, OCIAD1, LRRC66, USP46, SCFD2, FIP1L1, UFM1, GSX2,*

KIT, KDR, SRD5A3, PDCL2 and *CREP135*. Figure 3.4 shows a Manhattan plot for this region, with variants colour-coded according to predicted functional impact using SNPEff. Based on association statistics, none of the variants in the top 10 orders of magnitude are predicted to change the protein-coding sequence of these genes, although there is a modestly associated splice region variant in *KIT* (Chr6 g.71906518T>C rs109750754, $p=1.94\times10^{-23}$). Given that the primary signals highlight non-coding variants, a QTL mechanism that incorporates one or more gene expression-based effects seems most likely.

Multiple segregating QTL at the KIT locus. One explanation for the dispersed nature of the chromosome 6 QTL is that this locus comprises multiple, overlapping effects. Linkage disequilibrium (LD) analysis between the top variant (Chr6 g.64210286A>G rs451683615) and the next three most strongly associated variants (Chr6 g.71722665C>T rs463810013, Chr6 g.71873479T>C rs109512689 and Chr6 g.71873455A>C rs385773341) supports this hypothesis, with rs451683615 being in relatively low LD with the other variants (maximum $R^2=0.35$). Furthermore, when rs451683615 was fitted as a fixed effect, the signal on chromosome 6 still exceeded the genome-wide significance threshold ($p=5\times10^{-8}$), with the two strongly correlated *KIT* variants ($R^2=0.91$) rs208251862 (Chr6 g.71692344C>A; $p=7.1\times10^{-19}$) and rs463810013 ($p=1.5\times10^{-18}$) now being the top variants (Fig 3.4b). When the rs463810013 variant was fitted as a fixed effect to represent these effects, rs451683615 once again became the most significant variant ($p=3.054\times10^{-25}$; Fig 3.4c), and when both rs451683615 and rs463810013 were fitted as fixed effects, a small signal was still detected near *KIT* (smallest $p=3.31\times10^{-11}$ for Chr6 g.72007252A>T rs109258078; Fig 3.4d). These results

suggest that the signal observed on chromosome 6 is likely the result of two or more QTL, and/or alternatively, the consequence of one or more structural variants that are not well tagged, and therefore cannot be easily accounted for by fitting biallelic SNPs in the association models.

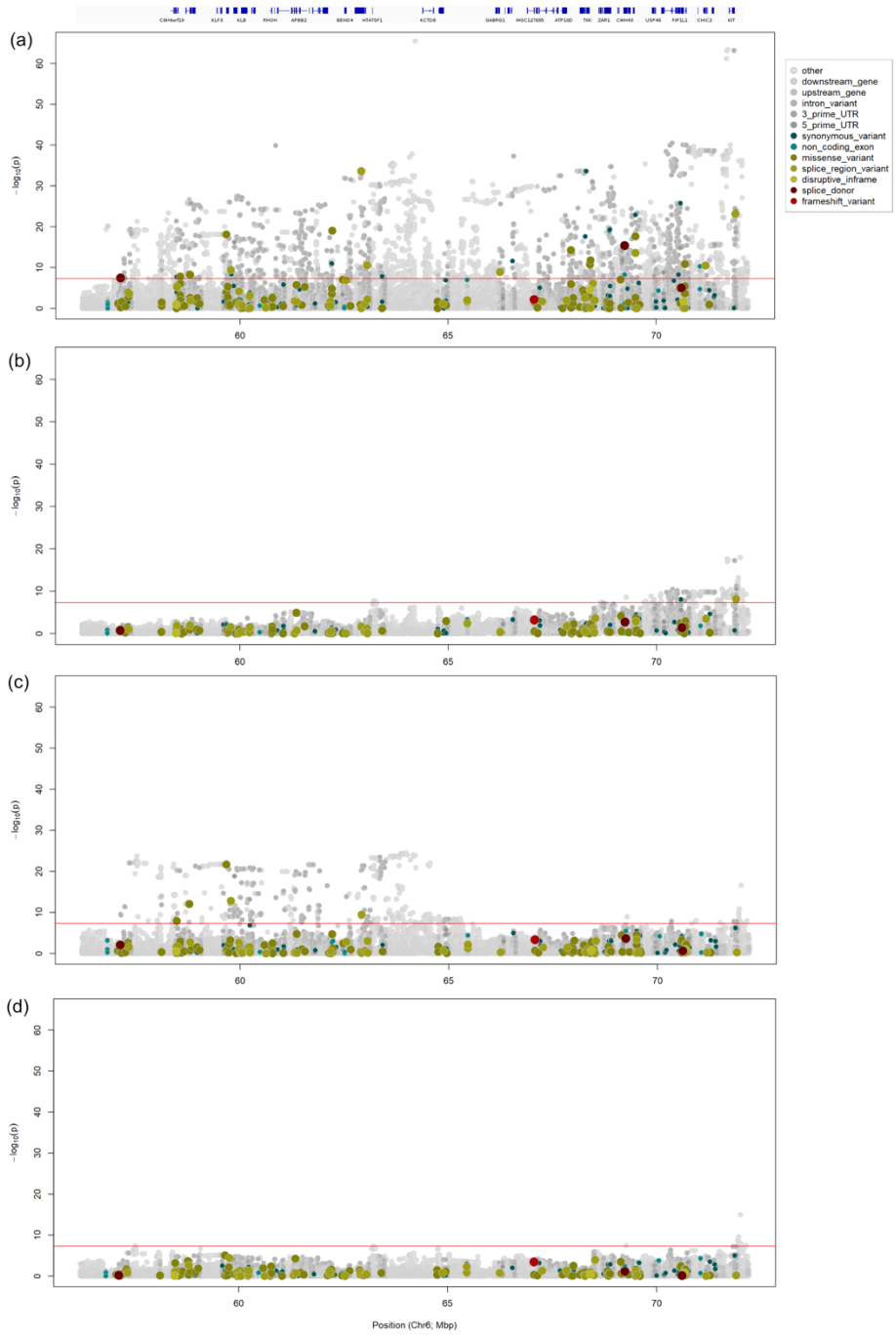


Figure 3.4 QTL analysis of chromosome 6 with variants colour-coded according to predicted functional impact using SNPEff. (a) 16 Mb window of imputed whole-genome sequence association data centered around the top variant Chr6 g.64210286A>G (rs451683615) with the corresponding annotated gene track above. (b) 16 Mb window of

imputed whole-genome sequence association data with rs451683615, (c) Chr6 g.71722665C>T (rs463810013) and (d) both rs451683615 and rs463810013 fitted as fixed effects. The red line indicates the genome-wide significance threshold $p=5\times10^{-8}$.

Structural variant analysis at the chromosome 6 locus. Given the ambiguity of the association signals at the chromosome 6 locus, and the implication of *KIT* structural variants that have a role in other coat characters in cattle (e.g. white face piebaldism in Hereford (30) and colour-sidedness in Belgian Blue, Brown Swiss, and other breeds (31)), we performed a sequence-based structural variant analysis to attempt to identify segregating candidate mutations for this QTL. This analysis was conducted using the same population of 565 whole-genome-sequenced animals as that used for sequence imputation prior to GWAS, and we focused on a broad 20 Mb region (60 to 80 Mb) to capture the dispersed nature of the association peak. This region included the top 10 variants shown in Table 3.1, for which the CNVnator software (28) was used to call structural variants within the interval based on a 1 kb sliding window approach with 500bp overlaps (see Methods). This analysis revealed a large number of candidate polymorphic intervals (N=39,960). We used correlation analysis between estimated copy numbers and genotypes from the top two associated GWAS variants (Chr6 g.64210286A>G rs451683615 and Chr6 g.71722665C>T rs463810013) to prioritize the variants for subsequent investigations. Of the top 10 most highly correlated variants for each of the two tag SNPs, these intervals represented six discrete structural variants (and some variants could be merged because they spanned adjacent intervals). Table 3.3 shows the position, mutation-type, and LD correlation coefficient of these six variants for the tag SNP of interest, with LD values based on re-calling of the intervals following merging and manual boundary refinement. Visualization of sequence alignments

suggested legitimate polymorphic structural variation for all six variants, with four of these showing clear multi-modality in read depth (Fig 3.S2). Notably, LD analysis between the six structural variants and the 124,445 other sequence variants within the 20 Mb chromosome 6 interval showed that five of the six variants were better tagged by other chromosome 6 polymorphisms, which all showed limited phenotypic association by comparison with the top associated tag SNPs rs451683615 and rs463810013 (Table 3.3). One exception was a 330bp duplication at Chr6:72,060,120-72,060,450bp, where this variant was best tagged by a SNP that is largely equivalent to rs463810013 (rs385773341 Chr6 g.71873455A>C; R^2=0.98 with rs463810013; Table 3.3). None of the six structural candidates mapped to protein coding sequences, although the apparent 330bp duplication was also the polymorphism nearest to *KIT* (albeit 142 kb downstream). Assessment of the potential function for this variant did not present any obvious regulatory implication, since the duplication was devoid of noteworthy site-wise conservation or GERP-annotated constrained elements. Acknowledging the fact that our read-depth-based analysis of structural variation may represent the complexity of the identified candidate mutations, these data likely exclude five of six of the structural variants as candidates for the white spotting QTL. The potential role of the sixth candidate variant is unknown, and although the duplication was best represented by the top GWAS tag variants, its overall correlation was still low (maximum R^2=0.43). This observation, and the fact that copy number genotypes were not clearly differentiated for this variant (Fig 3.S2), lead us to suggest that physical genotyping and more detailed investigation will be required to further assess the nature and candidacy of this polymorphism.

Table 3.3 Description and LD summary statistics for the candidate structural variants that are most highly correlated with tag SNPs rs451683615 (Chr6 g.64210286A>G) and rs463810013 (Chr6 g.71722665C>T)

Region spanning CNV	Type	rs451683615 correlation (R^2)	rs46381013 correlation (R^2)	Closest gene	Maximum R^2	SNP ID	GWAS p-value
Chr6:64,092,201-64,092,752bp	Deletion	0.172	0.099	KCTD8	0.544	rs110545184	$3.24x10^{-22}$
Chr6:65,557,508-65,559,004bp	Deletion	0.102	0.066	GNPDA2	0.876	rs384078363	$3.74x10^{-5}$
Chr6:65,657,051-65,657,595bp	Deletion	0.128	0.089	GNPDA3	0.746	rs383024906	$2.79x10^{-11}$
Chr6:68,269,498-68,270,804bp	Deletion	0.171	0.164	NFXL1	0.569	rs456305543	$5.89x10^{-34}$
Chr6:71,310,834-71,312,202bp	Deletion	0.065	0.163	GSX2	0.695	rs466525306	$4.78x10^{-12}$
Chr6:72,060,120-72,060,450bp	Duplication	0.22	0.431	KIT	0.432	rs385773341	$8.08x10^{-61}$

CNV: copy number variant
R^2: linkage disequilibrium correlation coefficient
SNP ID: single nucleotide polymorphism accession number

3.4.1.3 Chromosome 2

The top variant at the chromosome 2 locus (Chr2 g.111576221A>C rs109979909, p=1.27×10^{-13}) maps to intron 1 of the *FARSB* gene. Considering all the genes in a 1 Mb interval centred on rs109979909, the *PAX3, MIR2284Y-5, FARSB, LOC538702, MOGAT1, ACSL3, RPSL3, RPS6* and *KCNEE4* genes map to this region. In particular, *PAX3* is a striking candidate, since it encodes a *MITF* transcription factor (see 'Chromosome 22' section above) and was proposed as a causal gene for the 'splashed white' coat phenotype in horses (4). Variant effect prediction for all variants in the 1 Mb interval (Chr2:111,076,221-112,076,221bp) revealed a candidate causal missense

mutation in *PAX3*, that codes for a threonine to methionine substitution at amino acid position 424 (rs208582518; p.Thr424Met; Fig 3.5; (32)). Although the p.Thr424Met variant shows a comparatively weaker association than the top associated variant at this locus (p=2.72×10^{-11} versus smallest p=1.27×10^{-13}), it is sufficiently strongly associated to remain a compelling candidate mutation for the QTL. Additional inspection of the sequence alignments across the 1 Mb region centred on rs109979909 did not show any evidence of segregating structural variants as alternative candidates at this locus.

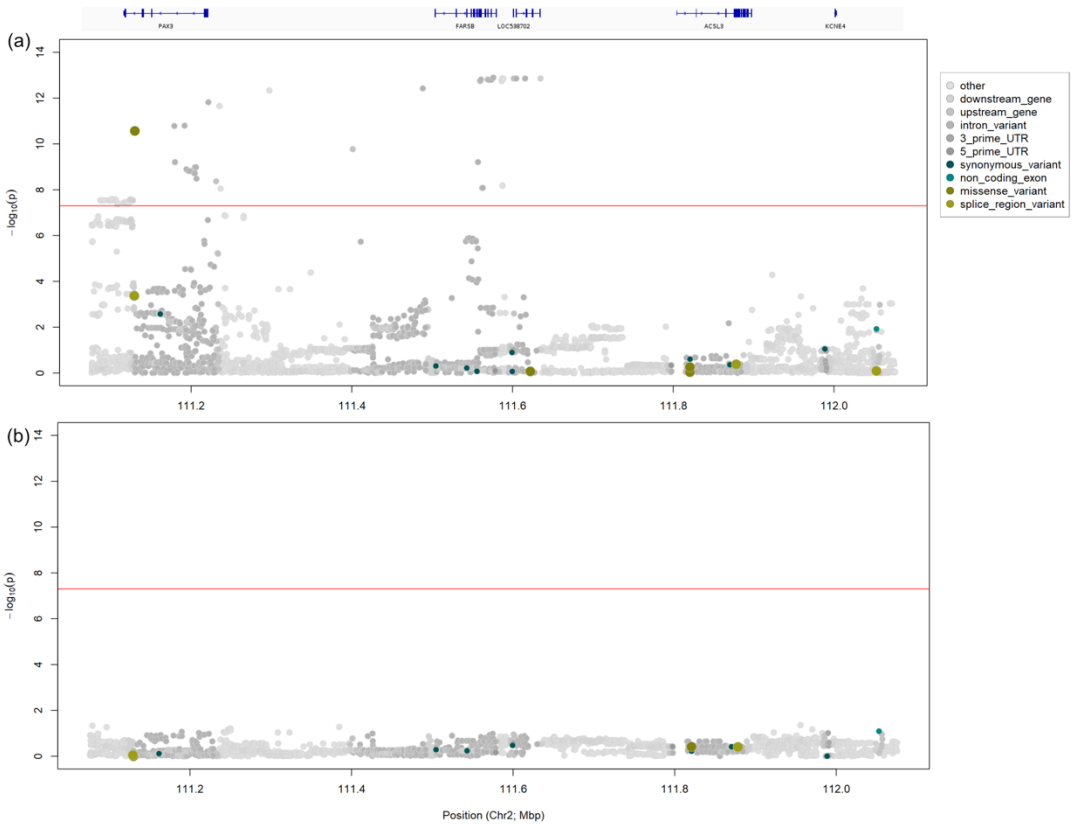


Figure 3.5 QTL analysis of chromosome 2 with variants colour-coded according to predicted functional impact using SNPEff. (a) 1 Mb window of imputed whole genome sequence association data centred around top variant Chr2 g.111576221A>C (rs109979909) with the corresponding annotated gene track above. (b) 1 Mb window of imputed whole-genome sequence association data with rs109979909 fitted as a fixed effect. The red line indicates the genome-wide significance threshold p=5×10^{-8}.

A novel candidate causal PAX3 missense mutation. The p.Thr424Met (rs208582518) variant maps to exon 9 of the *PAX3* gene. When fitted as a fixed effect in the association model, the variant accounted for the majority of the signal at this locus (smallest p=0.0149 for Chr2 g.111955758G>A rs41718011 for this model; Fig 3.5b). The p.Thr424Met variant is located within the transactivating domain of the PAX3 transcription factor, which is also identified as a constrained element from the GERP 32-way amniote alignments. The variant has a site-wise GERP score of 1.72, and when assessing the predicted functional impact of the missense variant using the SIFT algorithm (33) that is integrated as part of the Ensembl Variant Effect Predictor (32), this SNP is predicted to be 'deleterious' (score 0.01, low confidence). Likewise, p.Thr424Met is predicted to be 'possibly damaging' (score 0.86) by the Polyphen-2 functional prediction tool (34,35), and multiple alignment of PAX3 protein sequences representing a range of vertebrates also shows conservation of the threonine residue and surrounding amino acid acids in mammals (Fig 3.6). Overall, the *PAX3* p.Thr424Met missense mutation is a compelling candidate causal mutation for the white spotting phenotype, although the strong association of other non-coding variants leaves open the possibility of expression-based effects, which again operate most likely through the *PAX3* gene.

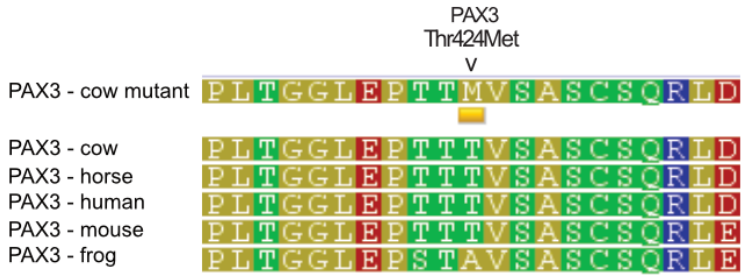


Figure 3.6 Region around the p.Thr424Met mutation. Wild-type threonine at position 424 is conserved across cow (Bos taurus), horse (Equus caballus), human (Homo sapiens) and mouse (Mus musculus) PAX3 orthologues.

3.4.2 Breed, frequency, and effect size characteristics of the three major QTL

White spotting is a characteristic trait in HF and has been under selection for many generations. Although some J animals in New Zealand show white spotting, it is far less frequent in this breed. Thus, we expect that the alleles that are associated with a greater proportion of white spotting are more frequent in HF. Based on the allele frequencies of the top tag variants for each of the three major QTL in 589 purebred HF, and 274 purebreed J, we obtained the frequencies shown in Table 3.4 (see Methods: Study populations for breed definitions). Here, we have denoted the white-increasing allele as *Q*, and the white decreasing allele as *q*.

Table 3.4 Q allele frequencies for the top variant at each QTL for 589 purebred Holstein-Friesians and 274 purebred Jerseys

Genomic position	Variant reference ID	*q* allele	*Q* allele	HF *Q* frequency	J *Q* frequency
Chr22 g.31769747A>G	rs209784468	G	A	0.97538	0.3431
Chr6 g.64210286A>G	rs451683615	G	A	0.99236	0.6332
Chr2 g.111576221A>C	rs109979909	C	A	0.98557	0.6953

Given the large sizes of the effect of the QTL, it is interesting to examine how '*Q*' (more white) or conversely '*q*' (less white) alleles might combine across loci to impact the phenotype. To investigate the QTL in this way, 'stacked' genotypes were derived for each animal based on the top-associated tag variants representing the chromosome 2, 6, and 22 loci. In this way, animals could be categorized based on the number of '*Q*'

alleles presented (possible range from 0 to 6). This analysis focused on a subset of 699 F_2 cows (½HF×½J) to minimize possible confounding by admixture, where animals were also all derived from the same research group. The smallest number of *Q* alleles carried in this population was two ('2*Q*'; N=10 cows), none of these cows displaying visible white colour on their coat (based on pictures that show only a single side view). By comparison, animals that carry six *Q* alleles (i.e., homozygous *Q* for all three loci; '6*Q*'; N=160) displayed a striking increase in white spotting. Figure 3.7 compares the 10 2*Q* animals (left panel) with a random selection of 10 6*Q* animals (right panel), and highlights the major impact of these QTL. The mean percentage of white spotting value was 0% for the 10 2*Q* animals and 32.6% for the 160 6*Q* animals (or 36.9% for the subset of 10 6*Q* animals shown in Fig 3.7). These observations give some clue as to the somewhat counterintuitive finding that *Q* alleles for two of the three QTL are the major alleles in J animals. Although this breed is best known for its solid, light brown coat, in F_2 animals, only those with a large number of *Q* alleles showed substantial proportions of white spotting on their coat. Figure 3.S3 and Table 3.S2 also show a breakdown of the *Q* allele counts in purebreds, based on the 589 HF and 274 J animals referenced above. In this purebred dataset, the percentage of J animals with 6*Q* alleles is only 1.8%, whereas in HF it reaches 91.7%. This is consistent with the observation that the numbers of J animals in New Zealand with prominent white spotting are small and the numbers of those that have splashes of white or white accents are larger. It is also noteworthy that the *Q* alleles for the three major QTL are reference alleles in the UMD3.1 genome assembly, which is based on a single Hereford cow. The population frequencies of these variants in the Hereford breed are unknown, and although this breed is not as characteristically spotted as the Holstein breed, Herefords are well known for their white faces (attributed to another mutation in *KIT* (30)), with substantial

white markings concentrated on the belly, brisket, neck, and back.

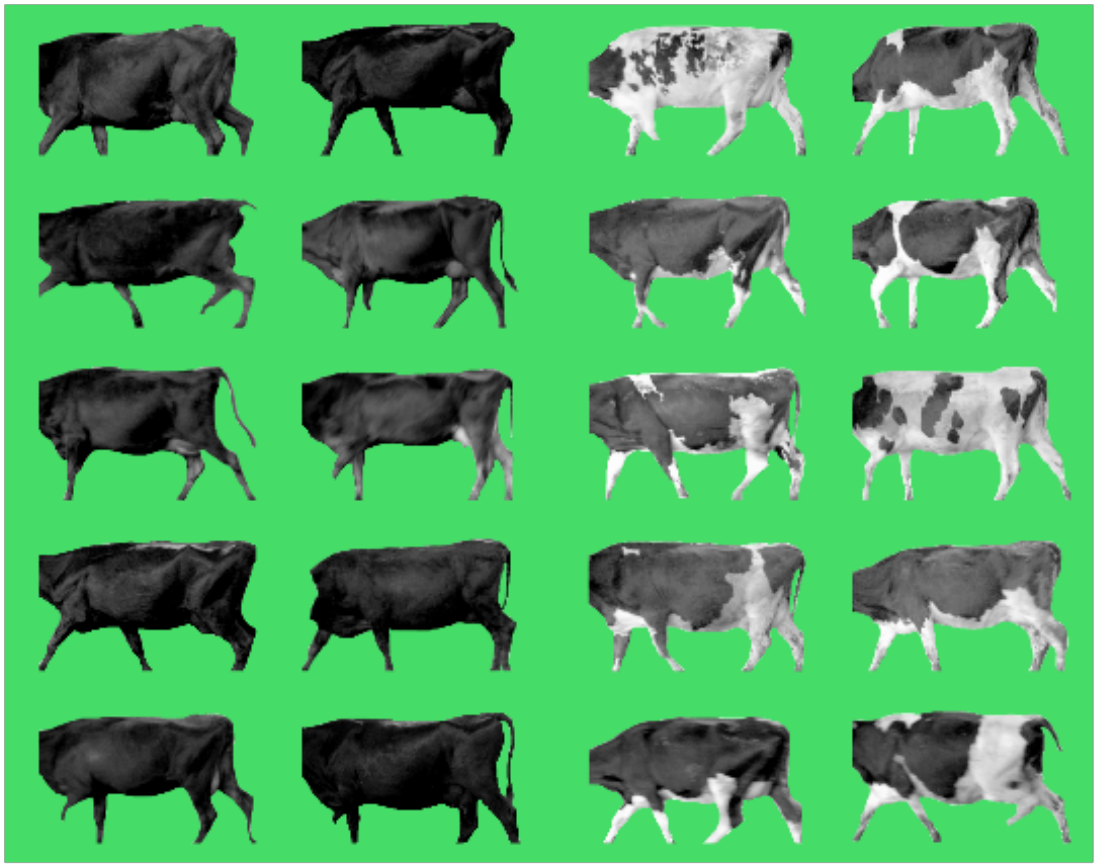


Figure 3.7 Black and white images of 10 ½HF×½J cows carrying the smallest number of Q alleles observed (2Q; left), contrasted with 10 ½HF×½J cows carrying the maximum number of Q alleles at the three major loci (6Q; right).

3.5 Discussion

We present the first association analysis for white spotting in dairy cattle using imputed whole-genome sequence data. This study comprises the largest GWAS for this phenotype, to date, providing details of the genetic effects on white spotting in a population of approximately 3,000 HF, J, and their crosses. We provide evidence for the implication of the *KIT*, *MITF* and *PAX3* genes in white spotting of the coat, and further

suggest regulatory and missense variants that potentially explain the effects of the *MITF* and *PAX3* genes.

MITF is the only plausible candidate for the QTL on chromosome 22, which encodes a transcription factor that has been shown to impact pigmentation in cattle (12,36), mice (37), horses (4,5), dogs (38,39), humans (40), and most recently ducks (41). It is also the only gene located near the top associated variant (Chr22 g.31769747A>G rs209784468), which is situated in intron 2 of MITF transcripts based on the analysis of skin RNA-seq data. The rs209784468 variant falls within a conserved genomic region, which, in conjunction with its status as the lead associated variant, makes rs209784468 a candidate causal variant for this QTL. Given that this SNP and other lead variants are non-coding, and given the lack of other candidate variants that map to protein-coding sequences, we hypothesize that the mechanism underlying the QTL on chromosome 22 is a modulation of the expression of *MITF*. However, how this effect manifests itself during development is unknown. MITF is required during embryonic development to stimulate the transition of neural crest cells into melanocyte precursors (42). If the *MITF* gene is not expressed within the small window during which transition is meant to take place, future expression of *MITF* cannot rescue melanocyte development (42). Impaired functionality or expression of the *MITF* gene during development will result in a reduced number of melanocytes, and manifest itself as white spotting on the coat (42). However, impaired functionality of the *MITF* gene within the mature hair follicle may also impair melanocyte survival and differentiation (29), thus decreasing the number of pigment producing melanocytes. In humans and mice, loss-of-function mutations in *MITF* cause severe symptoms including: coloboma, osteopetrosis,

microphthalmia, albinism and deafness (43,44). Disruptive mutations in *MITF* also cause Tietz syndrome, which is characterized by depigmentation of the skin, hair, iris and severe hearing loss, and Waardenburg syndrome type 2A, which is characterized by patchy depigmentation of the skin and bi- or unilateral deafness in humans and mice (37,40,45). Interestingly, mutations with a strong effect have also been observed in cattle (36,46). The white spotting *MITF* variant that we describe in this study represents a common allele (or nearly fixed in the case of HF animals), with no known effects on hearing or other undesirable phenotypes. The fact that this variant causes a less severe phenotype than the variants with a strong effect fits with an expression-based mechanism for this QTL, however it would still be interesting to compare the phenotype of the segregating individuals for the QTL identified in the current analysis with the phenotypes of individuals with more severe *MITF* syndromes (e.g. hearing loss). In terms of functional analyses, to unambiguously test the role of the rs209784468 SNP and other linked candidates, experiments analogous to those performed in an investigation of human hair colour loci (47) could be performed. Cell-culture-based analyses or studies on model organisms could be conducted to perturb the candidate loci that have an effect on gene expression or pigment formation/melanocyte function.

The most significant variant for the QTL on chromosome 6 mapped to a region 7.5 Mb upstream of the *KIT* gene. Although seemingly too far away to cause this signal, the *KIT* gene is perhaps the single most famous and well-characterized pigmentation gene. There are 19 reported mutations within or near the equine *KIT* gene that cause either complete depigmentation, or white spotting (3,5,6,48), and there are approximately 76 known *KIT* alleles in mice that cause dominant or semi-dominant white spotting (9,49).

A *KIT* translocation mutation has also been identified as the causative mutation for 'colour sidedness' and the white coat phenotype in Belgian Blue and White Galloway cattle (31,50). Although it is possible that the white spotting QTL in the current study is underpinned by contributions from other genes, these facts make *KIT* worthy of consideration as the likely causal agent underlying the chromosome 6 signals. Thus, the inconsistency of the mapping data may instead represent an amalgamation of multiple signals at the locus, and/or some other complexity that is not well represented by our imputed genome sequence dataset. Indeed, when the lead variants were consecutively fitted in our association analyses, no single variant could account for the signal. Given the precedent regarding the *KIT* structural mutations that influence coat phenotypes, we also conducted a sequence-based structural analysis of a broad, 20 Mb region encompassing *KIT* and the top tag variants from the GWAS. This analysis did not reveal any obvious candidate but it is possible that these efforts were confounded by errors in the genome assembly around *KIT*, an observation highlighted through analyses by Whitacre et al. (30). If such confounders exist, breed-specific *de novo* assemblies and sequence information based on long-read sequencing technologies, such as single-molecule sequencing (51), may be helpful in future investigations of the locus. Additional future work could also attempt to fine map the effects in alternative breeds in which fewer QTL could be segregating, or alternatively conduct functional analyses as mentioned in the previous section for the associated variants that map to intron 4 of *KIT* itself.

To our knowledge, the observation of a likely role for *PAX3* in white spotting of the coat in cattle is a novel finding. The top variant for this QTL on chromosome 2 mapped

to a region 0.3 Mb upstream of the *PAX3* gene, although bioinformatic prediction of variant effects revealed a highly associated p.Thr424Met missense mutation that could underlie this QTL. Previous studies have reported variants in *PAX3* that cause pigmentation phenotypes in humans (52), mice (53) and horses (4,5) and variation in ambilateral circumocular pigmentation in the Fleckvieh breed of cattle (54). The latter phenotype describes pigmentation of the area that encircles the animals' eyes in breeds that otherwise have a white head, which raises the possibility that white spotting in HF is influenced by the same QTL that is involved in ambilateral circumocular pigmentation in Flekvieh cattle. In humans, as for some mutations in *MITF,* protein-changing variants in *PAX3* have been shown to cause a similar form of Waardenburg syndrome, which is characterized by wide set eyes, hearing loss and regions of depigmentation in the iris, hair and skin (52,55). Studies in humans and mice have demonstrated that the *PAX3* gene encodes a transcription factor that binds directly to the proximal M promoter of the *MITF* gene, thus facilitating expression of *MITF* (29,55–57). Studies of different spontaneous and radiation-induced *PAX3* mutations in Splotch mice have suggested that *PAX3* is required for proper development of neural crest cells, expansion of melanoblast populations, and prevention of melanoblast terminal differentiation (53). Thus, if the function of the PAX3 protein is altered, *MITF* transcription and activity may be impaired, which in turn may have an impact on regional melanocyte populations and melanogenesis, resulting in an increased proportion of white spotting on the animal's coat. It is also interesting that Hayes et al. (1) observed an association between variants that are located next to the bovine *PAX5* gene and the proportion of black on the coat. We did not observe a genome-wide significant signal on chromosome 8, although this association was demonstrated in Australian Holsteins (1); the highlighted tag SNP in their study was not tested for

association here because it was nearly fixed in our population (MAF < 0.001) and was excluded from the dataset. Unlike *PAX3*, the associations of *PAX5* and *MITF* with melanogenesis are unclear, but the implication of these two structurally related transcription factors in independent GWAS should be analysed in future work. Regarding the other major QTL identified, functional studies are required to confirm a causative effect of the *PAX3* p.Thr424Met mutation, and confirm the molecular mechanism through which this QTL acts.

3.6 Conclusions

Our results add strength to previous analyses that suggest the involvement of the *KIT* and *MITF* genes in white spotting of the coat in cattle, and reveal a new QTL for this trait at the *PAX3* locus. The genes identified highlight the commonality of the mechanisms that underlie the modulation of skin and hair pigmentation in animals, in which all three genes are key regulators of melanocyte development, migration, and differentiation. Moreover, these three genes have already been implicated in the modulation of pigment phenotypes in diverse species. In addition, the sizes of the effect of the major QTL being substantial, there is potential for selection of whiter or darker animals, depending on the farmers' preferences.

3.7 Declarations

3.7.1 Ethics approval

All animal experiments were conducted in strict accordance with the rules and guidelines outlined in the New Zealand Animal Welfare Act 1999. Most data were generated as part of routine commercial activities that are outside the scope of those requiring formal committee assessment and ethical approval (as defined by the above guidelines). Approval was sought for coat scoring procedures that were not based on pre-existing photographs, and subsequently approved by the AgResearch Animal Ethics Committee, Hamilton, New Zealand (approval AEC 14090).

3.7.2 Consent for publication

Not applicable

3.7.3 Availability of data and material

Phenotypic data representing the white spotting phenotype were uploaded as a submission to the Dryad database (doi:10.5061/dryad.tqjq2bvtf) (58). Sequence-based genotype data representing the three QTL of interest were uploaded under the same submission ID. Additional genome-wide data are available upon reasonable request following execution of a transfer agreement, and with permission of Livestock Improvement Corporation.

3.7.4 Funding

This work was supported by the Ministry for Primary Industries (Wellington, New Zealand), which co-funded the work through the Primary Growth Partnership. External funders had no role in the design of the experiment, the collection, analysis, or interpretation of the data, or writing the manuscript.

3.7.5 Competing interests

AY, CC, GW, KT, LM, TJJJ, TJL, SRD, BH, RS, ML are employees of Livestock Improvement Corporation, a commercial provider of bovine germplasm. The remaining authors declare that they have no competing interests.

3.7.6 Authors' contributions

SJ performed most of the bioinformatic and statistical analyses with help from ER, KT, TJL and TJJJ; SJ, AY, CC, GW, LM and ML were involved in data collection; KT conducted sequence imputation; SJ, ML, and SRD conceived the study and experiments; BLH, DG, ML, RGS, RJS and SRD were involved in the supervision of the project; SJ and ML wrote the manuscript. All authors read and approved the final manuscript.

3.7.7 Acknowledgements

The authors would like to acknowledge all farm owners and managers who took part in

our study, and in particular Joyce Voogt for her valuable insights into farmer opinions. We would like to acknowledge Fiona Brown, Nicolas Lopez-Villalobos, Danny Donaghy and Martin Correa Luna from Massey University and Sandeep Seernam from AgResearch for their help during the data collection process. Lastly, we would like to acknowledge Stella Sim, Esther Donkersloot and Neil Macdonald from LIC for providing photographs used in this research.

3.8 References

1. Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME. Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 2010;6(9):e1001139.

2. Mort RL, Jackson IJ, Patton EE. The melanocyte lineage in development and disease. Development. 2015;142(7):1387–1387.

3. Brooks SA, Lear TL, Adelson DL, Bailey E. A chromosome inversion near the KIT gene and the Tobiano spotting pattern in horses. Cytogenet Genome Res. 2007;(119):225–30.

4. Hauswirth R, Haase B, Blatter M, Brooks SA, Burger D, Drögemüller C, et al. Mutations in MITF and PAX3 cause “splashed white” and other white spotting phenotypes in horses. PLoS Genet. 2012;8(4):e1002653.

5. Hauswirth R, Jude R, Haase B, Bellone RR, Archer S, Holl H, et al. Novel variants in the KIT and PAX3 genes in horses with white-spotted coat colour phenotypes. Anim Genet. 2013;44(6):763–5.

6. Haase B, Brooks SA, Tozaki T, Burger D, Poncet PA, Rieder S, et al. Seven novel KIT mutations in horses with white coat colour phenotypes. Anim Genet. 2009;40(5):623–9.

7. Marklund S, Moller M, Sandberg K, Andersson L. Close association between sequence polymorphism in the KIT gene and the roan coat color in horses. Mamm Genome. 1999;10(3):283–8.

8. Karlsson EK, Baranowska I, Wade CM, Salmon Hillbertz NHC, Zody MC, Anderson N, et al. Efficient mapping of mendelian traits in dogs through genome-wide association. Nat Genet. 2007;39(11):1321–8.

9. Baxter LL, Hou L, Loftus SK, Pavan WJ. Spotlight on spotted mice: A review of white spotting mouse mutants and associated human pigmentation disorders. Pigment Cell Res. 2004;17(3):215–24.

10. Liu L, Harris B, Keehan M, Zhang Y. Genome scan for the degree of white spotting in dairy cattle. Anim Genet. 2009;40(6):975–7.

11. Fontanesi L, Scotti E, Russo V. Haplotype variability in the bovine MITF gene and association with piebaldism in Holstein and Simmental cattle breeds. Anim Genet. 2012;43(3):250–6.

12. Hofstetter S, Seefried F, Häfliger IM, Jagannathan V, Leeb T, Drögemüller C. A non-coding regulatory variant in the 5′-region of the MITF gene is associated with white-spotted coat in Brown Swiss cattle. Anim Genet. 2019;50(1):27–32.

13. Berry SD, Davis SR, Beattie EM, Thomas NL, Burrett AK, Ward HE, et al. Mutation in bovine β-carotene oxygenase 2 affects milk color. Genetics. 2009;182(3):923–6.

14. Littlejohn MD, Henty KM, Tiplady K, Johnson T, Harland C, Lopdell T, et al. Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. Nat Commun. 2014;5:1–8.

15. Karim L, Takeda H, Lin L, Druet T, Arias JAC, Baurain D, et al. Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. Nat Genet. 2011;43(5):405–13.

16. Lopdell TJ, Tiplady K, Struchalin M, Johnson TJJ, Keehan M, Sherlock R, et al. DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. BMC Genomics. 2017;18(1):1–18.

17. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, et al. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. Sci Rep. 2016;6:25376.

18. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

19. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43(5):491–8.

20. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084–97.

21. Berry SD, Lopez-Villalobos N, Beattie EM, Davis SR, Adams LF, Thomas NL, et al. Mapping a quantitative trait locus for the concentration of β-lactoglobulin in milk, and the effect of β-lactoglobulin genetic variants on the composition of

milk from Holstein-Friesian x Jersey crossbred cows. N Z Vet J. 2010;58(1):1–5.

22. Koufariotis LT, Chen YPP, Chamberlain A, Jagt C Vander, Hayes BJ. A catalogue of novel bovine long noncoding RNA across 18 tissues. PLoS One. 2015;10(10):1–24.

23. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.

24. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80–92.

25. Aken BL, Achuthan P, Akanni W, Amode MR, Bernsdorff F, Bhai J, et al. Ensembl 2017. Nucleic Acids Res. 2017;45(D1):D635–42.

26. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018;46(D1):D754–61.

27. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28(12):1647–9.

28. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974–84.

29. D'Mello S, Finlay G, Baguley B, Askarian-Amiri M. Signaling pathways in melanogenesis. Int J Mol Sci. 2016;17(7):1144.

30. Whitacre L. Structural variation at the KIT locus is responsible for the piebald

phenotype in Hereford and Simmental cattle. University of Missouri--Columbia; 2014.

31. Durkin K, Coppieters W, Drögüller C, Ahariz N, Cambisano N, Druet T, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature. 2012;482(7383):81–4.

32. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.

33. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–82.

34. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Protoc Hum Genet. 2013;76(1):7–20.

35. Adzhubei I, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Publ Gr. 2010;7(4):229–39.

36. Bourneuf E, Otz P, Pausch H, Jagannathan V, Michot P, Grohs C, et al. Rapid discovery of de novo deleterious mutations in cattle enhances the value of livestock as model species. Sci Rep. 2017;7(1):11466.

37. Hou L, Pavan WJ. Transcriptional and signaling regulation in neural crest stem cell-derived melanocyte development: do all roads lead to Mitf? Cell Res. 2008;18:1163–76.

38. Baranowska Körberg I, Sundström E, Meadows JRS, Rosengren Pielberg G, Gustafson U, Hedhammar Å, et al. A simple repeat polymorphism in the MITF-

M promoter is a key regulator of white spotting in dogs. PLoS One. 2014;9(8):e104363.

39. Schmutz SM, Berryere TG, Dreger DL. MITF and white spotting in dogs: A population study. J Hered. 2009;100(suppl_1):S66–74.

40. Léger S, Balguerie X, Goldenberg A, Drouin-Garraud V, Cabot A, Amstutz-Montadert I, et al. Novel and recurrent non-truncating mutations of the MITF basic domain: genotypic and phenotypic variations in Waardenburg and Tietz syndromes. Eur J Hum Genet. 2012;20(5):584–7.

41. Zhou Z, Li M, Cheng H, Fan W, Yuan Z, Gao Q, et al. An intercross population study reveals genes associated with body size and plumage color in ducks. Nat Commun. 2018;9(1):2648.

42. Opdecamp K, Nakayama A, Nguyen M, Hodgkinson C, Pavan W, Arnheiter H. Melanocyte development in vivo and in neural crest cell cultures: crucial dependence on the Mitf basic-helix-loop-helix-zipper transcription factor. Development. 1997;124(12):2377–86.

43. George A, Zand DJ, Hufnagel RB, Sharma R, Sergeev Y V, Legare JM, et al. Biallelic mutations in MITF cause coloboma, osteopetrosis, microphthalmia, macrocephaly, albinism, and deafness. Am J Hum Genet. 2016;99:1388–94.

44. Steingrímsson E, Moore KJ, Lamoreux ML, Ferré-D'Amaré AR, Burley SK, Sanders Zimring DC, et al. Molecular basis of mouse microphthalmia (mitf) mutations helps explain their developmental and phenotypic consequences. Nat Genet. 1994 Nov;8(3):256–63.

45. Shibahara S. Microphthalmia-associated transcription factor (MITF): Multiplicity

in structure, function, and regulation. J Investig Dermatology Symp Proc. 2001;6(1):99–104.

46. Philipp U, Lupp B, Mömke S, Stein V, Tipold A, Eule JC, et al. A MITF mutation associated with a dominant white phenotype and bilateral deafness in German Fleckvieh cattle. PLoS One. 2011;6(12):4–9.

47. Guenther CA, Tasic B, Luo L, Bedell MA, Kingsley DM. A molecular basis for classic blond hair color in Europeans. Nat Genet. 2014;46(7):748–52.

48. Brooks SA, Bailey E. Exon skipping in the KIT gene causes a Sabino spotting pattern in horses. Mamm Genome. 2005;16(11):893–902.

49. Smith CL, Blake JA, Kadin JA, Richardson JE, Bult CJ, Mouse Genome Database Group. Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. Nucleic Acids Res. 2018 Jan 4;46(D1):D836–42.

50. Brenig B, Beck J, Floren C, Bornemann-Kolatzki K, Wiedemann I, Hennecke S, et al. Molecular genetics of coat colour variations in White Galloway and White Park cattle. Anim Genet. 2013;44(4):450–3.

51. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Von Haeseler A, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461–8.

52. Pingault V, Ente D, Dastot-Le Moal F, Goossens M, Marlin S, Bondurand N. Review and update of mutations causing Waardenburg syndrome. Hum Mutat. 2010;31(4):391–406.

53. Kubic JD, Young KP, Plummer RS, Ludvik AE, Lang D. Pigmentation PAX-ways: The role of Pax3 in melanogenesis, melanocyte stem cell maintenance, and

disease. Pigment Cell Melanoma Res. 2008;21(6):627–45.

54. Pausch H, Wang X, Jung S, Krogmeier D, Edel C, Emmerling R, et al. Identification of QTL for UV-protective eye area pigmentation in cattle by progeny phenotyping and genome-wide association analysis. PLoS One. 2012;7(5):e36346.

55. Bondurand N, Pingault V, Goerich DE, Lemort N, Sock E, Caignec C Le, et al. Interaction among SOX10, PAX3 and MITF, three genes altered in Waardenburg syndrome. Hum Mol Genet. 2000;9(13):1907–17.

56. Watanabe A, Takeda K, Ploplis B, Tachibana M. Epistatic relationship between Waardenburg Syndrome genes MITF and PAX3. Nat Genet. 1998;18(3):283–6.

57. Potterf SB, Furumura M, Dunn KJ, Arnheiter H, Pavan WJ. Transcription factor hierarchy in Waardenburg syndrome: Regulation of MITF expression by SOX10 and PAX3. Hum Genet. 2000;107(1):1–6.

3.9 Appendix

Table 3.S1 Absolute number of animals genotyped per SNP Chip and number of SNPs per chip.

Genotyping Platform	Animals	SNPs
50kv1	600	53126
50kv2	1297	53629
GGP50k	1051	48156
GGP50v1.1	109	48161
GGPHDv2	156	138419
GGPv1	334	8729
GGPv2	202	20012
GGPv2.1	6	20015
GGPv3	726	31813
GGPv3.1	100	31945
GGPv4	180	37092
HD	458	772235

Some cattle were genotyped on more than one panel, so included in multiple categories. The number of SNPs per panel presented in this table reflect numbers prior to filtering based on quality metrics.

Table 3.S2 Number of purebred Jerseys and Holstein-Friesians carrying 0-6Q alleles and corresponding mean percentage of white value.

	Number of Q alleles						
	0	**1**	**2**	**3**	**4**	**5**	**6**
Jersey N	1	13	47	85	91	32	5
Jersey – mean percentage white	0	0	0.08	0.8	1.5	4.2	30.4
Holstein-Friesian N	0	0	0	1	4	44	540
Holstein-Friesian – mean percentage white	-	-	-	0	4.3	15.0	26.6

The mean percentage of white value reported is representative of raw phenotype measurements in purebred J and HF cattle from the mapping population. No fixed effects have been fitted to account for population structure or other confounding effects during this calculation.

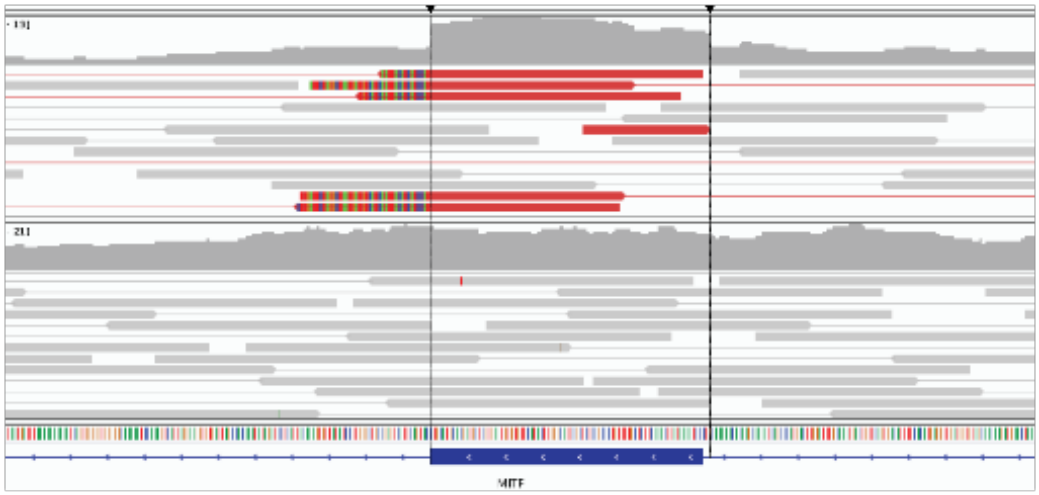


Figure 3.S1 Read depth anomalies at intron-exon boundaries of MITF around exon 4 suggest the presence of a pseudogene. The top sequence alignment track represents a whole-genome sequenced animal heterozygous for the Chr22 g.31769331C>T (rs110881545) variant, for which read-depth is increased across the exons and soft-clipped reads show evidence of mismatches to neighbouring exon structures.

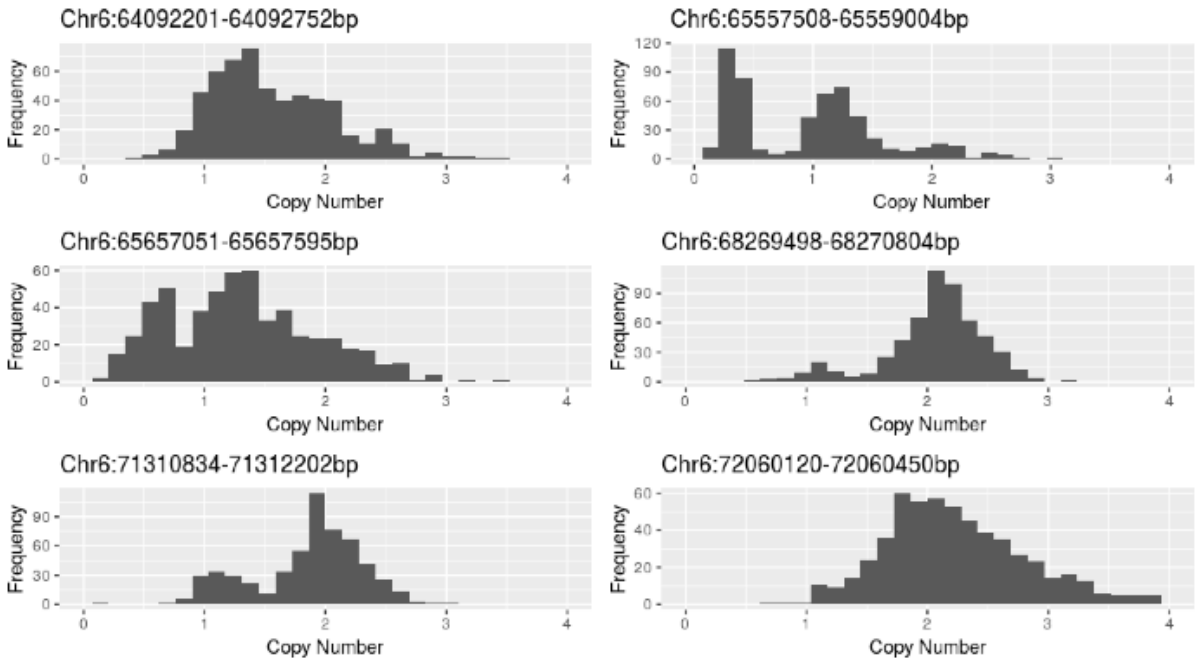


Figure 3.S2 Frequency of CNVnator assigned copy number across 565 sequenced cattle for each of the six candidate structural variants identified at the chromosome 6 locus. Four of the six structural variants show clear evidence of multimodality.

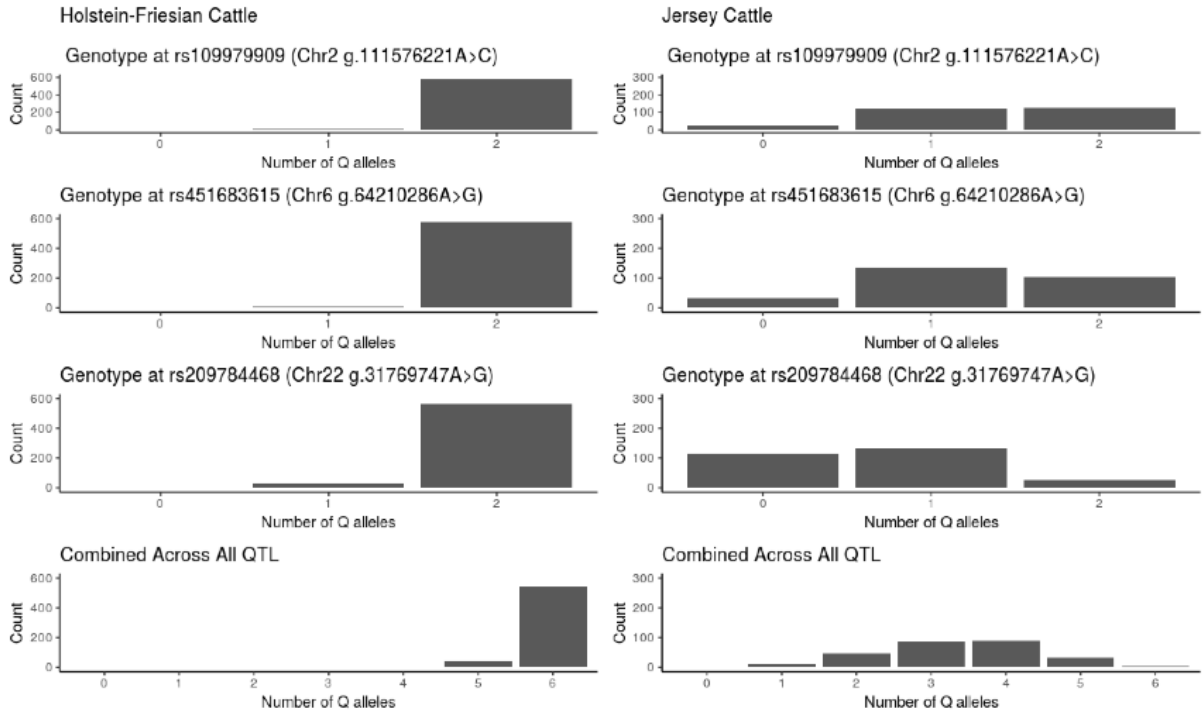


Figure 3.S3 Distribution of Q allele counts for each tag variant and combined across loci in cattle identified as purebred Holstein-Friesian (left) and pure-bred Jersey (right) within the population used for mapping.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

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:	Swati Jivanji
Name/title of Primary Supervisor:	Dorian Garrick
In which chapter is the manuscript /published work:	Chapter Three

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Jivanji S, Worth G, Lopdell TJ, Yeates A, Couldrey C, Reynolds E, Tiplady K, McNaughton L, Johnson TJ, Davis SR, Harris B, Spelman R, Snell RG, Garrick D, Littlejohn MD. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. Genetics Selection

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Swati Jivanji
Date:	25/11/2021
Primary Supervisor's Signature:	Dorian Garrick
Date:	25 Nov 2021

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.

CHAPTER FOUR

Novel structural and epistatic mutations at the KIT locus are associated with white patterning traits in bovine breeds.

IN PREPARATION

Novel structural and epistatic mutations at the *KIT* locus are associated with white patterning traits in bovine breeds.

Swati Jivanji[1*], Emma Wilkinson[2], Emily Mears[3], Anna Yeates[4], Chad Harland[4], Charlotte Gray[4], Christine Couldrey[4], Gemma Worth[4], Janelle Moody[3], Lorna McNaughton[4], Mike Keehan[1], Tony Fransen[4], Tracey Monehan[4], Yu Wang[4], Richard Spelman[4], Dorian Garrick[1], Russell Snell[3], Richard Mort[2] & Mathew Littlejohn[1,4].

[1]School of Agriculture and Environment, Massey University, Palmerston North, New Zealand

[2]School of Biomedical and Life Sciences, University of Lancaster, England

[3]School of Biological Sciences, University of Auckland, New Zealand

[4]Livestock Improvement Corporation, Newstead, New Zealand

*Corresponding author

Email: swati.jivanji.1@uni.massey.ac.nz

4.1 Abstract

Coat colour and patterning traits in domestic species have undergone intense selection, and have become breed-defining characteristics of many populations. White spotting in cattle is one such trait and is a hallmark of Holstein-Friesian cattle. Previous studies have implicated the involvement of *PAX3*, *MITF*, and *KIT* in the proportion of white spotting, with candidate causal mutations having been proposed at the *PAX3* and *MITF* loci. However, identification of a causal mutation at the *KIT* locus has remained elusive. Using a combination of both short- and long-read sequencing strategies we have identified a novel 6,948bp deletion 114 kb upstream of the *KIT* gene as a strong candidate for white spotting of the coat. Bioinformatic analyses suggest that the long-form of this structural variant is highly conserved amongst mammals and represents the ancestral allele, harbouring a number of regulatory features including a MITF transcription-factor binding site. Phylogenetic analyses show that this polymorphism segregates in many bovine breeds, and is near fixed in characteristically spotted populations, further suggesting its role as a modulator of white coat patterning. We postulated that this variant might underlie epistatic effects apparent in Hereford cattle, where F_1 crosses occasionally produce calves that lack the pure white faces characteristic of that breed. While association analyses did not support a role for the *KIT* non-coding element, genome-wide analysis revealed a strong epistatic interaction with the *MITF* locus, highlighting the same variant previously implicated in white spotting of Holstein-Friesian and Jersey cattle. The findings in this study contribute new insight into the genetic basis and selection of white spotting in cattle and highlight interactions between coat colour loci.

4.2 Introduction

Coat colour and patterning traits have interested breeders for hundreds, if not thousands of years, as demonstrated by early Lascaux cave drawings of cattle with white spotting (1). Striking coat patterning traits, such as white spotting, white-face, belted, colour-sidedness, and speckled were captured by early breeders due to their uniqueness and easy means for breed identification, and intense selection has driven these traits to fixation in many breed populations. The genetic variants that cause these traits have largely been resolved over the past decade (2–5), with the exception of the white spotting pattern characteristic of Holstein cattle (6). White spotting appears to have an oligogenic architecture, where we previously identified three major effect genes, *PAX3*, *MITF* and *KIT,* that together explained approximately 57% of the phenotypic variance in white spotting (7). Of note, the observed effect across the three QTL appeared to be approximately 30%, which is likely indicative of a non-additive, epistatic mode of inheritance. Two of the identified quantitative trait loci (QTL) presented plausible causative variants for these effects, however the QTL on chromosome 6 presented no compelling candidate mutations and exhibited a peculiar, highly dispersed association signal. A structural variant analysis revealed a 330bp duplication approximately 142 kb downstream of the *KIT* gene that was best tagged by a single nucleotide polymorphism (SNP) mapping near the top of the association peak, albeit with a moderate correlation (maximum R^2=0.432) (7). Ultimately, fine-mapping analysis of the *KIT* locus remained inconclusive in that study.

Several other coat patterning traits have been attributed to structural variations at the *KIT* locus in cattle. The colour-sidedness trait was one of the first cattle coat patterning traits suggested to be caused by structural variation at the *KIT* locus (2). That

dominantly inherited trait, observed in Belgian Blue and Brown Swiss cattle, was proposed to be caused by a reciprocal translocation of a duplicated region at the *KIT* locus to a region on chromosome 29 via a circular intermediate. The mutation is hypothesised to have a gain-of-function effect resulting from dysregulated expression of the translocated *KIT* gene (2). Brenig et al. (3) later reported that the initial translocation of a region encompassing the *KIT* gene to chromosome 29 was also present in White Galloway and White Park cattle, where one copy was suggested to cause a speckled phenotype, and two copies were suggested to cause a white phenotype with black ears, muzzle, eyes and feet. The white-face trait observed in Hereford cattle has also been attributed to a structural variant at the *KIT* locus, where Whitacre (4) proposed that a serial duplication approximately 45 kb upstream of the *KIT* gene likely causes animals to have a pure white-face, and functions via a dominant Mendelian effect. More recently, a 9.4 kb deletion at Chr6:70,417,067-70,417,114bp, a 310 kb duplication of Chr4:85,174,891-85,174,937bp, and subsequent insertion of this duplicated sequence into the site of the deletion on chromosome 6 was found in Gloucester cattle (8). This mutation has been associated with a colour-sidedness, or 'line-back' phenotype in Gloucester cattle (8), Austrian Pinzgaur cattle, and some Tux-Zillertaler cattle (9). Given the precedent for structural variation at the *KIT* locus influencing coat patterning traits in cattle, in the current study we aimed to investigate the previously identified (7) 330bp *KIT* candidate mutation further. Using a combination of short-read sequence data and long molecule sequencing, we show that the previously identified duplication event downstream of the *KIT* gene is a sequence-mapping artifact. The predicted duplication was instead explained by a novel 6.9 kb insertion upstream of the *KIT* gene that is a strong candidate underlying white spotting of the coat. We also demonstrate novel epistatic effects between the Hereford white-face *KIT* mutation and a previously

identified *MITF* mutation, where one copy of the *MITF* variant appears to be sufficient to modify this otherwise dominant trait.

4.3 Results

4.3.1 Identification of novel structural variation at the *KIT* locus

Based on a previous, automated structural variant analysis of the chromosome 6 locus, we identified a 330bp duplication as moderately correlated with tag SNPs of the white spotting QTL (7). To investigate this candidate further, we performed a manual, more detailed sequence-based analysis of the region, inspecting sequence alignments from animals of contrasting QTL genotype. This analysis revealed discordant read-pairs and possible spurious mapping in cattle heterozygous or homozygous for the 'solid coloured' allele (Fig 4.S1; animals differentiated using the rs451683615 tag variant reported by Jivanji et al. (7)). These discordant read-pairs had a mapped insert size of approximately 400 kb, suggesting that the previously highlighted 330bp duplication (7) might represent a larger, uncharacterised structural mutation of unknown composition. Notably, mate-pairs of these discordant reads implicated a new region upstream of the *KIT* gene, where the forward mate mapped to Chr6:70,396,258-70,396,788bp (hereafter referred to as the 3' site), and the reverse mate mapped upstream of the *KIT* gene to Chr6:70,052,058-70,053,039bp (hereafter referred to as the 5' site). Soft-clipped reads were observed at both identified regions (Fig 4.S1b), though these sequences did not appear to be captured in the reference genome, and thus suggested against a straightforward inversion or translocation event.

Since these alignment anomalies appeared to be polymorphic across animals, we sought to classify 152 genome-sequenced individuals based on discordant read status (targeted due to having >10× average read-depth). Here, unaligned genome sequence files were queried using the unique split-read sequences to identify animals that were (or were not) variant at the 5' and 3' sites of interest (see Methods). Assessment of the relationship between these 'split-read genotypes' and the previously reported top QTL tag single nucleotide polymorphisms (SNPs; rs451683615 and rs463810013), found a strong and improved correlation (R^2>0.75) over that previously demonstrated with the 330bp duplication reported by Jivanji et al. (7) (Table 4.S1). These data suggested this new mutation as potentially responsible for the white spotting QTL, and motivated investigation to further characterise the variant.

4.3.2 Molecular characterisation of the structural variant

Given the constraints of whole-genome sequence data obtained with short read lengths, we used long-range polymerase chain reaction (PCR) and Oxford nanopore minION sequencing to characterise the candidate structural variant junctions. As the soft-clipped base identity did not support a simple inversion and/or duplication mutation, we amplified each putative junction independently, targeting three Jersey bulls, three Holstein-Friesian bulls, and two Hereford bulls based on their genotypes at the rs451683615 and rs463810013 QTL tag variants (Table 4.S2). Long-range PCR of a 4.7 kb region encompassing the 3' candidate structural variant site (Chr6:70,394,202-70,399,130bp) did not reveal any differences in amplicon sizes between individuals, with minION sequence data conforming to that represented in the ARS-UCD1.2 reference genome. However, amplification of a 6.3 kb region encompassing the 5' candidate site (Chr6:70,048,369-70,050,884bp), highlighted an approximately 13 kb

amplicon in two of the three Jersey bulls, and an amplicon of expected size in all other bulls (Fig 4.1). Analysis of minION sequence data representing these amplicons confirmed a novel 6,948bp insertion at the 5' site in all Jersey cattle (two bulls being homozygous for the insertion and one heterozygous).

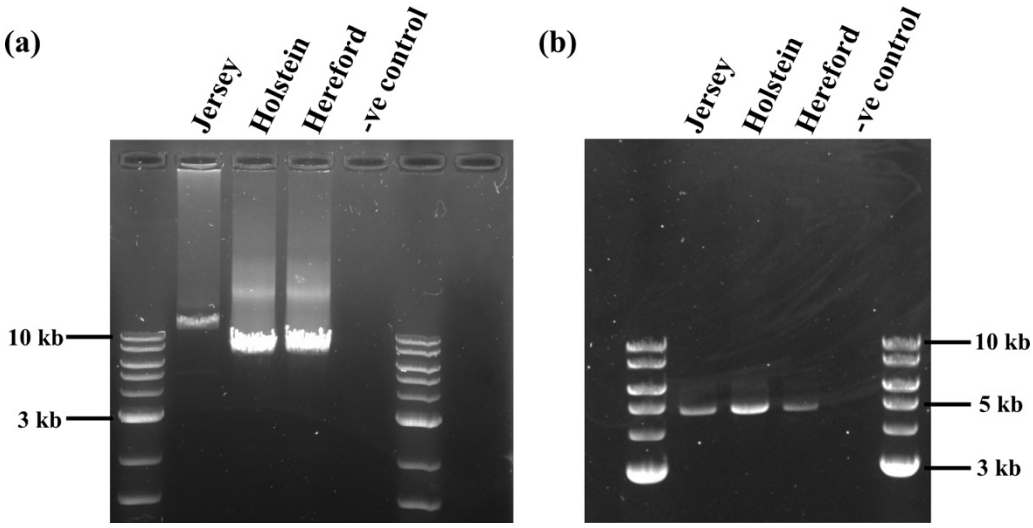


Figure 4.1 Amplicons from long-range polymerase chain reaction across the candidate structural variant sites in Jersey, Holstein, and Hereford bulls. (a) Amplification of a 6.3 kb fragment at the 5' candidate structural variant site (Chr6:70,048,369-70,050,884bp) showed a larger fragment (~13 kb) in Jerseys compared to the other bulls. (b) All bulls had the expected 4.7 kb fragment at the 3' candidate structural variant site (Chr6:70,394,202-70,399,130bp).

Amplicons representing the 5' and 3' regions presented co-locating biallelic variants, confirming that PCR was unlikely confounded by allele drop-out across the two regions of interest. These observations supported the assumption that a larger translocation and/or inversion event between the loci was unlikely, and that the discordant reads observed at the 3' site were instead an artefact of read mapping caused by the absence of the 5' 6.9 kb insertion from the reference genome. To test this hypothesis, we incorporated the novel 6.9 kb insertion into the ARS-UCD1.2 reference genome and remapped genome-wide short-read data in several samples of relevant QTL genotype

(see Methods). As suspected, discordant reads were no longer observable at the 3' locus in these animals, and previously unmapped reads now mapped to the novel inserted sequence (Fig 4.2). Notably, the novel 6.9 kb insertion was flanked by two near-identical 42bp fragments with homology to a bovine retrotransposable (RTE-BovB) long interspersed nuclear element (LINE). This sequence was also present at the 3' site downstream of *KIT* (and 1,905 other locations across the ARS-UCD1.2 genome – see Methods), suggesting the likely cause of the read-mapping artefacts.

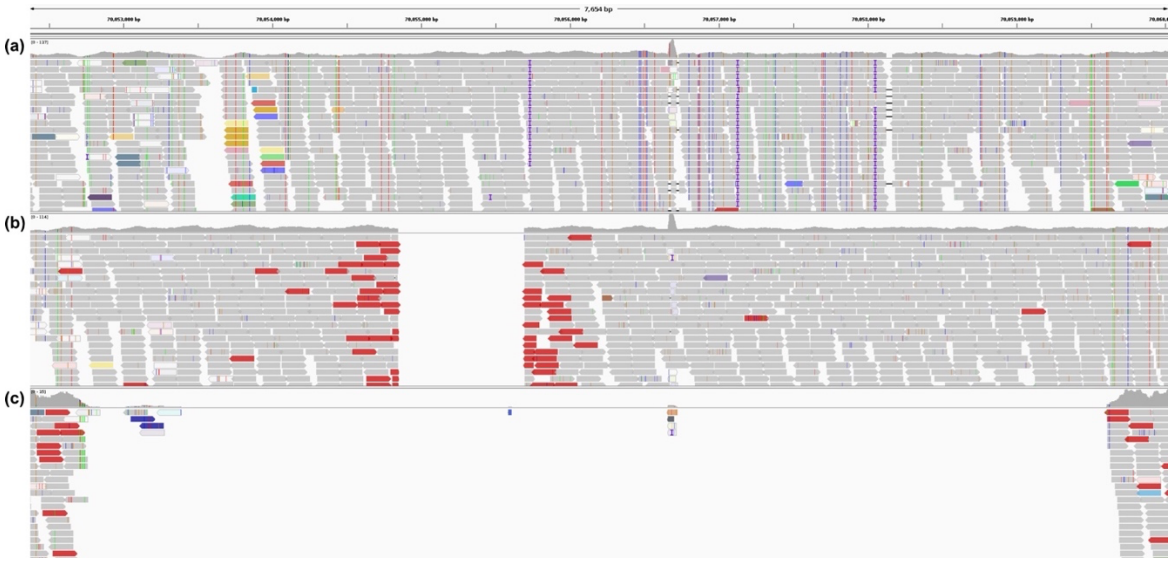


Figure 4.2 Short read sequence data mapped to the bespoke reference genome with the ancestral allele sequence incorporated between ARS-UCD1.2 Chr6:70,052,697bp and Chr6:70,052,698bp. The cattle mapped to this region represent (a) the ancestral allele, (b) the 850bp deletion allele, and (c) the 6.9 kb deletion allele. The sequence reads highlighted in red have an insert size larger than expected (i.e., >150bp), indicative of a deletion.

4.3.3 The 5' structural mutation is an evolutionarily conserved ancestral allele

Given that the bovine genome reference is based on a Hereford animal, and both Holstein-Friesian and Hereford cattle are spotted and lack the novel 6.9 kb structural variant, we considered the long allele as the likely ancestral, solid-coloured form. In this context, the Holstein-Friesian and Hereford (i.e., reference) alleles represent the deleted

state, so we hereafter refer to the mutation as a deletion. This non-coding sequence was distant to the *KIT* gene, locating approximately 114 kb upstream of the transcription start site. While this considerable genomic distance might otherwise undermine the candidacy of the variant, assessment of sequence conservation revealed syntenic regions in various other organisms including human and mouse (Fig 4.3), suggesting a regulatory role for the sequence. Here, strong conservation of regions nested within the 6.9 kb deletion are apparent, including ~200bp of genomic DNA near contiguous to human, with higher average nucleotide conservation than bovine and human *KIT* protein coding sequences (Fig 4.3; 94.2% versus 88.9% identity respectively). Notably, this same region contains a variety of regulatory features including a ChIP-seq-annotated MITF transcription factor binding site (mapping to Chr4:54,571,293-54,571,587bp on the human GRCh38/hg38 reference assembly, 86 kb upstream of the human *KIT* gene). Given the critical role of *MITF* in melanocyte biology (10), and its status as the gene responsible for the other largest-effect white spotting QTL in cattle (7), we considered knock-out of this sequence a strong candidate mutation for white spotting in our population.

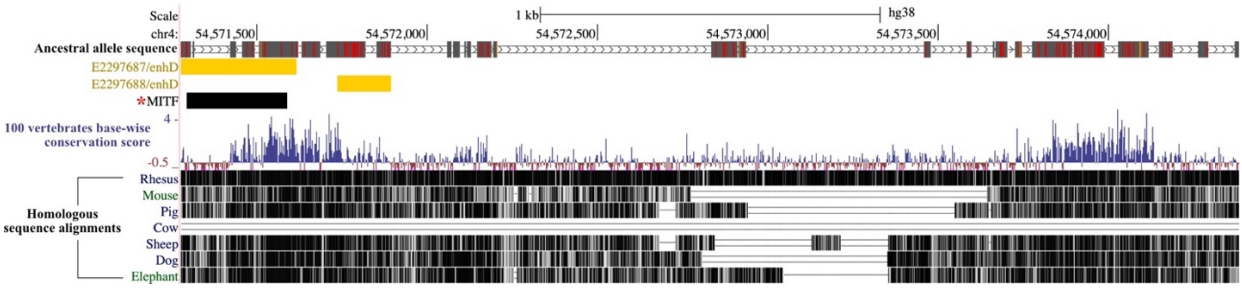


Figure 4.3 The novel 6.9 kb sequence aligned to the human GRCh38/hg38 reference genome with functional element and conservation annotations from the UCSC genome browser. The solid grey bars indicate contiguous syntenic sequence between the ancestral allele with red lines indicating SNPs. The ancestral allele sequence overlaps distal enhancer-like signatures (indicated in yellow), and a ChIP-seq identified MITF transcription factor binding site (indicated with a red asterisks). This same sequence appears to have a high base-wise conservation score, and is

highly conserved in a variety of mammals. As also indicated by the mammalian multi-species alignments (bottom 7 tracks) this sequence is absent from the Hereford-derived cow reference genome.

4.3.4 Association analysis between *KIT* SV deletion haplotypes and white spotting

To test whether this variant was associated with white spotting in New Zealand (NZ) dairy cattle, we aimed to impute the 6.9 kb deletion into a population of 2,967 mixed breed cattle with pre-existing coat phenotypes (7). Aside from the 6.9 kb deletion, we also noted the existence of an intermediate deletion allele of 850bp from the minION sequence data (Fig 4.2). Although this intermediate deletion haplotype contained the same conserved, candidate regulatory sequences as that carried by the longer-form ancestral allele (see Fig 4.S2), we aimed to represent this haplotype for imputation to enable association testing of all possible structural alleles. To form the imputation reference genotypes, the genomes of 1,127 NZ dairy cattle were first remapped to the modified ARS-UCD1.2 genome assembly containing the long-form ancestral allele (referenced above). Structural variant diplotypes were then called using a curated approach, and imputation was performed with haplotypes represented as a single triallelic mutation (see Methods). Within-breed frequencies of these haplotypes are indicated in Table 4.S3, where notably, purebred Holstein-Friesian cattle appear fixed for the 6.9 kb deletion allele (i.e., the allele proposed to increase white spotting).

Since our hypothesis was that the chromosome 6 white spotting QTL was due to deletion of a highly conserved, non-coding regulatory element carried by both the long and intermediate form ancestral alleles, our primary association analysis merged these haplotypes for comparison with the 6.9 kb deletion allele. Association analysis was

conducted for these diplotypes in conjunction with 152,071 other variants mapping to a 20 Mb interval previously shown to capture the white spotting QTL (7). Figure 4.4 shows the regional Manhattan plot for this analysis. Notably, the 6.9 kb deletion ranked amongst the most highly associated variants for white spotting (p=4.46×10^{-100} versus p=3.21×10^{-101} for the lead variant Chr6 g.70210094A>C). The effect sizes for the 6.9 kb deletion were substantial, estimated as an increase of 11.9±0.2% in white spotting per allele. When the 6.9 kb deletion was fitted as a fixed effect in these association models, the significance of most (though not all) of the variants within the 20 Mb interval was lost (smallest p=8.48×10^{-11} for Chr6 g.70343862A>T). This residual signal was also seen when the top-associated variant was fitted as a fixed effect (Chr6 g.70210094A>C*;* smallest p=7.64×10^{-11} for Chr6 g.70343862A>T)*,* suggesting allelic heterogeneity, imputation error, or unaddressed population stratification as potentially responsible for these effects. Further association analysis comparing the 6.9 kb deletion to either the long-form or intermediate form ancestral alleles similarly showed near-top ranked association for the 6.9 kb deletion variant (Fig 4.S3; see Methods). Comparison of effects between the long and intermediate form ancestral haplotypes did not reveal significant differences between alleles (Fig 4.S3). These findings support the hypothesis that the full length 6.9 kb deletion is necessary to cause increased white spotting of the coat, though we note that the numbers of animals comparing long, and intermediate length ancestral alleles was limited (N=85).

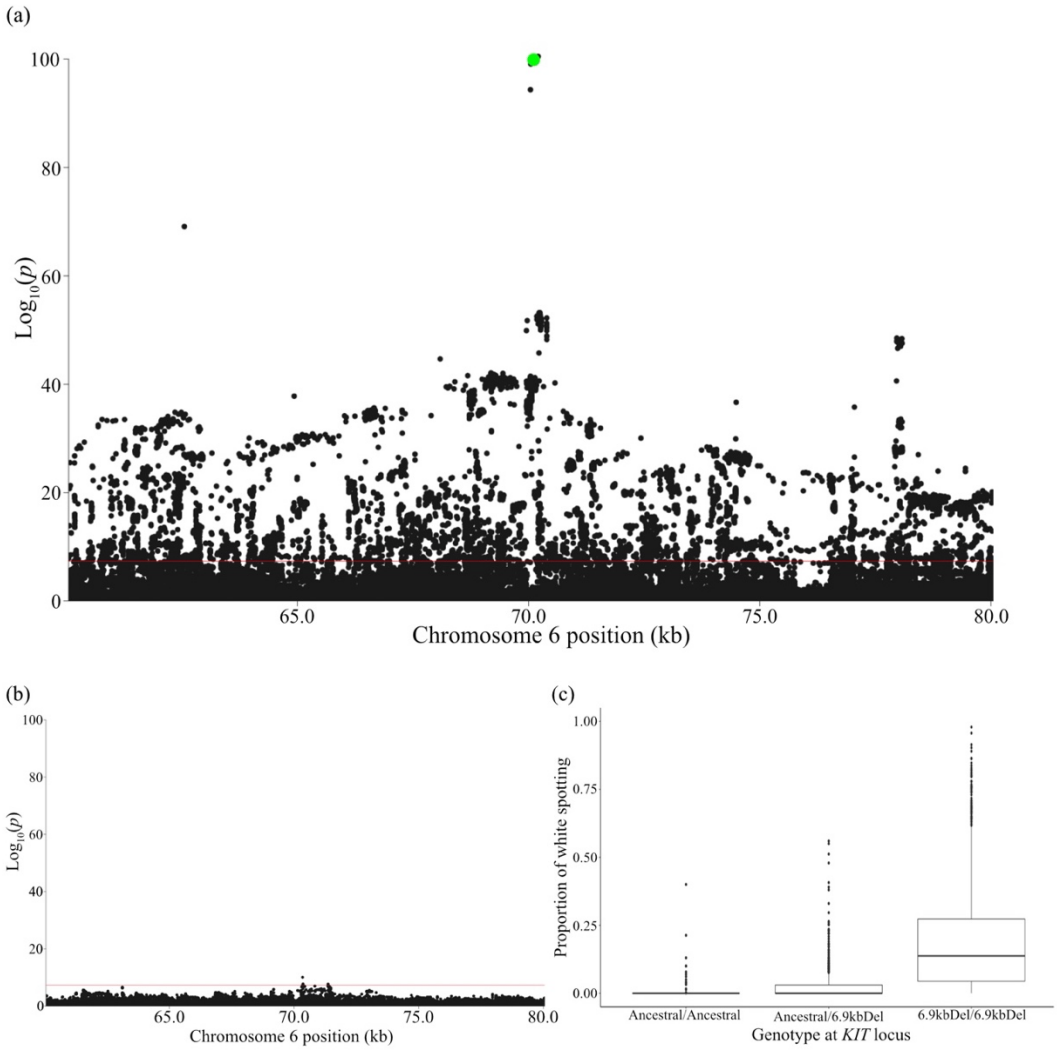


Figure 4.4 Association between the KIT structural variant and the proportion of white spotting. (a) Association results for the proportion of white spotting, where the KIT structural variant was imputed as a biallelic variant representing the ancestral allele or 6.9 kb deletion allele. The KIT structural variant (indicated in green) had a p-value of 4.46×10^{-100}, and the red line indicates the significance threshold $p=5\times10^{-8}$. (b) Almost all of the association signal is removed when the KIT structural variant is fitted as a fixed effect. (c) The boxplots show the proportion of white spotting based on genotype at the KIT locus for cattle with the either the ancestral allele, or the 6.9 kb deletion allele.

4.3.5 The *KIT* 6.9 kb deletion segregates in other breeds

Although Holstein cattle are perhaps the most well-recognised white-spotted breed, many other breeds also have prominent white markings. To investigate whether the non-

coding deletion upstream of *KIT* segregated in other breeds, and how mutation status might align with coat pigmentation characteristics iconic of these breeds, we downloaded sequence data for a variety of spotted and non-spotted breeds from the NCBI Sequence Read Archive (see Methods). We aligned these data to our bespoke chromosome 6 reference genome, with structural variant genotypes derived for 548 animals representing these 13 breeds (Table 4.S5). Notably, all characteristically spotted breeds were fixed, or near fixed, for the 6.9 kb deletion allele (frequencies ranging from 0.97-1). The reverse was true in the majority of non-spotted breeds, and although the 6.9 kb deletion did segregate in Jerseys, Red Angus, Charolais, and Bos indicus cattle (frequencies ranging from 0.03-0.33; Table 4.S5). All of these breeds are known to present some individuals with white markings on the coat (or wholly white coats in the case of Charolais and some Bos indicus animals; (11)). The intermediate length, 850bp deletion was observed only in non-spotted breeds (Limousin, Jersey, Red Angus, Angus, Gelbvieh, and Charolais), and was the minor allele in all cases (frequencies ranging from 0.08-0.25). Since this apparent relationship between mutation status and white marking characters might be due to shared ancestry more generally, we next performed phylogenetic analyses of the 13 breeds. Here, we based genomic comparisons either on sequences extracted from the deletion locus (10 kb centred on the *KIT* structural variant (SV) site; Chr6:70,051,190-70,061,190bp), or based on chromosome 6 sequence identity overall (see Methods). When considering only the locus harbouring the non-coding deletion, spotted breeds were found to cluster together (Fig 4.5), but these relationships could not be discerned when considering chromosome-wide sequence identity (Fig 4.S4). These findings support the hypothesis that the 6.9 kb deletion contributes toward the coat patterning characteristics that have been selected in these breeds, or at least the immediate locus that harbours this variant.

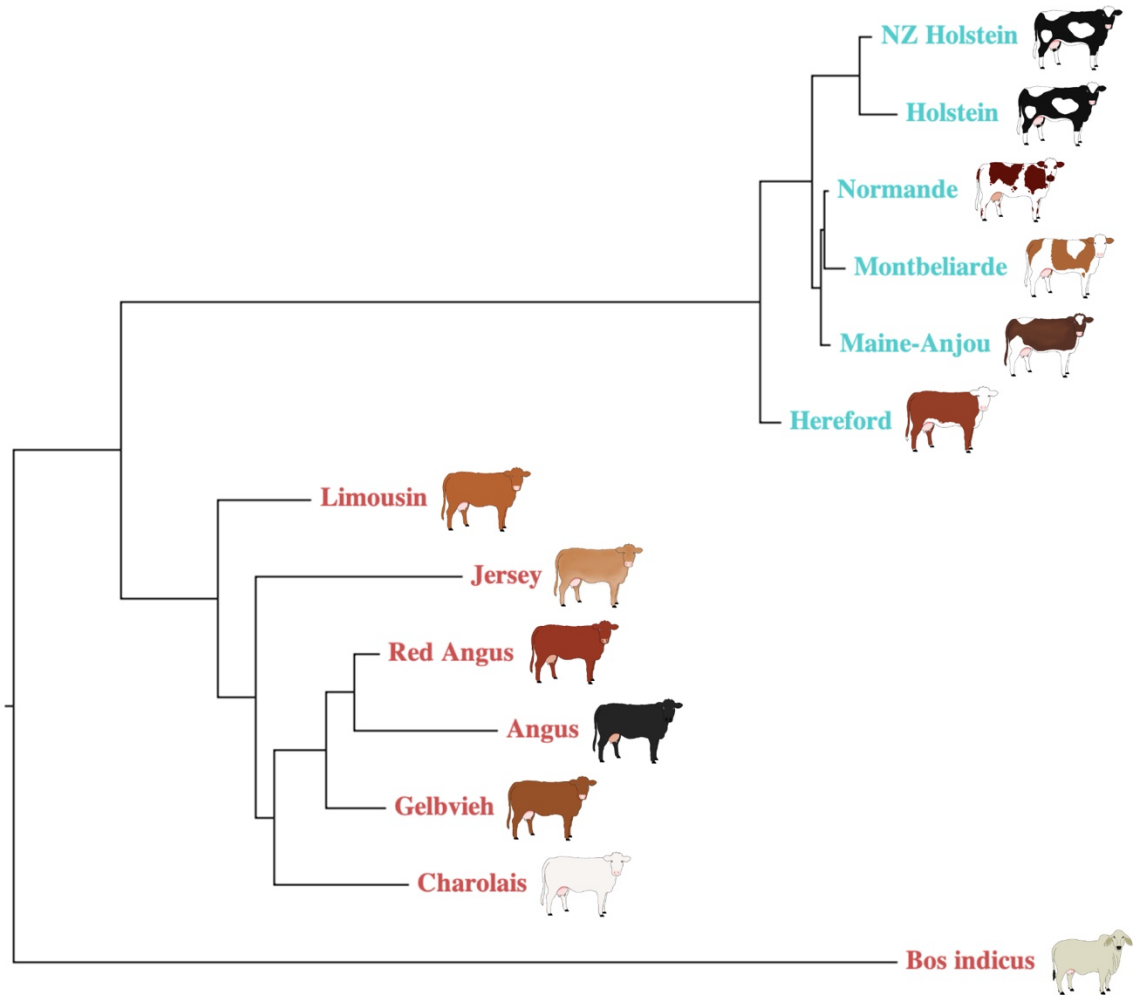


Figure 4.5 Mash-based phylogenetic tree for spotted (blue) and non-spotted (red) cattle across 10 kb from the bespoke chromosome 6 region Chr6:70,051,190-70,061,190bp incorporating the KIT SV. The tree was constructed using sketch sizes of s=1000, and k-mer sizes of k=21.

4.3.6 Epistasis at the *KIT* locus – *MITF* modulation of the Hereford 'white-face' trait

Several other breed-defining coat characteristics have also been mapped to the *KIT* locus in cattle, including the 'colour-sided' trait of Belgian Blue and White Park cattle (2,3), and the 'white-face' trait in Hereford cattle (4). Regarding the latter, we were aware of cases of incomplete penetrance for this otherwise dominantly inherited trait, where crosses of some Jerseys and Angus cattle produced calves with a broken, or

'splotchy' coloured face (Fig 4.6a). Since the novel *KIT* 6.9 kb deletion might be expected to interact with the *KIT* serial duplication proposed to underlie the Hereford white-face trait, we conducted an association analysis on 128 Hereford-cross calves that presented a mixture of splotchy and pure white faces (see Table 4.S5 for breed compositions). Genotype data were derived using a custom SNP-chip that contained probes for the 6.9 kb deletion mutation, where all calves were found to be homozygous or heterozygous for the 6.9 kb deletion allele. The *KIT* 6.9 kb deletion did not appear to be associated with the splotchy-face trait (smallest p=0.005 within 20 Mb of the SV for Chr6 g.64280973T>A). However, a strong association signal was detected on chromosome 22 (smallest $p=2.41\times10^{-20}$ for Chr22 g.31651404T>C; Fig 4.6b), presenting a near-equivalently associated *MITF* candidate causal mutation previously implicated in white spotting in Holstein-Friesian, Jersey, and crossbred cattle (Chr22 g.31651379A>G, rs209784468, $p=5.97\times10^{-20}$; (7)). When the rs209784468 SNP was fitted as a fixed effect in the association model, all significance at the chromosome 22 locus was removed, suggesting that the signal was likely representative of a single biallelic effect (smallest residual genome-wide p=0.00088 for Chr22 g.50885680C>G). The putative mutation appeared to operate in a Mendelian fashion, where nearly all calves carrying the 'G' allele showed splotchy faces, and nearly all pure white-faced animals were homozygous for the 'A' variant (Fig 4.6c; Table 4.S5). None of the Hereford crossed calves in our population were homozygous for the rs209784468 G allele, and the allele frequency for the G allele was 0.03 in our Hereford sequence dataset (N=35).

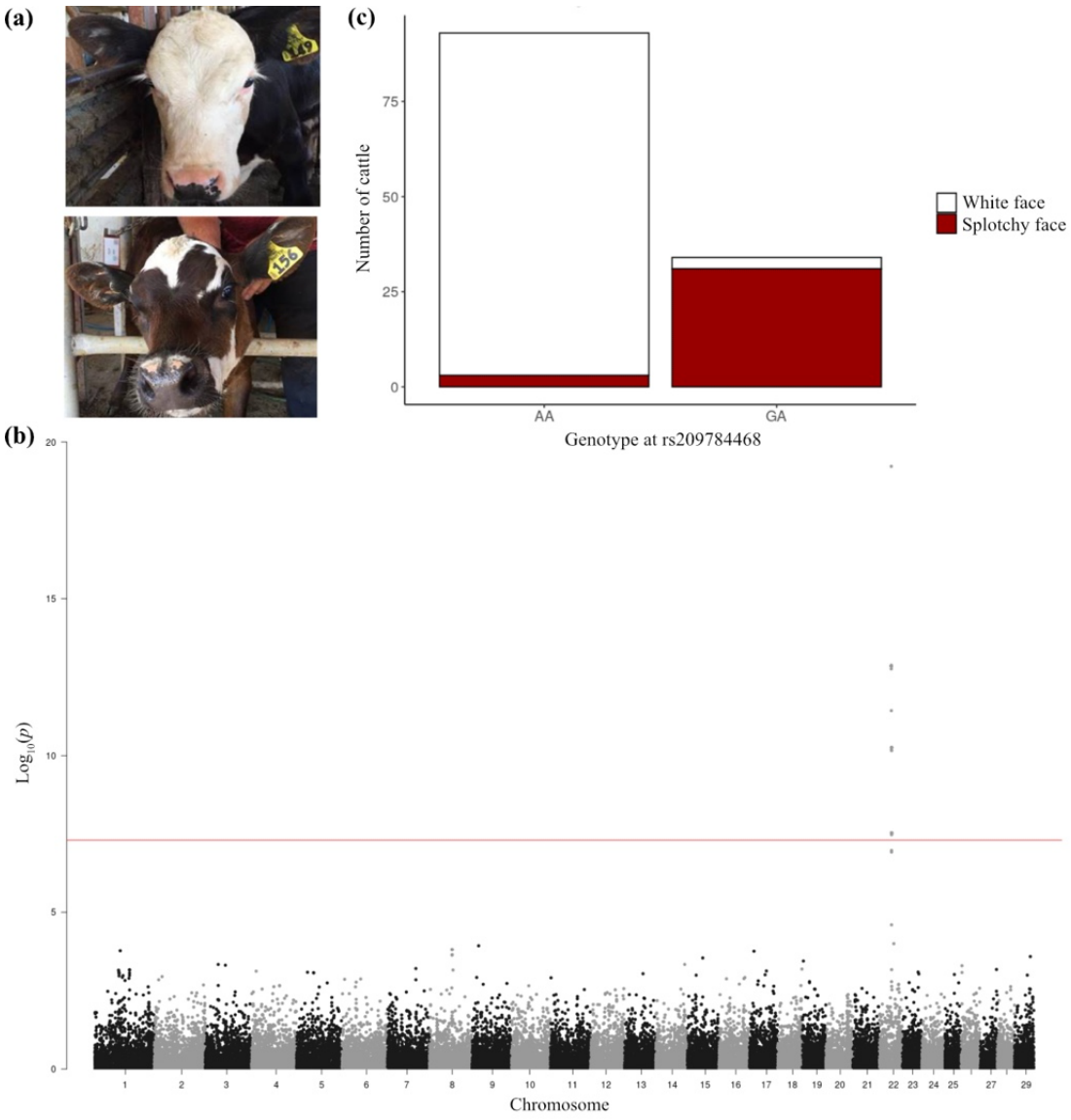


Figure 4.6 Association analysis results for the splotchy-face trait in Hereford-cross calves. (a) Image of white-faced calf (top) and splotchy-faced calf (bottom). (b) Manhattan plot based on the GWAS result for the splotchy-face trait. A single association signal is observed on chromosome 22, with the top variant mapping to Chr22 g.31651404T>C, $p=2.41\times10^{-20}$. The red line indicates the genome-wide significance threshold $p=5\times10^{-8}$. (c) The number of cattle with white faces (white) or splotchy faces (red) plotted by their genotype at candidate causal mutation rs209784468 (Chr22 g.31651379A>G).

4.4 Discussion

We present the first study to report a candidate causal mutation at the *KIT* locus in the white spotting trait characteristic of Holstein cattle. We provide strong bioinformatic and statistical evidence to implicate a novel 6,948bp stretch of sequence approximately 114 kb upstream of the *KIT* gene as an ancestral, evolutionarily conserved regulatory sequence that, when deleted, likely causes white spotting in Holstein cattle and other breeds. We also provide evidence of an epistatic interaction between two coat-colour mutations, a serial duplication upstream of the *KIT* gene previously hypothesised to cause the white-face trait in Hereford cattle (4), and a mutation in the *MITF* gene previously implicated in the proportion of white spotting (7). Heterozygous co-inheritance of these mutations appears sufficient to cause a broken white face (i.e., a splotchy face) in Hereford crossbreed calves.

Previous investigation of white-spotting in Holstein, Jersey, and crossbred cattle presented a peculiar, highly dispersed signal at the chromosome 6 locus, and automated structural variant analysis of this locus identified a 330bp duplication as moderately correlated with the tag SNPs at the chromosome 6 QTL (7). In this study, we performed a manual, more detailed sequence-based analysis of the 330bp duplication and observed alignment anomalies that implicated structural variation (SV) at a site approximately 114 kb upstream of the *KIT* gene, 400 kb away from the initially reported 330bp duplication. Long-range PCR and Oxford nanopore minION sequencing of both candidate SV regions found that the predicted duplication event was a sequencing artifact, but identified a novel 6.9 kb sequence that mapped between Chr6:70,052,697bp and Chr6:70,052,698bp on the ARS-UCD1.2 reference genome (12). This novel SV sequence is flanked by a 42bp sequence homologous with a bovine retrotransposable

element (RTE-BovB) long interspersed nuclear element (LINE), also present at the predicted duplication site (i.e., the 3' candidate SV site). Therefore, it seems likely that the BWA MEM alignment software mapped sequence reads from the 6.9 kb ancestral allele to the 3' candidate SV site due to homology across the 42bp sequence and proximity to their mate-paired reads, given the absence of the full length 6.9 kb allele from the Hereford-derived reference genome (13). Observation of homologous repeats either side of the deleted 6.9 kb sequence also gives some clue as to the likely genesis of this mutation, where the variant may have arisen due to non-allelic homologous recombination between sister chromatids during meiosis. This SV causing mechanism typically occurs when two regions that share sequence homology, but are not alleles of each other, cross-over due to a normal recombination event (14). In this instance, it is possible that the ancestral allele was deleted due to an unequal crossing-over event at one of the flanking repeat sites.

Bioinformatic analysis of the 6.9 kb sequence found an overlapping, highly conserved MITF transcription factor binding site, and genetic analyses of a population of 2,976 cattle with quantitative white coat spotting data showed that the SV wholly accounted for the white spotting QTL observed at the *KIT* locus. We hypothesise that deletion of the MITF transcription factor binding site reduces the transcriptional efficiency of *KIT* during bovine embryonic development. Consequently, melanocyte migration, proliferation and/or survival would likely be impaired, causing some regions of the coat to be devoid of functional melanocytes and pigmentation (i.e., cause white spotting on the coat). Interestingly, we found that the 6.9 kb deletion was present in every spotted breed we investigated, including New Zealand Holstein-Friesian, American and French Holstein, Normande, Montbeliarde, Maine-Anjou, Hereford, and some Bos indicus

cattle. Furthermore, the 6.9 kb deletion was absent in typically solid-coloured Angus, Limousin, and Gelbvieh cattle. These data suggest that either the mutation occurred in a very early ancestor of these cattle, and then was selected for in modern spotted breeds, and selected against in solid-coloured breeds, or that the deletion event has occurred several times throughout the evolution of modern cattle breeds and has persisted due to selection.

Hereford cattle have a breed-defining white face that is inherited in a dominant manner, but when crossed with some solid-coloured cattle (e.g., Angus or Jersey) their calves sometimes have a broken white-face, or 'splotchy' face. The novel 6.9 kb ancestral allele, associated with a solid coat colour, mapped approximately 42 kb away from the serial duplication proposed to be causal for the white-face trait (4) and therefore made an interesting positional candidate for the splotchy-face trait. Instead, we found that the splotchy-face trait is likely caused by an epistatic interaction between the white-face duplication and a candidate causal mutation in intron 2 (rs209784468) of the skin-expressed *MITF* gene isoform, previously implicated in the proportion of white spotting trait in Holstein-Friesian cattle (7). The G allele of rs209784468 was previously associated with reduced white spotting on the coat in Holstein-Friesian, Jersey and crossbreed cattle (7), and in this study it was associated with adulteration of the white-face trait. The white-face mutation is suspected to impair *KIT* expression during development and result in no pigmentation in the face, where one copy of this mutation is sufficient to result in the manifestation of this trait. It is somewhat surprising, that a single copy of the G allele at rs209784468 is sufficient to rescue melanocyte migration, proliferation and/or survival and cause breakthrough pigmentation of the face. Although no animals in our study were homozygous for the rs209784468 G allele, we speculate

that such cattle may manifest greater pigmentation on the face, or perhaps have wholly pigmented faces.

Our results extend previous genetic investigations into the proportion of white spotting in cattle by providing a compelling structural candidate mutation at the *KIT* locus. The 6.9 kb deletion allele is yet another example of how structural variation at the *KIT* locus may define breed-specific coat patterning traits in cattle. At the time of writing, work was underway with our collaborators at the University of Lancaster to create a murine cell-based mutation model to investigate whether the 6.9 kb deletion modulates *KIT* expression during development and influences melanoblast physiology. The results of these experiments will provide valuable insight into how *KIT* transcription can be regulated by structural variants and distal enhancers, and how these may modulate melanoblast physiology to create the array of striking coat patterning traits we observe in mammals.

4.5 Methods

4.5.1 Cattle populations

The cattle populations described in this paper represented several cohorts. Table 4.S6 summarizes breed and cohort information and the respective analyses performed in this study. Briefly, a cohort of previously described sequenced cattle (N=565) were used as the initial discovery dataset (15), and sequence data for an additional 562 cattle were later incorporated and used as an imputation reference dataset (total N=1,127 whole genome-sequenced cattle). These animals consisted of commercially farmed, purebred Holstein-Friesian (HF), purebred Jersey (J) or crossbred HF × J cattle of both sexes,

where ‘purebred’ animals were defined based on a breed proportion of $^{16}/_{16}$ from a four-generation pedigree. Blood samples from three HF bulls, three J bulls, and two Hereford bulls were used to derive DNA for PCR and long-read sequencing. Genotype and phenotype data were available for 2,976 cattle (7), where that population had an overlap of 499 animals with the initial genome-sequenced discovery dataset. DNA sequence data from Angus, Red Angus, Charolais, Limousin, Maine-Anjou, Montbeliarde, Normande, Gelbvieh, Bos indicus, Hereford, and Holstein cattle were downloaded in the form of fastq sequence files from the NCBI Sequence Read Achieve (N=332), and combined with purebred HF and J cattle sequence data from the discovery dataset for phylogenetic analyses. Genotype and phenotype data were available for 128 white-faced and splotchy-faced Angus × Hereford (N=21), and HF × J × Hereford (N=107) calves, derived as described in the following sections.

4.5.2 Whole-genome sequence and genotype data

Whole-genome sequencing and read-mapping were performed on a population of 1,127 cattle as previously described (16,17). Briefly, DNA samples for all cattle were sequenced using 100-bp paired-end reads on the Illumina HiSeq 2000 platform. Read mapping was initially performed on a previously published cohort of 565 cattle (15) using the ARS-UCD2.1 genome build (12) and BWA MEM (v0.7.17) software (13). These data were used as the discovery dataset and were later combined with an additional 562 cattle, sequenced after the discovery of the candidate structural variant. The larger dataset (N=1,127 cattle) were mapped to our bespoke reference genome (described in the ‘Creation of a structural variant-augmented reference genome’ section below) using BWA MEM, but due to data storage limitations, only mapped data from chromosome 6 was retained for each iteration. The mean mapped read depth across

chromosome 6 for this dataset was 15×. The methods outlined above were also used to map publicly available sequence data from an additional 332 cattle representative of a variety of spotted and non-spotted cattle breeds (Table 4.S5) to our bespoke reference genome, resulting in a mean mapped read depth of 13× for that dataset.

Microarray-based genotype data were available for the cattle used in the proportion of white spotting association analysis (N=2,976). These data were generated by GeneSeek (Lincoln, NE, USA), using a variety of platforms including the Geneseek GGPv1, GGPv2, GGPv3, GGP50k, Illumina BovineSNP50 or BovineHD 777k SNP chips, as previously described by Jivanji et al. (7). Step-wise imputation to sequence resolution genotypes was performed using Beagle 5.0 software (18), and has recently been described by Reynolds et al. (19). Variants were subsequently filtered for imputation quality (variants with a dosage R^2<0.7 were removed), and rare variants (variants with a homozygous alternate count ≤ 5 were removed) to avoid potentially spurious association.

Tissue samples were obtained from ear tissue biopsies from 128 white-faced and splotchy-faced Angus × Hereford and HF × J × Hereford calves, and DNA extraction was conducted at GeneMark (Hamilton, New Zealand), using the Qiagen BioSprint kit. Genotyping was conducted using the GeneMark 50kv1 (GMK50kv1) SNP chip, which contained nine custom probes designed to genotype the *KIT* SV.

4.5.3 Identification and genotyping of novel structural variation at the *KIT* locus

As part of our previous analyses, we reported a candidate structural variant that mapped downstream of the *KIT* gene to BTA6:72,060,120-72,060,450bp on the UMD3.1 bovine

reference build (7). Paired-end short-read sequence data from 565 HF, J and crossbred cattle were mapped to the ARS-UCD1.2 reference genome (12) using the BWA MEM (v0.7.17) (13), and this region was visualised using IGV software (20). The alignments were coloured by insert size and pair orientation in IGV, and view settings were adjusted to allow for identification of soft-clipped reads.

A grep-based method was used to genotype cattle for sequence anomalies observed between Chr6:70,396,258-70,396,788bp and Chr6:70,052,058-70,053,039bp. Twenty-nucleotide search strings that encompassed 10bp from the soft-clipped region, and 10bp from the reference consensus region (Table 4.S7) from both candidate structural variant sites were used to query raw FASTQ sequence files. These analyses were restricted to cattle samples that had a minimum average read depth of 10× coverage (N=152 cattle). A BLAST search against the ARS-UCD1.2 bovine reference genome confirmed each search string to be unique, with no sequence homology with any other region of the genome identified. The search strings that incorporated the soft-clipped reads were considered representative of the structural variant form, also referred to as the 'alternate' form, and the corresponding 20bp ARS-UCD1.2 sequence string represented the 'reference' form, with reverse complements for each search string also incorporated into these definitions. The proportion of alternate search-string matches observed at any one site (the number of alternate form matches/total matches detected) was used as a proxy for the candidate structural variant genotypes at the 3' and 5' sites. The correlation between the proportion of alternate search-string matches observed at any one site and the genotype at tag SNPs rs451683615 and rs463810013 were computed using the dplyr (v0.7.8) package in R (21).

4.5.4 High-molecular-weight DNA extraction

To characterise candidate structural variant junctions, 10 mL blood samples were obtained from three J bulls, three HF bulls, and two Hereford bulls, selected based on their genotypes at previously identified white spotting-associated tag SNPs rs451683615 and rs463810013 (Table 4.S2). Our DNA extraction protocol embedded mononuclear cells in an agarose matrix prior to cell lysis, aiming to prevent excessive sharing of the DNA and permit efficient long-range PCR. Samples were collected in heparin tubes and processed on the same day. Briefly, peripheral blood mononuclear cells were isolated from whole blood by a series of red blood cell lysis and centrifugation cycles. Pelleted white blood cells were gently resuspended in an appropriate volume of PBS to give a final concentration of 2×10^7 cells per ml and warmed to 37°C. The cell suspension for each sample was gently mixed with 2% low melting agarose in PBS in equal volumes and cast into 100 µl moulds. Once the agarose-cell suspension plugs had solidified, they were halved and incubated in lysis solution at 50°C for 48 hours. The 50 µl agarose plugs were washed three times with wash buffer and stored in wash buffer at 4°C until required. Each 50 µl agarose plug was expected to have had approximately 1×10^6 cells prior to cell lysis, equating to approximately 6,600 ng of DNA per plug.

4.5.5 Long-range PCR and minION sequencing of candidate structural variant sites

Genomic DNA was extracted from the 50 µl agarose plugs using the NucleoSpin Gel and PCR Clean-up kit. Briefly, agarose plugs were placed into a clean tube with 200 µl NTI buffer. The plugs were incubated for 5-10 minutes at 50°C and gently mixed every 2-3 minutes until the gel slice was completely dissolved. The sample was then loaded

into a NucleoSpin Gel and PCR Clean-up column and centrifuged at 11,000 × g for 30 seconds. The flowthrough was discarded, and the column was washed with 500 µl of NT3 buffer. Samples were centrifuged for 30 seconds at 11,000 × g, the flowthrough was discarded, and this step was repeated. The silica membrane was dried by being placed into a clean collection tube, and centrifuged for 1 minute at 11,000 × g, then incubating at 70°C for 2-3 minutes to remove excess ethanol. The DNA was eluted using 15 µl of water warmed to 70°C. The sample was centrifuged for 1 minute at 11,000 × g, and this step was repeated before the final DNA sample was obtained.

Primer pairs were designed to amplify a 6,337bp region at the 5' candidate structural variant site (Chr6:70,048,369-70,050,884bp) and a 4,749bp region at the 3' candidate structural variant site (Chr6:70,394,202-70,399,130bp; Table 4.S7), selected to target 'cleanly aligned' sequences based on visualisation of genome sequence alignments. A touch-down PCR was conducted using the KAPA LongRange PCR kit (KapaBiosystems). The initial denaturation step was conducted at 95°C for 30 seconds, followed by denaturation at the beginning of each PCR cycle at 95°C for 30 seconds. The annealing step started at 70°C for the first cycle, and dropped by 1°C each cycle until it reached 60°C. The annealing temperature remained at 60°C for the remaining 25 cycles. The extension steps for the PCR targeting the 5' candidate structural variant site were conducted at 68°C for 13 minutes each, and the extension steps for the PCR targeting the 3' candidate structural variant site were conducted at 72°C for 4 minutes. The PCR products were loaded and run on a 1% agarose gel for 60 minutes at 100 V to estimate amplicon size.

The PCR amplicons were purified using AMPure XP beads and then used to construct a sequencing library using the SQL-LSK109 kit (Oxford Nanopore Technologies) as per the manufacturer's instructions. The first library, targeting the 5' candidate structural variant site, was constructed using 700 ng of DNA from across the eight samples, loaded onto a FLO-MIN106 flow-cell (Oxford Nanopore Technologies). Double the amount of DNA was used for the two Jersey samples that had an ~13 kb amplicon (140 ng per sample) compared to other samples that had ~6 kb amplicons (70 ng per sample) to compensate for the size-based sequencing bias of the minION sequencer. Amplicons representing the 5' structural variant site were sequenced for 2 hours. The second library, targeting the 3' candidate structural variant site, was constructed using 700 ng of DNA from across the eight samples in equal amounts (87.5 ng per sample), loaded onto a FLO-MIN106 flow-cell and sequenced for 40 minutes. The sequencing depth per sample can be found in Table 4.S2.

4.5.6 Creation of a structural variant-augmented reference genome

Sequence reads from the minION sequencer were base-called using Guppy basecaller (v4.0.14) (22), with the samples then separated based on their barcodes by Guppy barcoder (v4.0.14), and subsequently aligned to the ARS-UCD1.2 reference genome (12) using minimap2 (v2.14) (23). A consensus sequence for the 5' targeted region amplicon incorporating the insertion (13,285bp) was constructed from the J1 long read data using Shasta (24), and the consensus sequence added to the ARS-UCD1.2 reference genome as an alternative contig. Short read sequence data from two cattle previously identified to have the structural variant from the grep-based method (see 'Identification and genotyping of novel structural variation at the *KIT* locus'), and two cattle that had the deleted allele, were aligned to the modified reference genome using

BWA MEM (v0.7.17) (13). The sequence reads aligned to the amplicon-derived consensus were manually inspected in IGV (20) and obvious and easily resolved errors in the consensus sequence were manually corrected.

The final consensus sequence representing the 5' structural variant region was searched against the ARS-UCD1.2 reference genome using BLAST (25), to identify the likely site of insertion. The ARS-UCD1.2 chromosome 6 reference sequence was split at the candidate insertion point using SAMtools (26), and the novel structural variant sequence was inserted ~114 kb upstream of the *KIT* gene (between Chr6:70,052,697bp and Chr6:70,052,698bp). Notably, the inserted sequence was flanked by blocks of sequence 42bp long, identical except for one base-pair (TGAACTTCCTGATGTT[G>C]AAGCTGGTTTTAGAAAAGGCAGA). The G-allele variant of the 42bp sequence mapped at the 3' end of the ancestral allele and appeared to be a fragment of a bovine retrotransposable element (RTE-BovB) long interspersed nuclear element (LINE). A BLAST of 42bp sequence revealed that it was observed 1,905 times across the ARS-UCD1.2 reference genome, including at the previous 3' candidate SV site, mapping to Chr6:70,396,679-70,396,718bp. Short read sequence data from the four cattle previously used to correct errors in the consensus sequence, were used to confirm the insertion site. Sequence alignments were visually inspected in IGV, adjusted, and remapped until soft-clipped reads were no longer observed across the breakpoints. All minION sequence data was also remapped to the bespoke reference genome using minimap2, and the SV region was visualized in IGV. A co-locating 850bp deletion was observed in the J2 sample, mapping to Chr6:70,054,845-70,055,695bp in the bespoke reference. This haplotype also differed from the long-form ancestral allele by 71 other polymorphic variants.

4.5.7 Genotyping the *KIT* structural variant

Sequence alignments representing HF (N=280), J (N=188), and HF x J (N=659) cattle and the 332 other spotted and non-spotted breeds mapped to the bespoke reference genome were used to genotype the *KIT* SV. CNVnator (v0.3.3) (27) was used to predict the presence of the 6.9 kb deletion (Chr6:70,052,698-70,059,646bp) and the 850bp deletion (Chr6:70,054,845-70,055,695bp) based on average read-depth in these regions in their sequence context. The CNVnator predicted copy number calls across the 6.9 kb deletion and 850bp deletion alleles were confirmed or adjusted based on visual inspection of plots generated by Samplot (28) that summarised read-depth and split-read information across the SV site. If deletion status remained ambiguous, sequence reads were manually inspected in IGV to confirm the SV state.

4.5.8 Variant calling and imputation

Variant calling was conducted on sequence data from 548 Bos indicus, Angus, Red Angus, Charolais, Limousin, Maine-Anjou, Montbeliarde, Normande, Gelbvieh, Holstein, Hereford, NZ HF and NZ J cattle aligned to the bespoke chromosome 6 using the Genome Analysis Toolkit (GATK) HaplotypeCaller (v4.1.8.1) software (29). Default parameters were used to variant call all cattle with homozygous genotypes across the *KIT* SV (regardless of SV state). Sequence alignments for the remaining cattle that were either hemizygous for the 850bp deletion, or for the 6.9 kb deletion, were split at the SV junctions using SAMtools (26) so that the SV sequence could be interrogated separately. Variant calling across the sequence representing a hemizygous deletion state was conducted using HaplotypeCaller (v4.1.8.1) software (29) with ploidy set to one. The remaining segments of sequence aligned to chromosome 6 were variant

called using default parameters, and the resulting variant called files were concatenated using BCFtools (26).

The GATK HaplotypeCaller (v4.1.8.1) software (29) was used for variant calling on 280 HF, 188 J, and 659 crossbred sequenced cattle mapped to the ARS-UCD1.2 chromosome 6 reference. For the purpose of downstream imputation and association analyses, the *KIT* SV genotype (as established by methods described in the 'Genotyping the KIT structural variant' section) was summarised as one representative triallelic variant, and manually added to the variant call format (VCF) file at the non-variant Chr6:70,057,008bp site. The sequenced cohort, composed of 1,127 animals, was used as a reference dataset to impute the triallelic variant SV genotype into the phenotyped population (i.e., animals with white spotting data). The reference genome was phased using Beagle 5.1 (18), and imputation was conducted across Chr6:60-80 Mb. The 850bp deletion was imputed into the phenotyped population with an allelic R^2 of 1, and the 6.9 kb deletion imputed with an allelic R^2 of 0.99, suggesting all alleles were imputed accurately.

4.5.9 Phenotypes, population structure adjustments and association analyses

White spotting phenotypic data was available for 2,967 NZ dairy cattle (7). This population was divided into four groups based on their genotype at the *KIT* SV site, to test the association between the alternative SV states and the proportion of white spotting. The first group included animals with the long-form ancestral allele or the 6.9 kb deletion (N=2,596), the second group included animals with the long-form ancestral allele or the 850bp deletion (N=85), and the third group included animals with the 6.9 kb deletion or the 850bp deletion (N=2,507). The last group combined animals that had

either the long-form ancestral allele, or the 850bp deletion allele for contrast with those that had the 6.9 kb deletion allele (N=2,967). A breakdown of these groups by SV genotype can be found in Table 4.S9. Genomic relationship matrices (GRM) were generated using GCTA (v1.93.2beta) (30) to address population stratification due to breed and relatedness in the association models. These GRM were calculated based on a subset of 19,354 markers from the Illumina Bovine SNP50 platform, having been filtered based on minor allele frequency (those with a minor allele frequency <0.02 were removed), deviation from Hardy Weinberg equilibrium (those with a P<0.15 were removed), missing genotype rates (those with a genotyping rate <0.01 were removed), and high linkage disequilibrium (LD) with another marker on the panel (those with pairwise R^2>0.9 were removed). To avoid fitting variants that were in LD with the *KIT* SV genotype, markers from chromosome 6 were also excluded. Association analysis was performed on 152,072 sequence-resolution markers mapping between Chr:60-80 Mb. The GCTA software was used to conduct the mixed linear model-based association analysis (MLMA), which incorporates the GRM described above.

Face colour (white or splotchy; Figure 4.7a) was reported by farmers, or scored using photographs taken by farmers, on 21 Angus × Hereford calves and 107 HF × J × Hereford calves. A 'leave one chromosome out' approach was used to calculate 29 GRM, where each GRM lacked one autosome to avoid double fitting when testing the effect of candidate variants on that excluded autosome, as previously described by Jivanji et al. (7). The GRM were calculated using variants from the GMK50kv1 platform used to genotype these cattle (N=21,159 variants), with quality filtering applied as described above for the proportion of white spotting analyses. The GCTA

software was used to perform MLMA using the GMK50kv1 resolution genotype data.

For significance testing, we used a *p*-value of 5×10^{-8} as the significance threshold to account for multiple hypothesis testing. To assess whether candidate mutations of interest wholly explained association signals for the white spotting and face colour analyses, variant genotypes were fitted as fixed effects as part of subsequent association analyses.

4.5.10 Phylogenetic analysis

Consensus sequences for chromosome 6 were generated for 13 spotted and non-spotted breeds (see Table 4.S4) mapped to the bespoke reference genome, using BCFtools consensus (26). A 10 kb region encompassing the *KIT* SV (Chr6:70,051,190-70,061,190bp) was extracted from each consensus sequence file. Within this region, there were, on average, 99 polymorphic variants across all breeds. The 'mash sketch' function from Mash (v2.3) (31) was then applied to convert each sequence into MinHash sketches with the default *k*-mer (*k*=21) and sketch sizes (s=1000). Pair-wise Mash distances were calculated using the 'mash dist' function. A neighbour-joining tree was constructed from the distance matrix using QuickTree (v2.5) (32), and visualized in FigTree (v.1.4.4). These methods were then applied to the larger, whole chromosome 6 consensus sequences, which had an average of 797,267 polymorphic variants across breeds.

4.6 Declarations

4.6.1 Ethics approval

All animal experiments were conducted in strict accordance with the rules and guidelines outlined in the New Zealand Animal Welfare Act 1999. Approval was sought for genotyping and trait scoring of the Hereford cross calves used in this study, and subsequently approved by the AgResearch Animal Ethics Committee, Hamilton, New Zealand (approval AEC 15236). The scoring procedures for white-spotting that were not based on pre-existing photographs, were approved by the AgResearch Animal Ethics Committee (approval AEC 14090). All other data were generated as part of routine commercial activities that are outside the scope of those requiring formal committee assessment or ethical approval (as defined by the above guidelines).

4.6.2 Consent for publication

Not applicable.

4.6.3 Availability of data and material

The discovery dataset comprised 565 whole genome-sequenced cattle as previously published by Reynolds et al. (15) (https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP276986), and the phenotypic and genotype data used in the proportion of white spotting association analysis has been previously published by Jivanji et al. (7) (https://doi.org/10.5061/dryad .tqjq2bvtf). All minION sequence data reported in this paper, and bam files representative of the 30 Mb region on chromosome 6 inclusive of the *KIT* SV ancestral allele from larger sequenced dataset (N=1,127) will be uploaded to the Dryad database upon manuscript acceptance.

Genotype and phenotype data representing the white-face and splotchy-face calves were uploaded under the same submission ID. Additional sequence data are available on the NCBI Sequence Read Archive (https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=ERP010431, https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP017441, and https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP245473).

4.6.4 Funding

This work was supported by the Ministry of Business, Innovation and Employment Endeavor Fund CONT-57639-ENDRP-LIC (Wellington, New Zealand), and Livestock Improvement Corporation (LIC; Hamilton, New Zealand). External funders had no role in the design of the experiment, the analysis or interpretation of the data, or writing of the manuscript.

4.6.5 Competing interests

AY, CH, CG, CC, GW, LM, TF, TM, YW, RS, and ML are employees of Livestock Improvement Corporation, a commercial provider of bovine germplasm. The remaining authors declare that they have no competing interests.

4.6.6 Authors' contributions

SJ performed most of the bioinformatic and statistical analyses with help from MK and CH; SJ, AY, CG, CC, GW, LM, TF, and ML were involved in data collection; YW and CC conducted sequence imputation; SJ and EW conducted most of the lab work with help from EM, JM, and RM; SJ, ML, RS, and RM, conceived the study and

experiments; SJ, ML, and TM secured funding for the project; RS, DG, RS, and ML were involved in supervision of the project; SJ and ML wrote the manuscript.

4.7 Acknowledgements

The authors would like to acknowledge Massey University, and Livestock Improvement Corporation (LIC) for their support in this research, the University of Auckland for access to laboratory resources, and the New Zealand eScience Infrastructure (NeSI) for providing the computational resources required for the analyses described here. The authors would also like to acknowledge the farm owners who took part in our study, Lorna McNaughton and Tony Fransen, for aiding in tissue biopsy collection and trait scoring.

4.8 References

1. Olson TA. Genetics of colour variation. In: The genetics of cattle. 1999. p. 33–53.
2. Durkin K, Coppieters W, Drögüller C, Ahariz N, Cambisano N, Druet T, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature. 2012;482(7383):81–4.
3. Brenig B, Beck J, Floren C, Bornemann-Kolatzki K, Wiedemann I, Hennecke S, et al. Molecular genetics of coat colour variations in White Galloway and White Park cattle. Anim Genet. 2013;44(4):450–3.
4. Whitacre L. Structural variation at the KIT locus is responsible for the piebald phenotype in Hereford and Simmental cattle. University of Missouri--Columbia; 2014.

5. Awasthi Mishra N, Drögemüller C, Jagannathan V, Keller I, Wüthrich D, Bruggmann R, et al. A structural variant in the 5'-flanking region of the TWIST2 gene affects melanocyte development in belted cattle. PLoS One. 2017 Jun 28;12(6):e0180170.

6. Georges M, Charlier C, Hayes B. Harnessing genomic information for livestock improvement. Nat Rev. 2019;20(3):135–56.

7. Jivanji S, Worth G, Lopdell TJ, Yeates A, Couldrey C, Reynolds E, et al. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. Genet Sel Evol. 2019 Nov 8;51(1):1–18.

8. Artesi M, Tamma N, Deckers M, Karim L, Coppieters W, Van den Broeke A, et al. Colour-sidedness in Gloucester cattle is associated with a complex structural variant impacting regulatory elements downstream of *KIT*. Anim Genet. 2020 Jun 12;51(3):461–5.

9. Küttel L, Letko A, Häfliger IM, Signer-Hasler H, Joller S, Hirsbrunner G, et al. A complex structural variant at the KIT locus in cattle with the Pinzgauer spotting pattern. Anim Genet. 2019;50(5):423–9.

10. Kawakami A, Fisher DE. The master role of microphthalmia-associated transcription factor in melanocyte and melanoma biology. Lab Investig. 2017;97(6):649–56.

11. Porter V. Cattle - a handbook to the breeds of the world. Porter V, editor. Cattle - a handbook to the breeds of the world. Marlborough: Crowood Press; 2007.

12. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9(3):giaa021.

13. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler

transform. Bioinformatics. 2009;25(14):1754–60.

14. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathog 2008 11. 2008 Nov 3;1(1):1–17.

15. Reynolds EGM, Neeley C, Lopdell TJ, Keehan M, Dittmer K, Harland CS, et al. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. Nat Genet. 2021;53(7):949–54.

16. Littlejohn MD, Henty KM, Tiplady K, Johnson T, Harland C, Lopdell T, et al. Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. Nat Commun. 2014;5:1–8.

17. Littlejohn MD, Tiplady K, Fink TA, Lehnert K, Lopdell T, Johnson T, et al. Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. Sci Rep. 2016;6:25376.

18. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. Am J Hum Genet. 2018 Sep 6;103(3):338–48.

19. Reynolds EG, Lopdell TJ, Wang Y, Tiplady K, Harland C, Johnson T, et al. Non-additive QTL mapping of lactatReynolds, E. G. et al. (2021) 'Non-additive QTL mapping of lactation traits in 124,000 sequence-imputed cattle reveals novel recessive loci', bioRxiv. Cold Spring Harbor Laboratory, p. 2021.08.30.457863. doi: 10.1101/2021. bioRxiv. 2021 Sep 1;2021.08.30.457863.

20. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013 Mar 1;14(2):178–92.

21. Wickham H, François R, Henry L, Müller K. A Grammar of Data Manipulation [R package dplyr version 1.0.7]. Comprehensive R Archive Network (CRAN);

2021.

22. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019 Jun 24;20(1):129.

23. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.

24. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol 2020 389. 2020 May 4;38(9):1044–53.

25. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 2013;41(W1):W29–33.

26. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Jan 29;10(2):1–4.

27. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21(6):974–84.

28. Belyeu JR, Chowdhury M, Brown J, Pedersen BS, Cormier MJ, Quinlan AR, et al. Samplot: a platform for structural variant visual validation and automated filtering. Genome Biol 2021 221. 2021 May 25;22(1):1–13.

29. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GA Van der, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017 Jul 24;201178.

30. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 2011 Jan 7;88(1):76–82.

31. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al.

Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016 Jun 20;17(1):1–14.

32. Howe K, Bateman A, Durbin R. QuickTree: building huge Neighbour-Joining trees ofprotein sequences. Bioinformatics. 2002 Nov 1;18(11):1546–7.

4.9 Appendix

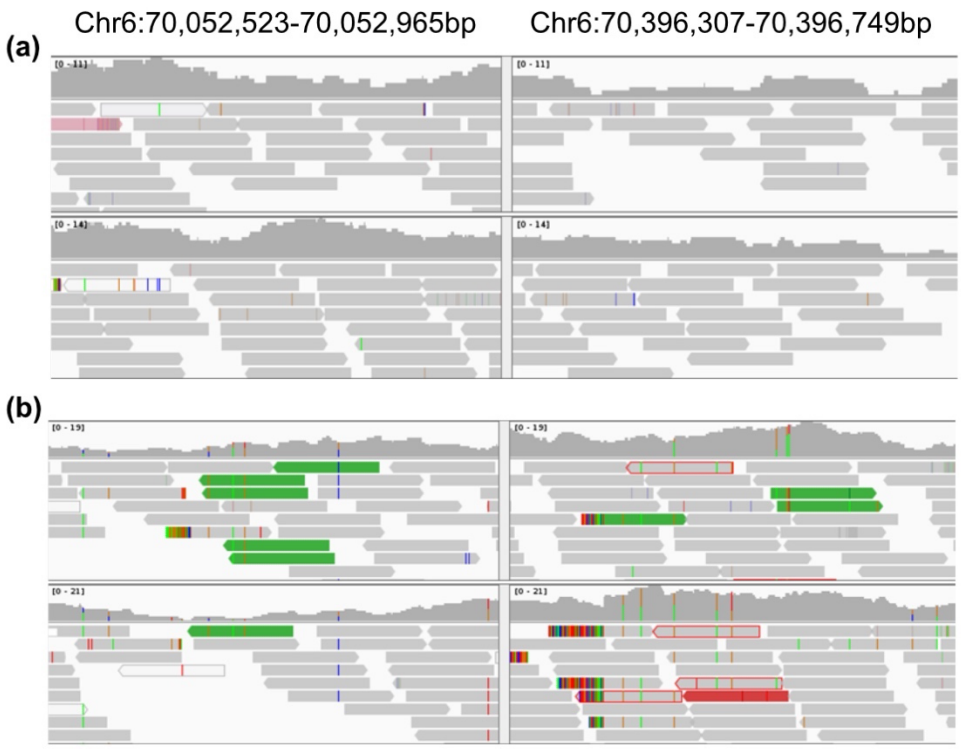


Figure 4.S1 Sequence alignments across candidate structural variant sites Chr6:70,052,523-70,052,956bp and Chr6:70,369,307-70,396,749bp. (a) Sequence alignments from two cattle (one per row) are visualised to demonstrate concordant mate-paired reads. Paired reads map pointing towards each other, where the forward (F) read in the mate-pair is expected to map in a 5' to 3' orientation, and the reverse read is expected to map in a 3' to 5' orientation (pair orientation = F1R2), with an insert size of 150bp. Read-pairs that meet these criteria appear grey. (b) The sequence reads highlighted in green are mate-paired reads that map 400 kb apart in a discordant R1F2 pair orientation. The string of bases within a sequence read that do not match the reference sequence at the mapped position, referred to as soft-clipped reads, are highlighted as blocks of coloured bases at the 5' end of both candidate structural variant regions. Reads highlighted with a red outline indicate reads where their mate-paired sequence read had not been mapped.

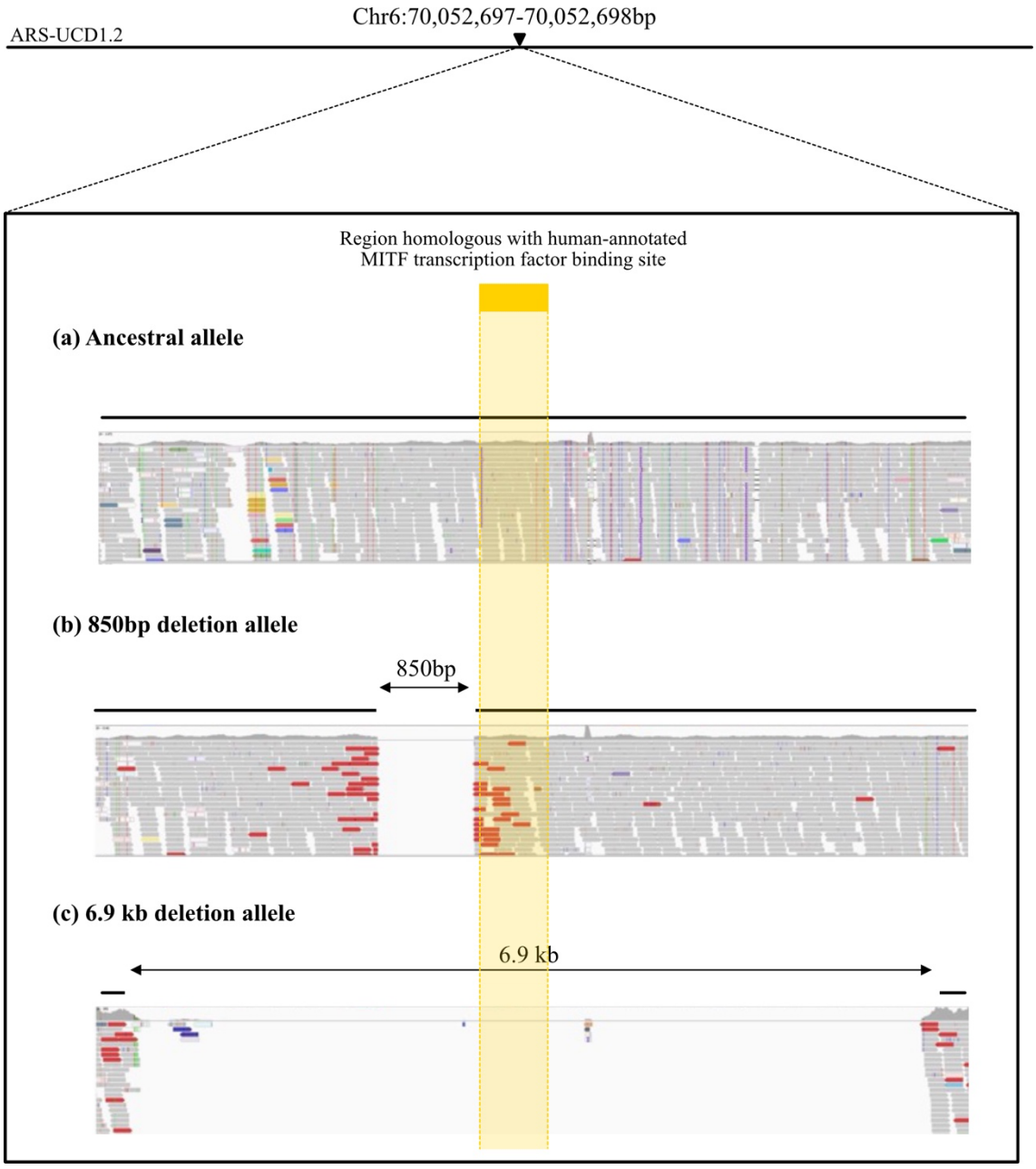


Figure 4.S2 Three possible KIT structural variant states identified between ARS-UCD1.2 Chr6:70,052,679bp and Chr6:70,052,698bp, with possible MITF transcription factor binding site highlighted (yellow). The (a) ancestral allele is a 6.9 kb insertion relative to the ARS-UCD1.2 reference genome, is present in solid-coloured animals, and overlaps a predicted MITF transcription factor binding site. The (b) 850bp deletion is also an insertion relative to the ARS-UCD1.2 reference genome and overlaps the predicted MITF transcription factor binding site. The (c) 6.9 kb deletion allele represents the ARS-UCD1.2 reference genome and is considered a deletion relative to the ancestral allele. This structural variant does not contain the MITF transcription factor binding site.

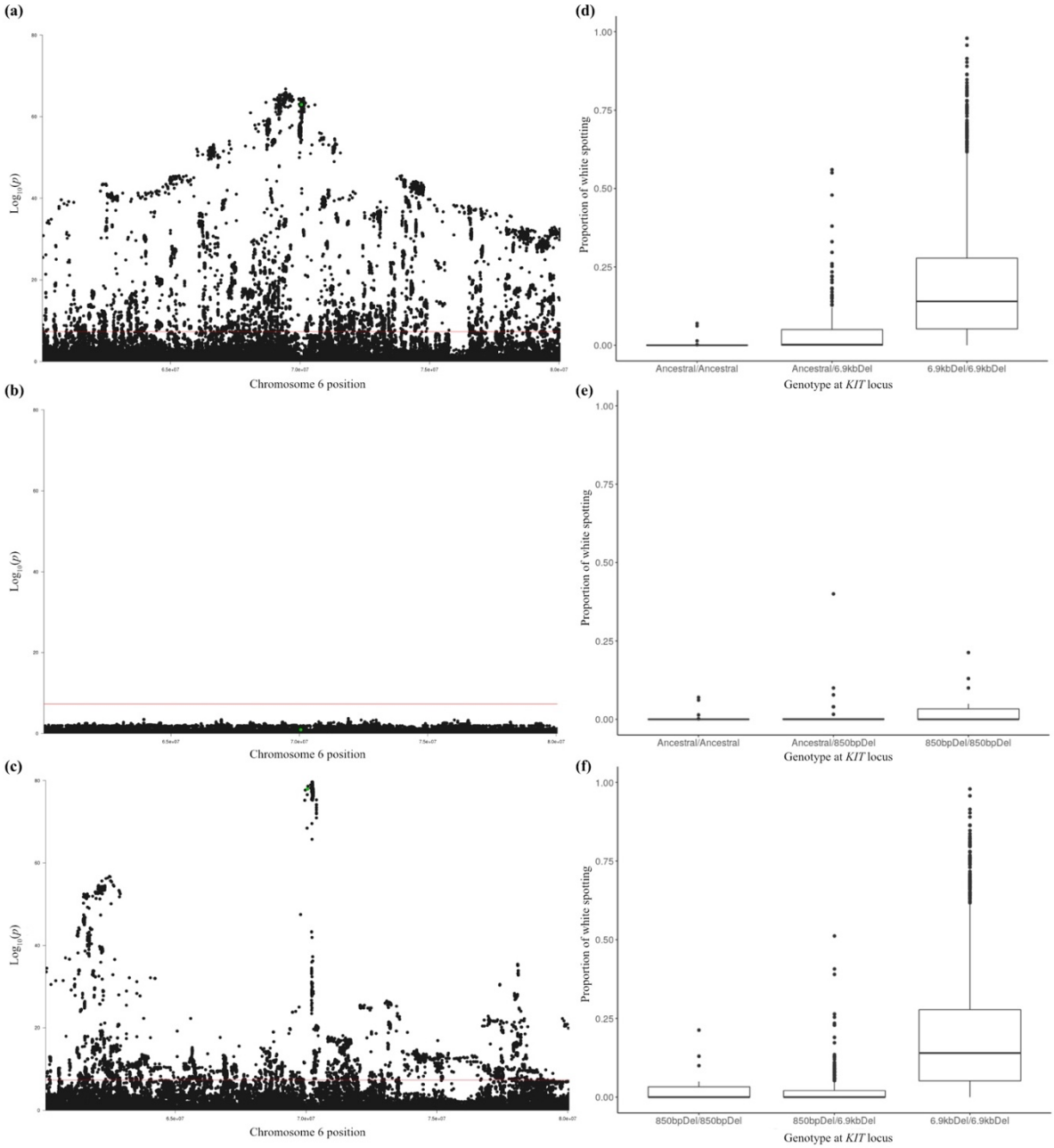


Figure 4.S3 Association and effect sizes of homozygotes of the three alternative KIT SV states on the proportion of white spotting. (a) Manhattan plot based on the association analysis results for the proportion of white spotting in cattle with either the ancestral or 6.9 kb deletion alleles. The variant representing the KIT SV (highlighted in green) has a p-value of 1.04×10^{-63}. (b) No association signal is seen in the Manhattan plot based on the association analysis results for the proportion of white spotting in cattle with either the ancestral or 850bp deletion alleles, and (c) Manhattan plot based on the association analysis results for the proportion of white spotting in cattle with either the 6.9 kb or 850bp deletion allele. The variant representing the KIT SV (highlighted in green) has a p-value of $p=1.25\times10^{-78}$. The red line indicates the significance threshold $p=5\times10^{-8}$. The boxplots show the proportion of white spotting based on genotype at the KIT SV for cattle (d) with the ancestral or 6.9 kb deletion allele, (e) with the ancestral or 850bp deletion allele, and (f) 850bp or 6.9 kb deletion alleles.

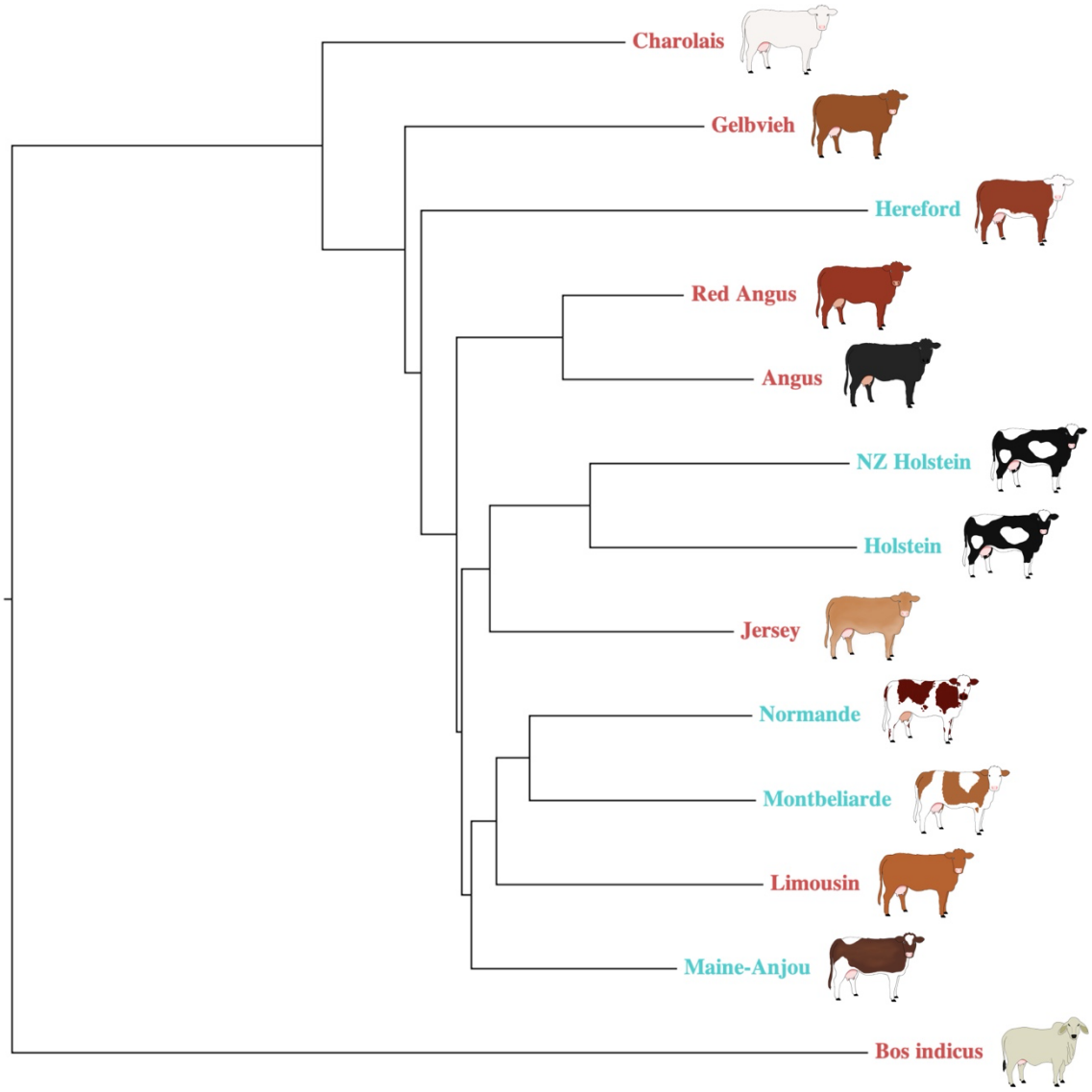


Figure 4.S4 Mash-based phylogenetic tree for spotted (blue) and non-spotted (red) cattle across across chromosome 6 constructed using sketch sizes of s=1000, and k-mer sizes of k=21.

Table 4.S1 Correlation between inferred structural variant genotypes at Chr6:70,052,523-70,052,956bp (5' site) and Chr6:70,369,307-70,396,749bp (3' site) and the proportion of white spotting tag SNPs rs451683615 and rs463810013.

	rs451683615	rs463810013
3' site genotyped with CNVnator*	0.22	0.43
3' site genotyped with grep-based method	0.77	0.79
5' site genotyped with grep-based method	0.80	0.87

*Previously reported by Jivanji et al. (7)

Table 4.S2 Additional information on bulls used for long-range PCR and Oxford nanopore minION sequencing targeting the 5' site (Chr6:70,048,369-70,050,884bp) and the 3' site (Chr6:70,394,202-70,399,130bp).

Animal	**Genotype at rs451683615**	**Genotype at rs463810013**	**PCR amplicon size**		**Average minION sequencing depth (×)**	
			5' site	3' site	5' site	3' site
Jersey 1	GG	TT	13,285bp	4,749bp	5,091	6,822
Jersey 2	GG	TT	13,285bp	4,749bp	4,848	5,098
Jersey 3	GG	CT	13,285bp/6,337bp	4,749bp	4,799	7,729
Holstein-Friesian 1	AA	CC	6,337bp	4,749bp	7,587	5,711
Holstein-Friesian 2	AA	CC	6,337bp	4,749bp	6,538	6,925
Holstein-Friesian 3	AA	CC	6,337bp	4,749bp	4,487	5,955
Hereford 1	AA	CC	6,337bp	4,749bp	6,769	5,127
Hereford 2	AA	CC	6,337bp	4,749bp	7,974	5,164

Table 4.S3 Structural variant state frequencies in 765 purebred Holstein-Friesian cattle and 387 Jersey cattle derived from reference population and the imputed dataset.

Breed	Ancestral allele	850bp deletion allele	6.9 kb deletion allele
Holstein-Friesian	0	0	1
Jersey	0.62	0.13	0.25

Table 4.S4 KIT structural variant frequencies in spotted and non-spotted cattle breeds.

	Ancestral allele	**850bp deletion**	**6.9 kb deletion**
Angus	0.92	0.08	0
Red Angus	0.78	0.19	0.03
Charolais	0.78	0.12	0.1
Limousin	0.75	0.25	0
Maine-Anjou	0	0	1
Montbeliarde	0	0	1
Normande	0	0	1

Gelbvieh	0.78	0.22	0
Bos indicus	0.67	0	0.33
Holstein	0	0	1
Hereford	0.03	0	0.97
NZ Holstein-Friesian	0	0	1
NZ Jersey	0.62	0.13	0.25

Table 4.S5 Number of calves included in splotchy face versus white face association analysis by breed composition, phenotype, and genotype at MITF candidate causal mutation Chr22 g.31651379A>G.

	Angus × Hereford	Holstein-Friesian × Jersey × Hereford	Genotype at Chr22 g.31651379A>G	
			AA	AG
White face	10	84	90	4
Splotchy face	11	23	3	31

Table 4.S6 Description of cattle populations used for analyses in this study, their sequencing or genotyping platforms and data availability.

Analysis	Population size	Number of cattle per breed		Sequencing/ genotyping platform	Data availability
Discovery dataset	565	Holstein-Friesian	116	Illumina HiSeq 2000 paired end sequencing	NCBI Sequence Read Archive accession identifier: SRP276986; previously described by Reynolds et al. (15)
		Jersey	95		
		Holstein-Friesian × Jersey	354		
Imputation reference dataset	1,127*	Holstein-Friesian	280	Illumina HiSeq 2000 paired end sequencing	To be uploaded upon paper acceptance
		Jersey	188		
		Holstein-Friesian × Jersey	659		
KIT SV lab-based characterisation	8	Holstein-Friesian	3	Oxford nanopore minION	To be uploaded upon paper acceptance
		Jersey	3		
		Hereford	2		
Proportion white association analysis	2,976**	Holstein-Friesian	592	Imputed to sequence	https://doi.org/10.5061/dryad.tqjq2bvtf; previously described by Jivanji et al. (7)
		Jersey	274		
		Holstein-Friesian × Jersey	2,110		

Phylogenetic comparison	548	Angus	82	Illumina HiSeq 2000 paired end sequencing & Illumina HiSeq 4000 paired end sequencing	NCBI Sequence Read Archive accession identifiers: ERP010431, SRP017441, and SRP245473 & NCBI Sequence Read Archive accession identifier: SRP276986; previously described by Reynolds et al. (15)
		Red Angus	29		
		Charolais	29		
		Limousin	10		
		Maine-Anjou	7		
		Montbeliarde	17		
		Normande	17		
		Gelbvieh	30		
		Bos indicus	3		
		Holstein	73		
		Hereford	35		
		NZ Holstein-Friesian	120***		
		NZ Jersey	96***		

Splotchy face association analysis	128	Angus × Hereford Holstein-Friesian × Jersey × Hereford	21 107	Illumina 50k SNP Chip	To be uploaded upon paper acceptance

*Includes all cattle from the discovery dataset

**Includes 499 cattle from the discovery dataset

***Derived from the discovery dataset

Table 4.S7 Twenty-nucleotide search strings used to initially genotype cattle for the 5' (Chr6:70,052,058-70,053,039bp) and 3' (Chr6:70,396,258-70,396,788bp) candidate structural variant sites.

	5' structural variant site	3' structural variant site
Search string with split reads	GGTCGCATTCATAGAAACATAGAAACTAGGTGAAGTGTGT	GAGGGGATATATGTATACCGTGAACTTCCTGATGTTGAAG
Reverse complement	ACACACTTCACCTAGTTTCTATGTTTCTATGAATGCGACC	CTTCAACATCAGGAAGTTCACGGTATACATATATCCCCTC
Reference	AGAAAGATGTTAACAGTAGGAGAAACTAGGTGAAGTGTGT	CTTCAGCAATAAGTGAACCATGAACTTCCTGATGTTCAAG
Reverse complement	ACACACTTCACCTAGTTTCTCCTACTGTTAACATCTTTCT	CTTGAACATCAGGAAGTTCATGGTTCACTTATTGCTGAAG

Table 4.S8 Long-range PCR primer sequences expected amplicon sizes, and targeted regions for the long-range PCRs conducted across the 5' and 3' candidate structural variant sites.

Targeted region (based on ARS-UCD1.2 reference genome)	**Forward primer sequence**	**Reverse primer sequence**	**Expected amplicon size based on reference**
Chr6:70,048,369-70,050,884bp	CTGAAGAAGGTGGGAAGGGTTT	CTGGAGAGAGGAGAGTGCAATG	6,337bp
Chr6:70,394,202-70,399,130bp	CTTGGGATCTAACACAGGCCAT	CAGCACAGATTCAGCCACATCT	4,749bp

Table 4.S9 Genotype groups used for association analyses and number of cattle per genotype within these groups.

Genotype group kept for association analysis	Homozygous allele 1*	Heterozygous	Homozygous allele 2*
Ancestral allele/6.9 kb deletion	23	399	2,174
Ancestral allele/850bp deletion	23	38	24
6.9 kb deletion/850bp deletion	2,174	309	24
Ancestral allele + 850bp deletion/6.9 kb deletion	85	708	2,174

* 'Allele 1' refers to the first SV state listed under 'Genotype group' for each group, followed by the SV state referred to as 'allele 2'.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

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:	Swati Jivanji
Name/title of Primary Supervisor:	Dorian Garrick
In which chapter is the manuscript /published work:	Chapter Four

Please select one of the following three options:

○ The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

○ The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

◉ It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Swati Jivanji
Date:	25/11/2021
Primary Supervisor's Signature:	Dorian Garrick
Date:	25 Nov 2021

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.

CHAPTER FIVE

Genome-wide association study of UV-protectant ambilateral circumocular pigmentation in Hereford cattle

IN PRESS AT THE NEW ZEALAND JOURNAL OF ANIMAL SCIENCE AND PRODUCTION

Genome-wide association study of UV-protectant ambilateral circumocular pigmentation in Hereford cattle

Jivanji S[1*], Kosch T[1], Littlejohn M[1] and Garrick D[1]

[1] School of Agriculture and Environment, Massey University, Palmerston North, New Zealand

[*]*Corresponding author*

Email: swati.jivanji.1@uni.massey.ac.nz

5.1 Abstract

Hereford cattle have a dominantly inherited white face that is characteristic of their breed. A small proportion of Hereford cattle present with regions of pigmentation around the eyes, known as ambilateral circumocular pigmentation (ACOP), which is hypothesised to provide UV-protectant properties to the underlying tissue. ACOP has been associated with reduced incidence of two diseases, bovine ocular squamous cell carcinoma, and bovine infectious keratoconjunctivitis, making it a desirable trait for selection. We conducted a genome-wide association study for ACOP in 605 American Hereford cattle using 43,789 genetic markers, and investigated the phenotypic association between face-colour and ACOP. The most highly associated variant from

our genome-wide association study mapped to Chr6 g.71059814G>A and explains most of the variation for this trait. This variant falls in close proximity to the well-known coat-colour patterning gene *KIT*, suggesting it may be the causal gene for this trait. Our results suggest that ACOP is not directly influenced by face-colour, and may be influenced by non-additive genetic effects. These results provide an opportunity to select for Hereford cattle with ACOP, enhancing welfare, and reducing the incidence of bovine infectious keratoconjunctivitis and bovine ocular squamous-cell carcinoma.

Keywords: Hereford cattle; ambilateral circumocular pigmentation; ocular squamous cell carcinoma; infectious keratoconjunctivitis; GWAS; pigmentation traits

5.2 Introduction

Hereford cattle are a popular beef breed easily identifiable by their dominantly inherited white face, a striking trait actively selected for by pedigree breeders as an indicator of 'true breeding'. The white-face trait manifests due to a complete lack of pigmentation (melanin), attributable to the lack of functional pigment cells (melanocytes) occupying the skin and hair within the face region. Melanin functions as an antioxidant and radical scavenger to protect the skin and underlying tissues against damage from ultraviolet (UV) radiation, so although the white-face trait is cosmetically attractive to breed enthusiasts, the lack of eyelid pigmentation increases the risk of damage from the sun and at least two diseases: ocular squamous cell carcinoma (OSCC), and bovine infectious keratoconjunctivitis (BIK) (1–3). The most common form of malignant tumour affecting cattle is OSCC which occurs on the eye and eyelids. This disease occurs at a high frequency in Hereford cattle (1). The occurrence of OSCC causes reduced longevity and economic loss as cancer-stricken cattle are usually rejected, or

condemned, at the meat processing plant as unfit for human consumption. Russell et al. (5) followed 396 Hereford cows over five years and observed that 51% of these cows were identified to have a lesion on their eye at least once in their life, and White & Moore (6) reported that OSCC accounted for approximately 11.6% of the beef carcass condemnations between 2003 to 2007.

Some white-faced Hereford, Simmental and Fleckvieh cattle have regions of pigmentation around the eyes known as ambilateral (both sides) circumocular (around the eyes) pigmentation. In Fleckvieh cattle, the frequency of ambilateral circumocular pigmentation (ACOP) is estimated to be around 36% (7), and in Hereford cattle it is suggested that approximately two thirds of the population have at least some pigmentation around the eyes or on the eyelids (8). ACOP is associated with reduced incidence of OSCC and BIK in white-faced cattle (9,10). ACOP has a reported heritability of between 0.41 and 0.50 (8), indicating that selection for this trait in white-faced cattle would increase the occurrence of ACOP and would likely reduce the incidence of disease. Although the heritability of ACOP and its correlation with OSCC and BIK have been extensively studied, to our knowledge only one study has investigated the genetics of this trait. Pausch et al. (11) estimated breeding values for ACOP in a Fleckvieh cattle population of 320,186 cows and 3,579 genotyped bulls. That study reported twelve quantitative trait loci (QTL) associated with ACOP across chromosomes 2, 5, 6, 11, 13, 14 and 22, and implicated well-established pigmentation genes *paired box 3* (*PAX3*), *proto-oncogene receptor tyrosine kinase* (*KIT*), and *micropthalmia-associated transcription factor* (*MITF*), among others. They also reported higher heritability values for ACOP (0.79±0.04) than previously reported by Anderson et al. (8).

The aim of our study was to investigate the genetic architecture of ACOP in Hereford cattle. We report a single significant signal on chromosome 6, and suggest *KIT* as the likely causal gene at this locus, with a single nucleotide polymorphism (SNP)-based heritability of 0.75. The results reported in this paper may inform future breeding decisions to select for ACOP in Hereford cattle, thereby enhancing welfare and profitability, especially in pasture-based farming systems where cattle are directly exposed to potentially damaging UV radiation.

5.3 Materials and Methods

5.3.1 Study population

Face colour (white, splotchy or speckled) and the presence of pigmentation around the eye (one eye, two eyes, or none) were visually scored from photographs of 706 Hereford cattle taken in August of 2019. The cattle were sired by 102 bulls in a progeny test programme, with 38 bulls having more than five offspring within the dataset. The cattle were raised on the Olsen Hereford Ranch in Harrisburg, Nebraska, USA. DNA samples from ear tissue biopsies were extracted and genotyped by GeneSeek (Lincoln, NE, USA) using an Illumina Bovine 50k SNP Chip. All SNP positions used and mentioned in this study are based on the ARS-UCD1.2 bovine genome build (12).

5.3.2 Population structure adjustments and GWAS

A genomic relationship matrix (GRM) was generated using 50k SNP Chip genotype data for each bovine autosome in GCTA (v1.93.2beta) to address possible population

stratification in the association model due to relatedness. We treated the ocular pigmentation trait as a binary variable and used a 'leave one chromosome out' approach to conduct a mixed linear model-based association analysis in GCTA, which incorporated the GRM as previously described by Jivanji et al. (13). We investigated the potential functional impact of all variants that surpassed the Bonferroni-corrected genome-wide significance threshold (p=5×10^{-8}) using the Ensembl Variant Effect Predictor (VEP) (14). In subsequent analyses, we fitted individual face-colour classes (white face, splotchy face, and speckled face), and the tag SNP genotype as separate covariates in the model.

5.3.3 Mode of inheritance and trait heritability

Dominant and recessive modes of inheritance for the ACOP trait were tested on the most significant signal identified in the GWAS (chromosome 6) by using the '--model' function in PLINK (v1.9) (15). In the dominant and recessive models, the minor allele (in this example the G allele) was used as the reference allele to define recessive and dominant encoded genotypes. The dominant model tested the effect of GG and AG versus AA, the recessive model tested the effect of GG versus AG and AA, and the additive model tested the regression of the number of G alleles on the presence or absence of ACOP. The ACOP trait was treated as a binary variable, as described above, and the association model was conducted on genetic markers across chromosome 6 (N=2,102). A principal component analysis-based approach was used to account for stratification due to relatedness in this model. Eigenvectors were calculated using all 50k SNP Chip markers, and the top ten principal components were fitted as covariates, based on visualisation of principal component clusters and a scree plot of the eigenvalues. The top ten principal components explain more than 90% of the genotype

variance in our population. A genome-wide GRM was used to estimate the SNP-based trait heritability in GCTA-GREML (v1.93.2beta), where the trait prevalence was set to 0.34 based on a study by Mészáros et al. (7).

5.4 Results

5.4.1 Genome-wide association analysis

Phenotypes of the Hereford population comprised 101 cattle with pigmentation around just one eye (unilateral circumocular pigmentation; UCOP), 240 cattle with pigmentation around both eyes (i.e., ACOP), and 365 cattle with no pigmentation around their eyes. The genome-wide association study (GWAS) for the absence or presence of any eye pigmentation (UCOP or ACOP; N=706 cattle), and that for the presence or absence of ACOP (N=605 cattle), both revealed the same signal. The results for both GWAS revealed a single signal on chromosome 6 that exceeded our multiple hypothesis testing threshold (p=5×10^{-8}), and mapped to the same top variant at Chr6 g.71059814G>A (UCOP or ACOP p=5.2×10^{-19}; just ACOP p=1.7×10^{-22}; Fig. 5.1). Since the p-value was more significant for the second analysis, where cattle with ACOP were compared to cattle with no circumocular pigmentation, this phenotype became the focus of the analyses that are presented below.

5.4.2 Influence of face colour on ACOP

Within our population, we observed Hereford cattle to have one of three different face-colour classes: white face, splotchy face, or speckled face (Table 5.1), where UCOP or ACOP were not considered in the allocation of animals to these three classes. The

occurrence of ACOP did not appear to segregate with any one face colour class, appearing to manifest regardless of face colour. In our association model, we accounted for face colour class and found that the top variant (Chr6 g.71059814G>A) remained at the top of the peak with very marginal shift in significance (white face $p=3.2\times10^{-22}$; splotchy face $p=2.7\times10^{-21}$; speckled face $p=1.2\times10^{-22}$).

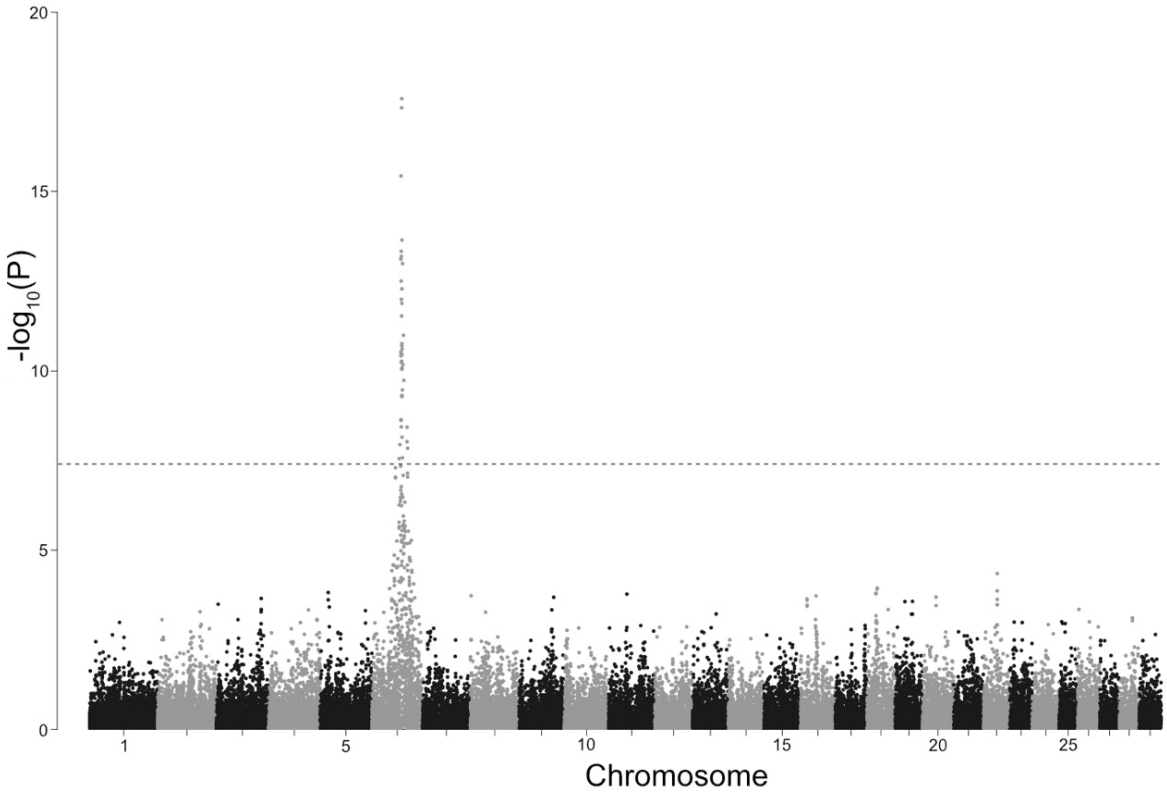
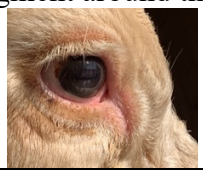


Figure 5.1 Manhattan plot based on genome-wide association analysis for the absence or presence of ambilateral circumocular pigmentation in Hereford cattle. There is a single signal that exceeds the significance threshold indicated by the dashed line ($p=5\times10^{-8}$) on chromosome 6, and the top variant maps to Chr6:71,059,814bp with a p-value of 1.7×10^{-22}.

Table 5.1 Image of each trait scored and number of cattle scored as having no pigment around their eyes versus pigment around both eyes (ambilateral circumocular pigmentation) for each face colour trait: white face, splotchy face, and speckled face. Cattle with both splotchy faces and speckles (N=7) were counted twice.

	White Face	Splotchy Face	Speckled Face
			
No pigment around the eyes 	332	2	28
Pigment around both eyes 	151	71	28

5.4.3 Analysis of chromosome 6 locus

The top variant at the chromosome 6 locus was a SNP at Chr6 g.71059814G>A (rs134749374; p=1.7×10^{-22}), which maps to intron 1 of the *Neuromedin-U* (*NMU*) gene, and is 805 kb downstream of the *proto-oncogene receptor tyrosine kinase* (*KIT*) gene. We detected 43 additional variants that exceeded the genome-wide wide significance threshold (p=5×10^{-8}), spanning approximately 20 Mb in the vicinity of the top variant. Our investigation of all genes that mapped to this region identified the *KIT* gene as the most likely causal gene. Assessment of the predicted functional impact of the top (N=44) variants using Variant Effect Predictor (VEP) did not present any variants with

obvious impacts on gene expression and/or protein function: one variant was predicted to be a 3' untranslated region variant, two variants were upstream gene variants, 11 variants were intron variants, and the remaining 30 variants were intergenic. Likewise, the top associated variant was not found to map to a highly conserved region that could support functional impact, as might be expected given the variants represented 50k SNP chip data only. Table 5.2 lists the top ten associated variants from chromosome 6 and their predicted functional effects. After fitting the top variant as a covariate in our association model, most of the significance was removed from the locus (smallest *p*-value was 0.0005 for Chr6 g.27370287A>G), suggesting a single biallelic QTL as responsible.

5.4.4 Mode of inheritance

The model used to test the contribution of additive, dominant, and/or recessive inheritance on the manifestation of ACOP identified the same top SNP from the previous additive model conducted in GCTA: Chr6 g.71059814G>A. The smallest *p*-value for this variant came from the additive model ($p=1.6\times10^{-37}$), followed by the dominant model ($p=2.0\times10^{-36}$), and then the recessive model ($p=4.8\times10^{-10}$). Although these results support an additive mode of inheritance, the small number of cattle with ACOP that were homozygous for the minor allele of this variant (N=29, minor allele frequency=0.27) may have limited our ability to resolve the mode of inheritance. When we examined the difference between phenotype (ACOP or no ACOP) by genotype class, we observed a larger difference in phenotype for the AA genotype, compared to the AG and GG genotypes. These results may allude to a non-additive mode of inheritance (Fig. 5.2). The heritability of ACOP estimated by GCTA-GREML was 0.75±0.07.

Table 5.2 Top 10 variants for the chromosome 6 locus detected in the genome-wide association analysis for the presence or absence of ambilateral circumocular pigmentation in Hereford cattle, with minor allele frequency, predicted variant effects from Ensembl Variant Effect Predictor, and gene the variant maps to.

Variant reference ID	Genomic position on Chr6	*p* value	Minor allele frequency	Predicted variant effect	Gene symbol
rs134749374	g.71059814G>A	1.7×10^{-22}	0.27	Intron variant	*NMU*
rs135243955	g.69234941C>T	2.6×10^{-18}	0.19	Intergenic	-
rs110185900	g.69224236C>T	4.6×10^{-18}	0.19	Intergenic	-
rs109877370	g.67582129C>T	3.7×10^{-16}	0.22	Intergenic	-
rs110547897	g.69596878C>T	2.2×10^{-14}	0.27	Upstream gene variant	*CHIC2*
rs132937166	g.68028962T>C	4.7×10^{-14}	0.2	Intergenic	-
rs133388993	g.68428584G>A	6.5×10^{-14}	0.32	Intergenic	-
rs109242003	g.67159718C>T	7.7×10^{-14}	0.13	Intergenic	-
rs29017695	g.71028209G>T	1.0×10^{-13}	0.17	Upstream gene variant	*PDCL2*
rs41653781	g.67991798A>C	3.1×10^{-13}	0.27	Intergenic	-

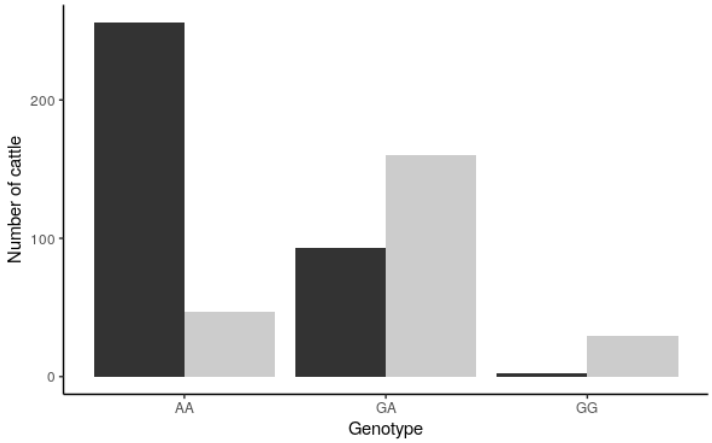


Figure 5.2 The number of cattle observed to have ambilateral circumocular pigmentation (grey) or no pigmentation around the eyes (black) by genotype class for tag variant Chr6 g.71059814G>A.

5.5 Discussion

To the best of our knowledge, we present the first genome-wide association analysis for ACOP in Hereford cattle. Using phenotype scores for 605 American Hereford cattle, and 43,789 genome-wide genetic markers, we identified a single significant association signal on chromosome 6, mapping to Chr6 g.71059814G>A (p=1.7×10^{-22}). Unexpectedly, the strength of association for this signal was not affected by face colour, suggesting that ACOP is controlled by different alleles to those that influence the colour of the face in Hereford cattle. We suggest that *KIT* is the most likely causal gene for ACOP, and given that the white-face trait has also been attributed to dominant mutations of the same gene (16), these findings suggest multiple segregating pigmentation QTL in these animals.

There were 44 variants that exceeded the genome-wide significant threshold (p=5×10^{-8}) on chromosome 6, with the most significant variant mapping 805 kb downstream of the *KIT* gene. Although this variant is nearly 1 Mb away from *KIT*, other variants with *p*-values of a similar magnitude mapped within 240 kb of the *KIT* gene. Given the use of a relatively sparse genotyping platform (43k SNP) for mapping, and the status of *KIT* as one of the most-frequently implicated and best-characterised pigmentation genes in mammals, these findings suggest *KIT* as the likely causal gene for the ACOP trait. The KIT protein is a tyrosine kinase mast/stem cell growth factor receptor that is involved in melanocyte stem-cell migration out of the neural crest during embryonic development. KIT facilitates melanocyte migration along the neural crest to the final destination in the skin. Mutations that alter *KIT* expression or activity are hypothesised to impact stem-cell migration, proliferation and/or survival. This impairs the ability of melanocyte

precursors to colonise their final destination, causing regions of the skin and hair to lack pigmentation and appear white (2,17,18). Bioinformatic prediction of the significant GWAS variants for functional impacts did not detect any likely causal mutations. However, when the most highly associated SNP Chr6 g.71059814G>A, was fitted as a fixed effect in the association model, this SNP accounted for the majority of the signal at this locus. Although it is possible that Chr6 g.71059814G>A may have some effect on gene expression or regulation, this is unlikely as the variant does not appear to be highly conserved amongst vertebrates, and represents but one of many candidates that would likely be represented if GWAS was performed with higher-density genotype data (e.g. whole genome sequence). More likely, Chr6 g.71059814G>A tags the causal effect for the manifestation of ACOP and is in linkage disequilibrium with the causal mutation. The ACOP trait has also been studied in Fleckvieh cattle, where Pausch et al. (11), reported several significant signals across the genome. The most significant signal was reported on chromosome 6, mapping to Chr6 g.70362480T>C (p=1.1×10^{-11}). That variant was not captured in our genotype data, but maps within 700 kb of our tag SNP and is likely tagging the same causal mutation. The heritability reported in our population (0.75±0.07), falls close to that reported by Pausch et al. (2012) (0.79±0.04), perhaps supporting a common genetic architecture. Future studies could use whole-genome sequencing of cattle with ACOP from both Fleckvieh and Hereford to fine map the causal mutation for ACOP, and establish if ACOP is caused by the same genetic effect in both breeds.

The dominantly inherited white-face trait in Hereford cattle has also been mapped to chromosome 6, locating upstream of the *KIT* gene. The white-face trait manifests if cattle carry one or more copies of a large tandem duplication approximately 50 kb

upstream of the *KIT* gene. The tandem duplication is hypothesised to impair *KIT* expression, preventing functional melanocytes from occupying the skin and hair of the face, resulting in the lack of pigmentation in this region (16). As ACOP is superimposed on the white-face trait, we investigated if face-colour influenced the presence/absence of ACOP. Fitting each face colour class (white, splotchy and speckled) as covariates in the model did not affect the GWAS signal. These results suggest that face colour cannot account for the signal observed at the chromosome 6 locus, and the manifestation of ACOP is caused by a different allele, perhaps at the same locus. Due to the molecular mechanism proposed to cause the white-face trait, and its dominant mode of inheritance, we investigated the possibility of a non-additive mode of inheritance for ACOP. Comparison of cattle with and without the ACOP trait by genotype class possibly supports a non-additive mode of inheritance, although the small number of cattle with ACOP that were homozygous for the minor allele limited our ability to formally resolve mode of inheritance. An epistatic mode of inheritance would not be entirely surprising as the GWAS signal maps near the *KIT* gene, the causal gene for the white face trait. It is possible that while the white face mutation impairs *KIT* expression, the ACOP mutation may have a positive effect on *KIT* expression and/or function, allowing functional melanocytes to migrate to, and occupy the eye area, pigmenting the surrounding skin and hair. These results are speculative and would require a larger population size to further investigate the mode of inheritance, and possible epistatic interaction between the white-face mutation and ACOP mutation.

Our results identify a strong association between Chr6 g.71059814G>A and the ACOP trait in American Hereford cattle, and implicate the likely involvement of the *KIT* gene. Our results also suggest genetic complexity with regard to the mode of inheritance, and

we suggest that ACOP in Hereford cattle and Fleckvieh cattle may be caused by the same mutation, but more work would be required to support this. Selection for ACOP in white-faced cattle could enhance welfare and profitability, by reducing the incidence of OSCC and BIK, especially in cattle raised in pasture-based grazing systems. Although this work has been done in American Hereford cattle, ACOP has also been observed in New Zealand Hereford cattle, and future studies could aim to investigate if the same genetic variants influencing ACOP segregate in local populations. To our knowledge, New Zealand Hereford breeders do not actively select for ACOP, but may benefit from considering ACOP associated genetic variants in their selection index. Future work may aim to investigate the causal mutation for ACOP in New Zealand Hereford populations and develop a gene test to select appropriate sires and increase the frequency of ACOP in future generations.

5.6 Acknowledgements

The authors would like to acknowledge the American Hereford Association for providing the genotype data and photographs used in this research.

5.7 References

1. Heeney JL, Valli VEO. Bovine ocular squamous cell carcinoma: An epidemiological perspective. Can J Comp Med. 1985;49(1):21–6.

2. D'Mello S, Finlay G, Baguley B, Askarian-Amiri M. Signaling pathways in melanogenesis. Int J Mol Sci. 2016;17(7):1144.

3. Seid A. Review on infectious bovine keratoconjunctivitis and its economic

impacts in cattle. J Dairy Vet Sci. 2019;9(5):555774.

4. Russell WC, Brinks JS, Kainer RA. Incidence and heritability of ocular squamous cell tumors in Hereford cattle. J Anim Sci. 1976;43(6):1156–62.

5. Berry S, Coppieters W, Davis S, Burrett A, Thomas N, Palmer D, et al. A Triad of Highly Divergent Polymeric Immunoglobulin Receptor (PIGR) Haplotypes with Major Effect on IgA Concentration in Bovine Milk. PLoS One. 2013;8(3):1–10.

6. White TL, Moore DA. Food for thought for food animal veterinarians. J Am Vet Med Assoc. 2009;235(8):937–41.

7. Mészáros G, Petautschnig E, Schwarzenbacher H, Sölkner J. Genomic regions influencing coat color saturation and facial markings in Fleckvieh cattle. Anim Genet. 2015;46(1):65–8.

8. Anderson DE, Chambers D, Lush JAYL. Studies on bovine ocular squamous carcinoma (“cancer eye”) III. Inheritance of eyelid pigmentation. J Anim Sci. 1957;16(4):1007–16.

9. Ward JK, Nielson MK. Pinkeye (bovine infectious keratoconjunctivitis) in beef cattle. J Anim Sci. 1976;49(2):361–6.

10. Davis KM. Digital analysis of eye pigmentation of Hereford, Hereford x Bos indicus or Hereford x Bos taurus cattle. Texas A&M University; 2013.

11. Pausch H, Wang X, Jung S, Krogmeier D, Edel C, Emmerling R, et al. Identification of QTL for UV-protective eye area pigmentation in cattle by progeny phenotyping and genome-wide association analysis. PLoS One. 2012;7(5):e36346.

12. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9(3):giaa021.

13. Jivanji S, Worth G, Lopdell TJ, Yeates A, Couldrey C, Reynolds E, et al. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. Genet Sel Evol. 2019;51(1):1–18.

14. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):1–14.

15. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

16. Whitacre L. Structural variation at the KIT locus is responsible for the piebald phenotype in Hereford and Simmental cattle. University of Missouri; 2014.

17. Adameyko I, Lallemend F, Aquino JB, Pereira JA, Topilko P, Müller T, et al. Schwann cell precursors from nerve innervation are a cellular origin of melanocytes in skin. Cell. 2009;139(2):366–79.

18. Mort RL, Ross RJH, Hainey KJ, Harrison OJ, Keighren MA, Landini G, et al. Reconciling diverse mammalian pigmentation patterns with a fundamental mathematical model. Nat Commun. 2016;7(1):1–13.

5.8 Appendix

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

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:	Swati Jivanji
Name/title of Primary Supervisor:	Dorian Garrick
In which chapter is the manuscript /published work:	Chapter Five

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Jivanji S, Kosch T, Littlejohn M, Garrick D. Genome-wide association study of UV-protectant ambilateral circumocular pigmentation in Hereford cattle. New Zealand Journal of Animal Science and Production. 2021 [In Press].

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Swati Jivanji
Date:	25/11/2021
Primary Supervisor's Signature:	Dorian Garrick
Date:	25 Nov 2021

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.

CHAPTER SIX

The genomes of precision edited cloned calves show no evidence for off-target events or increased de novo mutagenesis

PUBLISHED IN BMC GENOMICS (2021)

The genomes of precision edited cloned calves show no evidence for off-target events or increased *de novo* mutagenesis

Swati Jivanji[1*], Chad Harland[2], Sally Cole[3], Brigid Brophy[3], Dorian Garrick[1], Russell Snell[4], Mathew Littlejohn[1,2] and Götz Laible[3,5,6]

[1] School of Agriculture and Environment, Massey University, Palmerston North, New Zealand

[2]Livestock Improvement Corporation, Newstead, New Zealand

[3]AgResearch, Ruakura Research Centre, Hamilton, New Zealand

[4]School of Biological Sciences, University of Auckland, Auckland, New Zealand

[5]School of Medical Sciences, University of Auckland, Auckland, New Zealand

[6]Maurice Wilkins Centre for Molecular Biodiscovery, Auckland, New Zealand

*Corresponding author

Email: swati.jivanji.1@uni.massey.ac.nz

6.1 Abstract

Background: Animal health and welfare are at the forefront of public concern and the agricultural sector is responding by prioritising the selection of welfare-relevant traits in their breeding schemes. In some cases, welfare-enhancing traits such as horn-status (i.e., polled) or diluted coat colour, which could enhance heat tolerance, may not segregate in breeds of primary interest, highlighting gene-editing tools such as the CRISPR-Cas9 technology as an approach to rapidly introduce variation into these populations. A major limitation preventing the acceptance of CRISPR-Cas9 mediated gene-editing, however, is the potential for off-target mutagenesis, which has raised concerns about the safety and ultimate applicability of this technology. Here, we present a clone-based study design that has allowed a detailed investigation of off-target and *de novo* mutagenesis in a cattle line bearing edits in the *PMEL* gene for diluted coat-colour.

Results: No off-target events were detected from high depth whole genome sequencing performed in precursor cell-lines and resultant calves cloned from those edited and non-edited cell lines. Long molecule sequencing at the edited site and plasmid-specific PCRs did not reveal structural variations and/or plasmid integration events in edited samples. Furthermore, an in-depth analysis of *de novo* mutations across the edited and non-edited cloned calves revealed that the mutation frequency and spectra were unaffected by editing status. Cells in culture, however, appeared to have a distinct mutation signature where *de novo* mutations were predominantly C>A mutations, and in cloned calves they were predominantly T>G mutations, deviating from the expected excess of C>T mutations.

Conclusion: We found no detectable CRISPR-Cas9 associated off-target mutations in the gene-edited cells or calves derived from the gene-edited cell line. Comparison of *de novo*

mutation in two gene-edited calves and three non-edited control calves did not reveal a higher mutation load in any one group, gene-edited or control, beyond those anticipated from spontaneous mutagenesis. Cell culture and somatic cell nuclear transfer cloning processes contributed the major source of contrast in mutational profile between samples.

6.2 Introduction

The agriculture sector's response to demands for enhanced animal welfare, production, efficiency and sustainability is sometimes limited by available genetic variation within a particular population. Although favourable variation may be introgressed from other populations by cross-breeding, stabilising favourable variation by selective breeding regimes typically comes at the cost of losses in genetic gain and inbreeding depression. Gene-editing offers an attractive solution with its ability to directly introduce precise polymorphisms causal for favourable traits within a single generation. Acceptance of gene editing technologies is in part dependent on the occurrence of mutagenesis at sites other than the intended on-target site, or 'off-target' mutagenesis, and the ability to detect these events above baseline mutation levels.

The clustered regularly interspaced short palindromic repeat (CRISPR)-CRISPR associated (Cas) system is a versatile and popular gene-editing tool proven to be successful in large animal models (2). The most commonly used CRISPR-Cas9 system is derived from *Streptococcus pyogenes*, and uses the Cas9 endonuclease complexed with a guide RNA (gRNA) that identifies and binds to a 20-nucleotide target region (protospacer) immediately preceding a NGG protospacer-associated motif, or PAM. The endonuclease induces a

double stranded break 3bp upstream of the NGG site, which can either be repaired via non-homologous end joining, or a repair template coding for the desired polymorphism can be introduced to facilitate homology-directed repair (3,4). The potential for off-target mutations have been associated with non-unique matches and sequence mismatches distal from the PAM sequences at the 5' end of the gRNA (5–7). Structural variation at the targeted edit site (8–10), and unintended integration of the editing vectors (11,12) have also been associated with gene-editing and have raised concerns about the safety and applicability of these technologies in biomedicine and agriculture.

Off-target mutations have been investigated by amplification and sequencing of pre-selected sites identified by bioinformatic tools that highlight sequences with homology to the on-target site (13–15). This method may not be practical for large-scale screening, with the generation of a large number of possible non-unique matches. This approach also neglects to consider the potential for mutations to be introduced at sites with low on-target sequence similarity, and thus will not be able to identify such events. Whole genome sequencing (WGS) is a less biased approach to off-target mutation detection and enables analysis of single nucleotide variants (SNV), small insertions and deletions (indels), and some structural variants (SV), that may arise as a result of the use of CRISPR-Cas9 mediated gene-editing. However, since cells naturally accumulate *de novo* mutations through spontaneous mutagenesis during cell division, it is challenging to distinguish mutations attributable to the application of gene-editing technologies from those that occur spontaneously. To characterise any off-target mutagenesis, one approach is to quantify changes in detectable *de novo* mutation between gene-edited samples and controls, and then assess whether candidate variants do, or do not, sit in biologically plausible off-target sites.

This approach has been used to evaluate the presence and frequency of off-target mutations in gene-edited large animal models, generated from multiplexed single-cell-embryo injection, and their progeny (8,16). Wang et al. (8) and Li et al. (16) used a trio-based study design to investigate off-target effects of CRISPR-Cas9 and showed that the off-target mutation rate was negligible and the *de novo* mutation rate in edited animals was comparable to their non-edited controls. A WGS approach to off-target mutation detection was also used by Schaefer et al. (17) to identify off-target mutations in two gene-edited mice generated by single-cell embryo injection (18). Schaefer et al. (17) reported hundreds of off-target mutations by WGS comparison to a single untreated mouse, but this result was found to be flawed when the authors later reported no excess mutations upon conducting WGS analysis with additional mouse lines (19). These studies highlight the importance of considering inherited and spontaneous mutations when investigating off-target events, and the use of appropriate controls that enable these considerations to be made.

In this study, we conduct the first WGS analysis in cloned cattle generated from a gene-edited cell line to evaluate off-target events and *de novo* mutagenesis associated with the application of CRISPR-Cas9 mediated gene-editing and cloning to create live cattle for use in agriculture. We analysed WGS from a cell clone homozygous for a CRISPR-Cas9 induced 3bp deletion in the premelanosomal protein gene (*PMEL*), the parental (non-edited) primary fetal cell line that cell clone was derived from, as well as two edited and three control calves generated from these cells by somatic cell nuclear transfer. The 3bp deletion in the *PMEL* gene was proposed to cause coat colour dilution in Highland and Galloway cattle (20), and by introducing it into a Holstein-Friesian background, Laible et al. (1) simultaneously demonstrated causality of this mutation and introduced a favourable

trait within a single generation (1). Taking advantage of the clone-based study design, we used WGS and other molecular approaches to comprehensively screen for off-target SNVs, indels, and SVs that could be attributed to the use of CRISPR-Cas9 mediated gene-editing. We found no detectable CRISPR-Cas9 associated off-target mutations, and that the *de novo* mutation rate in calves generated from the gene-edited cell line was no different in calves generated from the non-edited cell line of same parental origin.

6.3 Results

6.3.1 Origin of the study material and analysis of whole genome sequence data

We used the recently described cloned calves that were edited for a *PMEL* coat colour dilution mutation (1) to investigate the precision of CRISPR-Cas9 gene-editing. For our in-depth genotype analysis, we applied WGS and included the parental non-edited cell line (BEF2), the gene-edited clonal cell line (CC14) derived from BEF2, three control clones (1802, 1803 and 1804) generated from BEF2 cells, and two gene-edited clones (1805 and B071) that were generated with CC14 donor cells (Fig 6.1). The average whole genome sequencing depth per sample was 50.7×, after alignment to the bovine reference genome ARS-UCD1.2 (21). Greater than 99% of the reads mapped to the reference genome, and more than 92% of the reads mapped with a map quality score of 60 across all samples except sample B071, which had approximately 80% of reads with a map quality score of 60. Variant calling using GATK HaplotypeCaller (22) identified 8,021,969 variants across the seven samples. Principal component analysis showed that samples did not appear to cluster together based on treatment group (edited versus non-edited; Fig 6.S1), and a pair-

wise genomic concordance test across the seven samples found 99.99% concordance between all pairs, consistent with clones originating from the same genetic background.

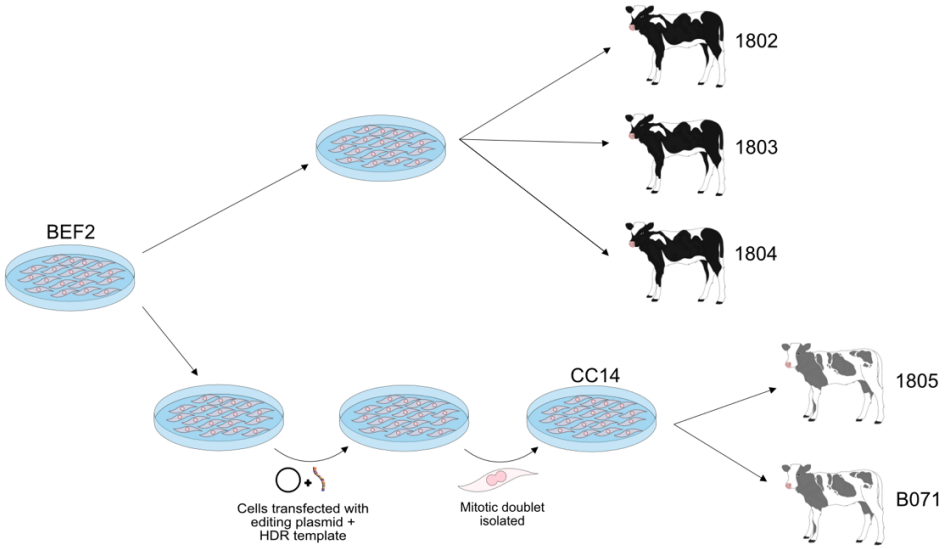


Figure 6.1 Relationship between the parental cell line BEF2, edited cell clone CC14 and edited and control calves. Shown is an experimental flow diagram from the parental cell line BEF2 to the two coat colour-diluted Holstein-Friesian calves homozygous for PMEL *p.Leu18del and three wild-type control calves. A subset of the male primary bovine fetal fibroblasts (BEF2) were transfected with a plasmid-encoded,* PMEL*-specific editor and a single stranded homology directed repair (HDR) template. Post transfection, a single mitotic doublet was used for the clonal isolation of CC14 with a homozygous* PMEL *p.Leu18del mutation. Two edited cloned calves (1805 and B071) and three non-edited control calves (1802, 1803 and 1804) were generated via somatic cell nuclear transfer using CC14 and BEF2 as donor cells, respectively. The 'named' samples are those that were sequenced in this study (i.e., BEF2, CC14, 1802, 1803, 1804, 1805, and B071).*

6.3.2 Identification of off-target mutations from WGS data

To identify mutations that may be the result of CRISPR-Cas9 mediated gene-editing, we applied a series of stringent filtering procedures (Fig 6.2). Variants relative to the reference genome that were identified to be monomorphic across all samples (N=7,670,567), and

those few sites with no short-read sequence coverage in the BEF2 parental cell line (N=14,947), were removed which reduced the 8,021,969 variant sites to 336,455 variants. To remove polymorphic variants that were present in BEF2 but are common to the wider cattle population, all variants that were segregating in a large sequenced New Zealand (NZ) dairy cattle population (see Methods) were also removed, further reducing the number of variants to 31,190. Variants that were present in the gene-edited cell line (CC14) and both gene-edited clones (1805 and B071), but absent in the parental cell line (BEF2) and all three control clones (1802, 1803 and 1804), were retained and variants with a map quality score of less than 60 were removed. This reduced the total number of candidates for variants induced by potential off-target events or spontaneous *de novo* mutagenesis to 457. Variants called to be heterozygous by GATK HaplotypeCaller (22) but identified to have an allele dosage significantly less than 0.5 in the CC14 cell line, 1805, or B071, were defined as candidate mosaic mutations and were filtered out, as it was likely that these mutations occurred after the gene-edited mitotic doublet was isolated (Fig 6.1). The remaining 218 candidate off-target/*de novo* mutations were then manually examined by visualisation of sequence reads in the Integrative Genomics Viewer (IGV). Using this filtering criteria (Fig 6.2), we identified 151 candidate mutations that may have resulted from off-target mutagenesis (131 SNVs and 20 indels; Table 6.S1). We also investigated SVs that may have been induced by the use of CRISPR-Cas9. Using a case-control design, DELLY (v0.8.1) (23) was used to predict the presence of SVs in CC14, 1805 and B071 that were absent in BEF2, 1802, 1803 and 1804. Using this approach, there were no detectable SVs that were present in the CC14 cell line and the two gene-edited cloned calves, yet absent in all control samples (Table 6.S2).

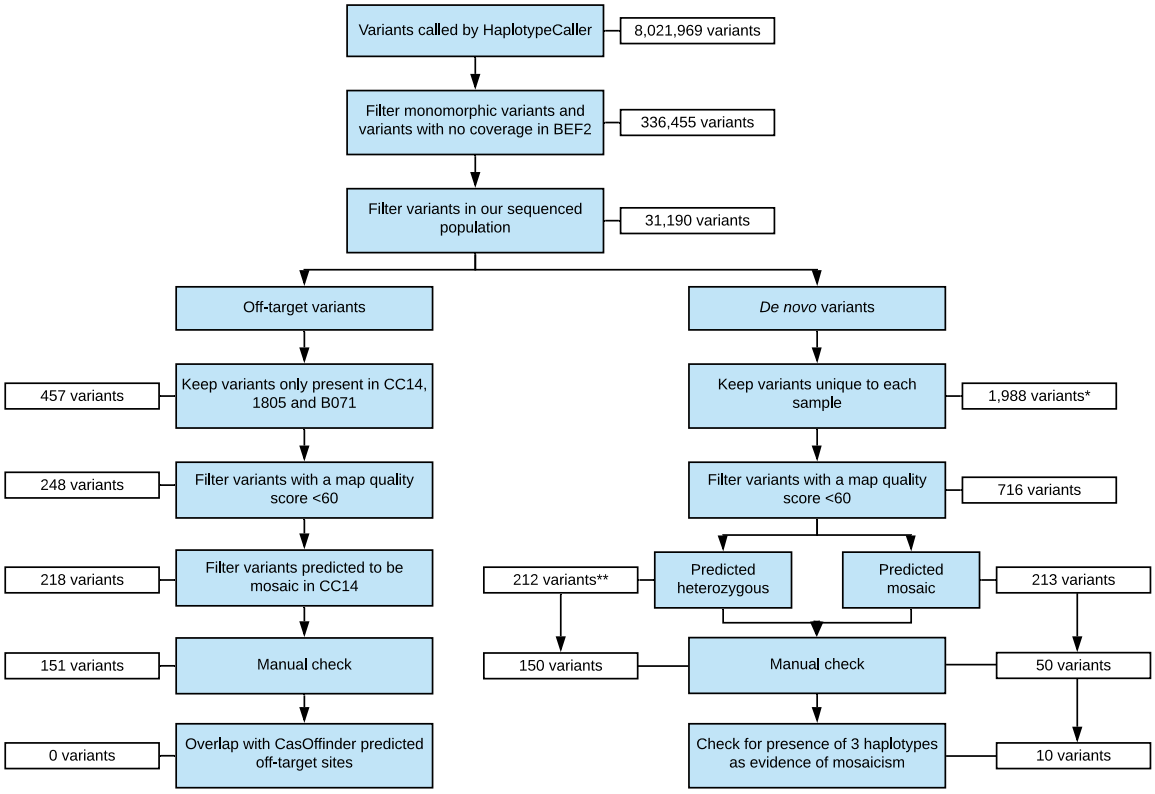


*Figure 6.2 Filtering criteria applied to raw variant calls to identify potential off-target mutations and spontaneous de novo mutations in the gene-edited cell line CC14. The white boxes adjacent the filtering criteria indicate the number of candidate mutations remaining after the filter was applied. *Variants were kept if also present in 1805 and/or B071 **Predicted heterozygous de novo mutations were also filtered for their presence in calves 1805 and B071.*

6.3.3 Identification of off-target mutations at predicted candidate loci

Potential genome-wide off-target sites were predicted based on on-target sequence similarity using Cas-OFFinder (13), where we allowed for up to five mismatches with the on-target site. Cas-OFFinder identified 1,166 potential off-target sites, none of which mapped within ±50bp of any of the 151 candidate mutations identified by our discovery pipeline. The sequence flanking each of the 151 candidate mutations was also manually inspected for evidence of sequence similarity with the gRNA and an adjacent PAM site,

with no matches or partial matches identified. To ensure that our filtering criteria had not excluded variants from the most likely off-target mutation sites, we also searched the unfiltered variant calls for matches with the sites identified by Cas-OFFinder. We found 230 (of 8,021,969) variants that mapped within 50bp of the 1,166 candidate off-target sites. Almost all (N=225) were filtered out due to being monomorphic across all samples, three sites were filtered out due to poor read quality, one captured the on-target mutation at the edited site, and one site was called in the sample of the non-edited control calf 1803. These steps provided reassurance as to the filtering criteria applied, and suggested that if CRISPR-Cas9 induced off-target mutagenesis had occurred, it had not done so at any of the most biologically plausible sites.

6.3.4 Long molecule sequencing of the on-target site

To investigate the on-target edit site for SVs and plasmid integration events, we conducted long-range polymerase chain reaction (PCR) to amplify approximately 8.8kb of sequence surrounding the edit site (Chr5:57,340,856bp-57,349,715bp) in the parental cell line (BEF2), the gene-edited cell clone (CC14), two gene-edited cloned calves (1805 and B071), and one control clone calf (1802). The amplicons were sequenced using the Oxford minION platform, generating an average sequence depth of 590× across the targeted region in each of the five samples, and minimap2 (24) was used to map the long sequence reads to the bovine ARS-UCD1.2 reference genome (21). Since structural variation might disrupt primer binding and lead to allele drop-out at the locus (i.e., a large hemizygous structural variant that could confound PCR), we looked for collocating variants to confirm biallelic amplification of the region. Manual inspection of the sequence reads in IGV revealed two

such biallelic SNVs (Chr5 g.57,343,664G>A and Chr5 g.57,348,336G>A) heterozygous in these samples, confirming that we captured both the maternal and paternal haplotypes across this region. Alignment of the long reads to the *PMEL*-specific CRISPR-Cas9 expression plasmid sequence using minimap2 (24) revealed no matches, suggesting that the editing plasmid was unlikely to have integrated at the on-target site.

6.3.5 Investigating evidence of plasmid integration

Although a PCR assay had previously failed to amplify a specific plasmid fragment (1), that approach assumes contiguous sequence representation of the plasmid template, and thus WGS data allows a more comprehensive analysis of potential integrations of the *PMEL*-specific CRISPR-Cas9 expression plasmid (and any potential fragments thereof). To investigate possible plasmid integration events at sites other than the on-target site, we added the sequence of the *PMEL*-specific CRISPR-Cas9 expression plasmid, that had been used for editing (a pX330 derivative), to the ARS-UCD1.2 reference genome (21) and re-ran sequence alignments using the Burrows-Wheeler Aligner (BWA; (25)) for the parental cell line (BEF2), gene-edited cell line (CC14), two gene-edited cloned calves (1805 and B071), and one control clone calf (1802). In all samples we observed a pile up of sequence reads in a G-rich repeat region at 828-873bp on the *PMEL*-specific editing plasmid. The mapping quality scores ranged between 0-35, suggesting these were mismapped reads, and of reduced interest given these were not polymorphic across the control and edited samples. No additional sequence reads were observed to map to the plasmid sequence for the two edited calves, control calf and parental cell line. Only for the CC14 sample, we found 46 additional sequence reads that appeared to map to the plasmid sequence (maximum

coverage of 8×). A *de novo* assembly of these reads indicated that these reads could not be assembled into a single contiguous sequence, and alignment to the bovine genome using BLAST (26) did not highlight any sequence overlap.

The limited read representation of *PMEL*-specific editing plasmid sequences mapped in CC14, and lack of these sequences in CC14-derived animals suggested bi- or mono-allelic integration in CC14 was unlikely, however we performed additional experiments to further investigate this possibility. Here, we designed two PCR primer pairs that together covered 1,365bp of the plasmid region, targeting sequence that overlapped the regions of homology identified from the short-read alignments. We designed a single primer pair targeted at Chr2:110,817,757-110,818,275bp, representing *Bos taurus* genomic sequence that would be expected to amplify in all samples. We created a mock plasmid-integrated DNA sample by spiking in 0.14 pg of the *PMEL*-specific editing plasmid into BEF2 gDNA, aiming to simulate a sample with a single integration event and thereby act as a positive control. These PCRs were conducted on DNA extracted from BEF2, B071, 1805, 1802, an aliquot of CC14 DNA previously extracted for WGS, and a fresh sample of DNA extracted from the CC14 cell clone. PCR amplification of the plasmid-specific 757bp and 690bp fragments returned a positive result in the plasmid and positive control sample, but a negative result in BEF2, both CC14 samples, B071, 1805 and 1802 (Fig 6.3). These results suggest that the short-read data seen to map to the plasmid sequence in CC14, was unlikely indicative of an integration event, and more likely due to low levels of sample contamination prior to WGS.

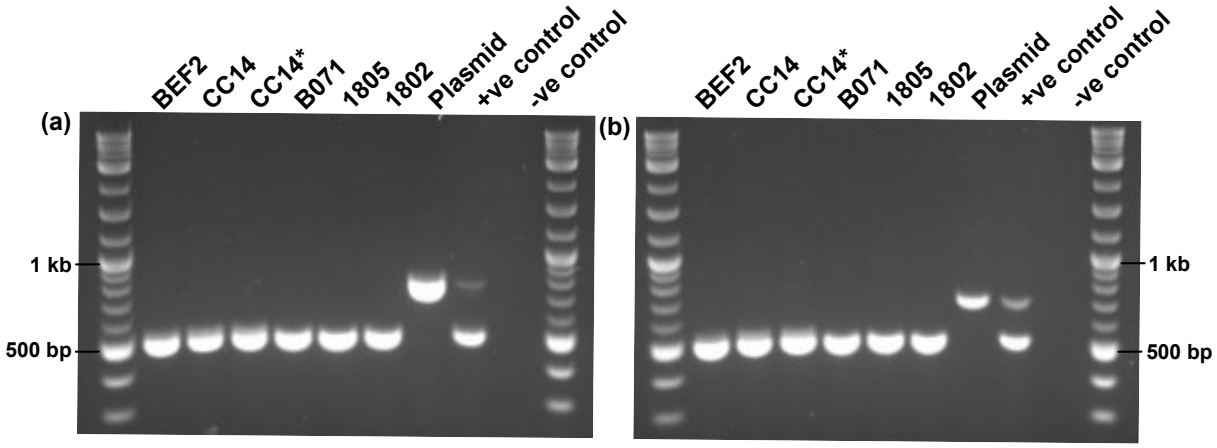


Figure 6.3 Absence of editing plasmid-specific fragments in genomic DNA extracted from the parental cell line (BEF2), the gene-edited cell clone CC14, DNA sent away for WGS of CC14 (CC14), and genomic DNA extracted from cloned calves B071, 1805, and 1802. Each PCR reaction contained two sets of primers and BEF2 spiked in with 0.14pg of plasmid DNA was used as the positive control. (a) Primer pair designed to amplify bovine Chr2:110,817,757-110,818,275bp (519bp), and another designed to amplify CRISPR-Cas9 expression plasmid-specific region 6,263-7,019bp (757bp); (b) Primer pair designed to amplify bovine Chr2:110,817,757-110,818,275bp (519bp product), and another to amplify plasmid-specific region 6,939-7,628bp (690bp). Fig 6.3 (a) and (b) have been cropped, and full-length gels are presented in Fig 6.S2.*

6.3.6 Analysis of *de novo* mutations in the cloned calves

The cloned calves used for this study were generated by somatic cell nuclear transfer with donor cells from either the parental cell line BEF2, or the gene-edited cell clone CC14 (1). To identify *de novo* mutations carried by each cloned calf, either originating from the donor cell or occurring during development of the calf, we applied the filtering criteria outlined in Figure 6.2. To differentiate between *de novo* mutations that likely occurred in cell culture and were subsequently inherited by the cloned calves, from *de novo* mutations that likely occurred during development of the cloned calves (i.e., after first cell division), we categorised *de novo* mutations as heterozygous or mosaic based on allele dosage at each

site (Table 6.1). A binomial probability function was applied to determine if the allele dosage at each variant site was consistent with a truly heterozygous genotype expected for a *de novo* mutation already present in the donor cell. When the allele dosage at a variant site was determined to be not statistically different from the expected allele dosage of 0.5, the variant was predicted to be a candidate heterozygous *de novo* mutation in the cloned calf, whereas if allele dosage was significantly less than 0.5, the variant was predicted to be a candidate mosaic *de novo* mutation that arose during development of the cloned calf. All variants were manually assessed in IGV software, after which a proportion of candidate *de novo* mutations were filtered out due to representing incorrect variant calls, most often due to errors based on proximity to polynucleotide regions, repetitive regions, miscalled variants in other samples, proximity to indels, or misalignment of reads. Table 6.1 shows the number of variants that remained after applying the filtering criteria outlined under '*de novo* variants' in Figure 6.2, where 'likely *de novo*' mutations are those that remained after the manual check.

Table 6.1 Number of candidate de novo mutations identified after each filter was applied to 31,190 filtered variants across the three control cloned calves and two gene-edited cloned calves

			Heterozygous *de novo* mutations		Mosaic *de novo* mutations	
Sample ID	Unique to each sample	Map quality = 60	Candidate *de novo*	Likely *de novo*	Candidate *de novo*	Likely *de novo*
1802	1,224	439	340	205	45	16
1803	1,402	550	409	276	82	14
1804	2,769	686	566	293	67	9
1805	1,145	433	313	197	52	8
B071	1,470	587	408	277	133	11

6.3.6.1 Heterozygous de novo mutations

The majority of *de novo* mutations present in the cloned calves appear to be heterozygous variants and are likely inherited from the donor cell used for somatic cell nuclear transfer. A pairwise comparison of the number of likely heterozygous *de novo* mutations inherited by each of the cloned calves (Table 6.1) suggests that the number of mutations observed in each clone is statistically different between six of the ten pairs (Table 6.2). The pair-wise comparison does not draw a distinction between the number of heterozygous mutations observed in the gene-edited compared to the non-edited calves, but rather appeared random. Based on these results, the number of heterozygous *de novo* mutations inherited by cloned calves generated from the gene-edited cell clone CC14 (1805 and B071) did not appear to be different than those in cloned calves generated from the non-edited, parental cell line BEF2 (1802, 1803 and 1804).

Table 6.2 Results (p-values) from two-proportion z-test comparing the difference in number of likely heterozygous (top) and mosaic (bottom) de novo mutations observed in the cloned calves.

	1802	1803	1804	1805	B071
1802		1.36×10^{-3}	9.07×10^{-5}	0.73	1.17×10^{-3}
1803	0.86		0.5	3.18×10^{-4}	1
1804	0.23	0.4		1.64×10^{-5}	0.53
1805	0.15	0.29	1		2.7×10^{-4}
B071	0.44	0.69	0.82	0.65	

6.3.6.2 Mosaic de novo mutations

The number of mosaic *de novo* mutations identified was more than a magnitude lower than the number of heterozygous *de novo* mutations identified (Table 6.1). These mutations, occurring during development of a calf, would be expected to be in complete, but imperfect linkage with the paternal or maternal haplotype (27), and we would therefore expect to see three haplotypes at the variant site. Each 'likely *de novo*' mosaic mutation (Table 6.1) was manually checked for evidence of a segregating bi-allelic variant on the same read, or read pair, to support the presence of three haplotypes and strengthen the evidence supporting a true mosaic mutation. Out of the total number of variants predicted to be likely true mosaic mutations: 8/16 variants in 1802, 6/14 variants in 1803, 5/9 variants in 1804, 2/8 variants in 1805, and 5/11 variants in B071 had evidence of three haplotypes and could be confirmed as true mosaic *de novo* mutations. A pair-wise significance test demonstrated that the difference in number of likely mosaic *de novo* mutations carried by each cloned calf (Table 6.1) does not appear to be statistically significant, regardless of the cell line of origin

(smallest *p*-value=0.15 between calves 1802 and 1805; Table 6.2). These results suggest that the *de novo* mutation rate during embryonic development does not significantly differ between cloned calves generated using donor cells from a cell clone edited by the CRISPR-Cas9 gene-editing tool, and those generated using a non-edited cell line of the same parental origin.

6.3.6.3 Comparison of mosaic de novo mutation rate in cloning compared to other reproductive technologies

We are unaware of any study to date that has attempted to quantify the *de novo* mutation rate in cloned animals. The most relevant comparison is a study by Harland et al. (28) which investigated the number of mosaic *de novo* mutations reported for generation of animals using other reproductive technologies. This study used whole genome sequence data from 131 three or four generation pedigrees to investigate *de novo* mutagenesis in cattle generated via artificial insemination (AI; N=35), multiple ovulation embryo transfer (MOET; N=44), and in vitro fertilisation (IVF; N=43). Comparison of mosaic *de novo* mutation in the cloned calves described in this study (N=5) to that in cattle generated from AI, MOET, and IVF in the Harland et al. (28) study suggest that the mosaic *de novo* mutation rate in cloned calves may be significantly higher than what is observed with the application of AI (p=0.0097) and MOET (p=0.012), but not significantly higher than that observed with IVF (p=0.13). Acknowledging the comparatively smaller sample size of our study, and the differences in sequencing platforms used between studies, these results support the hypothesis presented by Harland et al. (28), where increased cell handling and intervention may result in increased incidence of *de novo* mutagenesis.

6.3.6.4 De novo structural variants

A case-control design was used in DELLY (v0.8.1) (23) to identify candidate SVs in each cloned calf, but absent in the parental cell line BEF2. Each candidate SV that passed the DELLY quality control filter was manually examined for evidence of legitimate polymorphic structural variation (Table 6.3). The SVs identified in 1802, 1803, 1805 and B071 were all deletions, and 1804 carried two deletions and a duplication. All SVs identified were unique to the calf they were discovered in, suggesting that the SVs arose during development of the calf or in the cell they were derived from. The number of SVs identified in each calf did not appear to be statistically different (smallest p=0.51).

Table 6.3 Number of candidate structural variants (SV) identified in each cloned calf using BEF2 as a reference in DELLY.

Sample ID	SVs identified	Pass quality filter	Legitimate polymorphic SV
1802	21	7	1
1803	31	9	3
1804	39	10	3
1805	38	15	1
B071	32	15	2

6.3.7 Comparison of *de novo* mutations between experimental groups

Given the experimental structure presented, grouping samples by editing status *per se* was confounded by time in culture, where the gene-edited cell line CC14, and cells used to generate the gene-edited cloned calves 1805 and B071 were subject to more cell divisions than controls. This means that additional accumulated mutations are expected in these lines,

though it is nevertheless interesting to attempt to quantify these group effects. Here, we compared the number of mutations observed in the non-edited calves (1802, 1803 and 1804) and gene-edited samples (CC14, 1805 and B071), but absent in the parental cell line BEF2. Candidate mutations were filtered for map quality, and variants that deviated from the expected alternative allele depth were filtered out. As expected, a pair-wise comparison of the mutations observed to be present in each of these groups but absent in BEF2, suggested that CC14, 1805 and B071 collectively carry a greater mutation load (p=2.2×10^{-16}; Table 6.S3). We also compared the number of variants present in BEF2, but absent in the non-edited calves and the gene-edited samples to investigate the potential false discovery rate, but did not find significant differences between experimental groups (p=0.94; Table 6.S4).

6.3.7.1 Comparison of *de novo* mutation distribution and spectra

To further evaluate the candidate *de novo* mutations across experimental conditions, we categorised mutations according to the different stage of their occurrence, and compared mutation distribution and spectra of mutations occurring at each of these stages. *De novo* mutations that arose in cells post plasmid transfection (N=150; Table 6.1) were estimated based on heterozygous *de novo* mutations in the CC14 cell clone that were subsequently inherited by cloned calves B071 and 1805, but absent in all other samples. The number of *de novo* mutations that emerged during cell clone expansion were estimated based on the sum of mosaic *de novo* mutations in the gene-edited cell clone CC14, and heterozygous mutations that were present in any cloned calf, but not in CC14 or the parental cell line, BEF2 (N=1,298; Table 6.1). The smallest proportion of *de novo* mutations (N=58; Table

6.1) arose during the development of the cloned calves. Across the three groups, *de novo* mutations appeared to be randomly dispersed across most of the genome and were not observed to cluster in a group-dependent manner (Fig 6.4a), but a distinct spectra of mutations was observed between *de novo* mutations that were predicted to have arisen in the cloned calves and those that were predicted to have arisen post plasmid transfection or during cell culture (Fig 6.4b). Comparison of mutation spectra between the three groups revealed that C>A mutations were significantly enriched in cells post plasmid transfection and cells in culture for clonal expansion, compared to those in the cloned calves (p=3.213×10^{-6}), accounting for over 40% of total mutations observed in cells at the two stages of *in vitro* cell culture. The cloned calves appeared to be significantly enriched for T>G mutations compared to cells post plasmid transfection and cells in culture (p=3.213×10^{-6}). These mutations accounted for 31% of the total *de novo* mutations observed in the cloned calves. There were no significant differences in mutation spectra observed between cells post plasmid transfection and cells in culture.

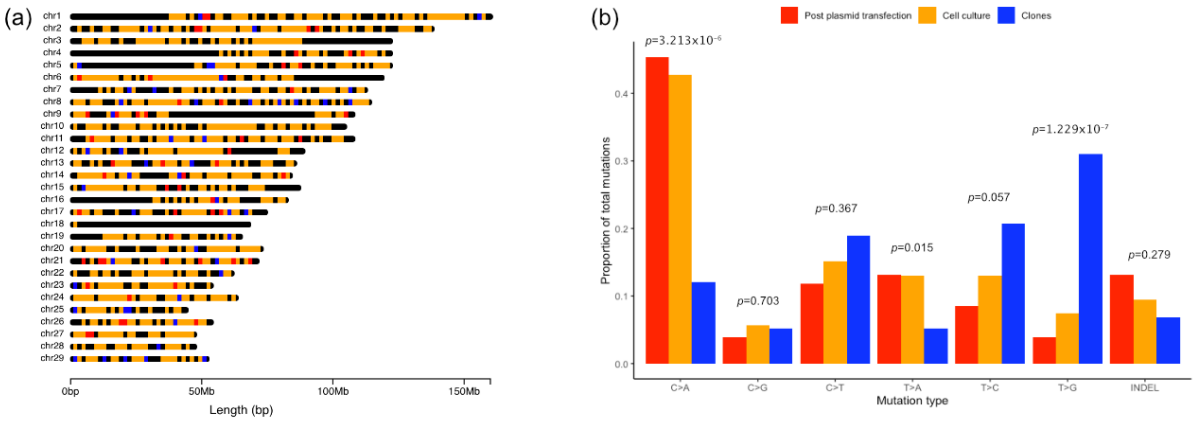


Figure 6.3 Distribution and spectra of de novo mutations predicted to have arisen in cells post plasmid transfection (red), during the cell culture expansion phase (orange) and during development of the cloned calves (blue). (a) Distribution of de novo mutations (SNVs and indels) across the bovine genome. (b) Proportion of de novo mutations within each mutation class observed between groups. A Fisher's exact test comparing the proportion of observed mutations per mutation class between groups, showed a significant difference in C>A mutations ($p=3.213\times10^{-6}$) and T>G mutations ($p=1.229\times10^{-7}$).

6.4 Discussion

We present the first study in cattle based on cloned calves produced by somatic cell nuclear transfer to evaluate unintended off-target mutations, SVs at the on-target site, and unintended integration of the editing plasmid associated with the application of CRISPR-Cas9 mediated gene-editing. Using WGS data and long molecule sequencing, we show that the application of CRISPR-Cas9 to induce a precise 3bp deletion in the *PMEL* gene did not produce detectable off-target events in the gene-edited cell clone (CC14) or the resultant gene-edited calves (1805 and B071). Furthermore, we provide primary evidence to suggest that CRISPR-Cas9 mediated gene-editing does not affect spontaneous mutagenesis or mutation spectra in subsequent cell divisions post-edit, and *de novo* mutagenesis in calves derived from the gene-edited cell clone appears to be equivalent to that of controls.

The filtering criteria that we used in this study was built around our clone-based study design where each sample originated from the same genetic background. The study design, combined with high-depth WGS data enabled direct comparisons between spontaneous mutagenesis that occurred in cells post plasmid transfection, in cell culture during cell clone expansion, and during development of the cloned calves. Application of these filtering criteria combined with manual sequence visualisation revealed no detectable CRISPR-Cas9 induced off-target SNVs, indels or SVs in the gene-edited cell clone or gene-edited cloned calves. Although integration events for circular, supercoiled plasmids are rare (29), it was possible that the whole plasmid, or parts of the editing plasmid, may have integrated at the on-target site or elsewhere in the genome (11,30,31). For this reason, we targeted a broad 8.8kb interval at the on-target site for high-depth long molecule sequencing. Although the results from this did not reveal evidence of an integration event or other structural variation, this did not rule out the possibility of whole or partial integration of the vector at an off-target site. To investigate this possibility, we added the *PMEL*-specific editing plasmid sequence to the reference genome and re-ran our short-read sequence alignment. Several reads from the CC14 gene-edited cell clone were found to map to this sequence, but our follow-up PCR analysis showed that these reads were most likely the result of sample contamination prior to WGS, rather than evidence of an integration event. While somewhat surprising, residual contamination events have been previously reported in other sequencing contexts (32), and so it seems plausible that the contamination noted here may have occurred during one of the many handling steps prior to WGS. Our findings highlight the importance of a methodical approach to investigating plasmid integration events when using double-stranded DNA to deliver editing tools such as CRISPR-Cas9, and the need for carefully designed experiments to ensure fragments of the plasmid do not persist in the

edited genome. Delivering editors as purified proteins (i.e., ribonucleoprotein complexes) could be assumed to minimise this risk and as such represent an appealing alternative to plasmid-based methods.

The number of heterozygous *de novo* mutations varied significantly between calves, but these changes could not be attributed to any single experimental condition (i.e., cell line, gene editing status, or donor cell origin). As we expect heterozygous *de novo* mutations in the cloned calves to have been inherited from their somatic donor cells, these results suggest that the use of CRISPR-Cas9 gene-editing is unlikely affecting the expected spontaneous mutation rate during clonal expansion of the gene-edited cell. The difference in observed heterozygous *de novo* mutations may instead be due to differences in the accumulation of unrepaired DNA damage across cells in culture, which could be induced by oxidative damage, damage due to UV exposure, other mutagens, or mechanical sheer (33,34). Indeed, when we examined the mutation spectra for heterozygous *de novo* mutations, we observed a significant enrichment in C>A mutations, a base-pair transversion associated with cells in culture and thought to be caused by oxidative stress (35–37). By contrast, the mosaic *de novo* mutations observed in the cloned calves appeared to be statistically equivalent regardless of their edited or non-edited status, although it is important to note that some somatic mutations may not be represented in the WGS data or discarded as sequencing errors due to their low-abundance or absence in the sampled tissue. It is also possible that the use of CRISPR-Cas9 may have introduced mutation or epigenetic changes that could affect genome stability. Although we did not see evidence of increased mutagenesis in our cloned calves generated from the expanded gene-edited cell line, epigenetic effects were not investigated, and thus the potential relevance of this class of

changes is unknown. When samples were grouped by experimental treatment, we observed a greater number of mutations in the edited group. However, the CC14 cell clone, and the CC14 derived cells used to generate calves 1805 and B071 underwent a bottleneck event where a single mitotic doublet was isolated and expanded, whereas the cells used to generate the control clones did not. The increased mutation load observed in this group is assumedly due to the greater number of population doubling events these cells experienced (1). Kuijk et al. (37) reported that pluripotent stem cells accumulated 3.5±0.5 mutations per genome, per population doubling. According to these estimates, it would be anticipated that the additional mutations observed in CC14, 1805 and B071 would arise from 16-21.4 additional cell population doubling events, consistent (at minimum) with the number of doubling events that were required to expand the single mitotic doublet to confluency (approximately 5×10^5 cells, and 20 population doubling events). Since our mutation analysis was retrospective to generation of the cloned calves and cell resources, the structure was not ideal to make comparisons between the edited and non-edited experimental groups, where editing status was confounded with time in culture. Thus, a design that also exposed non-edited cell lines to the expanded culture conditions experienced by the edited cell lines would have allowed more direct assessment of the influence of editing and cell culture on mutation rates, and is a design that could be considered for future such analyses.

When the frequency of *de novo* mutations was compared to that observed in cattle generated with the assistance of reproductive technologies such as AI, MOET and IVF (28), we observed that the average number of *de novo* mutations reported for the cloned calves was greater than the average number observed in the other groups, but not

significantly so when compared to IVF. The increased rate of mutagenesis observed in cloning and IVF compared to other reproductive technologies may be due to potentially suboptimal *in vitro* culture conditions and manipulations that are at the centre of these technologies. Analysis of the *de novo* mutation spectra revealed a marked difference in predominant mutation type between the cloned calves and cattle produced by natural matings and other reproductive technologies (38). We observed a significant enrichment in T>G transversion mutations in the cloned calves, where an excess of C>T transition mutations is usually seen in cattle born from a natural mating or other assisted reproductive technologies. The T>G mutation type has been observed to be enriched amongst mouse somatic mutations and thought to be attributable to less effective repair of thymine dimers, but the exact mechanism of mutagenesis remains unconfirmed (39), and why this transversion may be enriched in the cloned calves remains unclear. The results presented here must be interpreted with caution due to the small sample size used for comparison, and a larger dataset will be required to support these findings.

Our results demonstrate that naturally occurring, beneficial genetic variation can be introduced into animals that subsequently show levels of mutagenesis indiscernible from the *de novo* mutation rates of un-edited controls. Although gene-editing technologies such as CRISPR-Cas9 hold potential to accelerate introgression of favourable genetic variants across large populations, the widespread use of such technologies is limited in the agricultural sector due to uncertainties around the level of social acceptance and controversy surrounding perceived safety of the products from gene-edited animals. A major challenge is the ability to detect and quantify off-target mutagenesis above the background *de novo* mutation rate. We can identify candidate off-target events using

logical filtering criteria and evaluate each site for its biological plausibility based on sequence similarity with the on-target site, but it is more difficult to differentiate between off-target mutations and spontaneous *de novo* mutations at regions with little homology to the on-target site. Holstein-Friesian cattle have a baseline spontaneous *de novo* mutation rate of approximately 1.2×10^{-8} mutations per bp, per generation (28). Quantification of off-target mutations at sites of little homology would therefore require tens, potentially hundreds, of extra mutations in edited samples compared to controls, to observe a 'significant' increase in mutation rate above baseline levels. Spontaneous *de novo* mutations have been observed to follow a typical signature, where there is an expected excess of C>T mutations (27). Comparing mutation spectra between edited and non-edited samples may thus prove useful in evaluating the occurrence of unintended mutagenesis, although we are unaware of any studies that have specifically investigated the mutation profile of off-target events induced by gene-editing technologies. Development of sensitive tools that enable accurate detection of genuine off-target events, but also consider natural *de novo* mutation, may be difficult but will aid to establish the risk profile of gene-editing technologies and ultimately support informed consumer decisions.

6.5 Methods and Materials

6.5.1 Animal generation

All animals and cell lines described in the present study were generated as reported by Laible et al. (1). Briefly, male primary bovine fetal fibroblast cells (BEF2) were co-transfected with a modified pX330 transfection vector carrying Cas9 nuclease and *PMEL*-specific gRNA, and a homology-directed repair template using a Neon transfection system

(Invitrogen). Two days post transfection, mitotic doublets were manually selected, reseeded and cultured. Cell clones identified to be homozygous for the targeted 3bp deletion in exon 1 of the *PMEL* gene (p.Leu18del; Chr5:57,345,301-57,345,303bp) were further expanded (CC14). Donor cells from the biallelic cell clone CC14 and the wildtype parental cell line BEF2 were used to generate two *PMEL* gene-edited calves (1805 and B071) and three wildtype calves (1802, 1803 and 1804), respectively, via somatic cell nuclear transfer.

6.5.2 Whole genome sequencing and data analysis

Unedited male primary bovine fetal fibroblast cells (BEF2), edited fetal fibroblast cells homozygous for p.Leu18del (CC14), three control clone calves generated from BEF2, and two gene-edited clone calves generated from CC14 were chosen for whole genome sequencing. Genomic DNA was isolated from CC14 and BEF2 cells, and blood samples from each of the calves using a Nucleon BACC2 kit (Cytiva, Little Chalfont, UK). The genomic DNA samples were sequenced by Macrogen (Seoul, South Korea), with a targeted read depth of 60× per isolate. The samples were sequenced based on 150bp paired reads on the Illumina HiSeq X Ten platform and read mapping was performed using the ARS-UCD1.2 genome build (21) and the BWA MEM v0.7.17 software (40), resulting in mean mapped read depth of 50.7× across the genome (ranging between 44.7× to 54.8× across samples). SNV and indel calling was carried out using Genome Analysis Toolkit (GATK) HaplotypeCaller (v4.0.2.1) using default parameters (22), yielding an unfiltered dataset of 8,021,969 variants across the seven samples. Principal component analysis was conducted in PLINK (v1.9) (41) using whole genome sequence data across the seven samples, and GATK GenotypeConcordance (v4.0.2.1) (42) was used to test pair-wise genomic

concordance between samples using positions extracted based on the Illumina BovineSNP50 genotyping chip.

6.5.3 Identification of off-target mutations

All 8,021,969 variants called by HaplotypeCaller (v4.0.2.1) (22), were dummy coded (0 = no coverage; 1 = homozygous reference; 2= heterozygous; 3 = homozygous alternate). Variants identified to be monomorphic across all samples were removed, sites with no coverage in BEF2 were removed, and all variants present in an unrelated sequenced cattle population previously described by Jivanji et al. (43), and Lopdell et al. (44) (N=564, remapped to the ARS-UCD1.2 genome build (21) yielding 37,208,259 SNPs and 11,746,534 indels) were removed. Candidate off-target mutations were filtered according to the following criteria: (1) candidate mutations should be present in the CC14 cell clone and in both gene-edited clones, but absent in the BEF2 cell clone and three control clones, (2) should have a map quality score of 60, (3) should have an allele dosage of, or statistically equivalent to, 0.5 or 1 for the alternative allele in the CC14 cell clone and both gene-edited clones, and (4) manual inspection of sequence reads should show no evidence of miscalled or misaligned SNVs/indels at the candidate positions. Allele dosage was calculated for each variant by dividing the number of alternate reads by the total number of observed reads at each position. A binomial probability function was used to predict if the allele dosage was statistically equivalent to 0.5 for a heterozygous genotype, with a Bonferroni corrected *p*-value calculated as the significance threshold. In practice, these criteria would highlight a 60× depth site as being a potentially mosaic variant with a 10:50 depth ratio. All candidate off-target mutations were uploaded into IGV for visualisation (45), and the sequence

adjacent each candidate off-target mutation was visually inspected for sequence similarity with the gRNA (ATGGGTGTTCTTCTGGCTGT) and the presence of a 5'-NGG-3' PAM site.

Potential off-target sites were also predicted using Cas-OFFinder software (13). The online Cas-OFFinder tool was used to identify potential off-target mutations by searching the ARS-UCD1.2 genome build (21) for sequence similarity with the gRNA used to target the *PMEL* gene, allowing for up to five mismatches. Candidate off-target mutations predicted by the software were compared to a list of candidate off-target mutations identified by the filtering criteria described above, and also to the unfiltered variants called by GATK HaplotypeCaller (v4.0.2.1) (22).

Candidate SVs that may have arisen due to the application of CRISPR-Cas9 gene-editing were called and filtered using DELLY (v0.8.1) (23). A case-control approach was implemented in DELLY where the CC14 cell clone, clone 1805 and clone B071 were separately called as case samples with the parental cell clone, BEF2, used as the control. After initial SV calling, the wild type clones generated from BEF2 (1802, 1803 and 1804), were added as additional controls to further filter candidate SVs. All candidate mutations were manually inspected in IGV (45) to assess evidence of a legitimate SV at each of these sites.

6.5.4 Long molecule sequencing

Genomic DNA was extracted from cultured bovine cells for samples BEF2 and CC14, and from blood samples for 1805, B071 and 1802 as previously described by Laible et al. (1). Primers were designed to target Chr5:57,340,856-57,349,715bp (Table 6.S5), encapsulating 8,860bp around the *PMEL* on-target site. The PCR was conducted using the KAPA LongRange PCR kit (KapaBiosystems) with the following cycling conditions: 95°C for 3 minutes; 95°C for 30 seconds, 60°C for 30 seconds, and 68°C for 9 minutes for 35 cycles; and a final extension step of 68°C for 9 minutes. The PCR products were loaded and run on a 1% agarose gel for 60 minutes at 100 V to estimate amplicon size. Resultant amplicons were purified using AMPure XP beads and then used to construct a sequencing library using the SQL-LSK109 kit (Oxford Nanopore Technologies) as per the manufacturer's instructions. The library was constructed using 700ng of DNA from across the five samples, loaded onto a FLO-MIN106 flow-cell (Oxford Nanopore Technologies) and sequenced for 10 minutes, yielding an average 590× coverage over Chr5:57,340,856bp-57,349,715bp for each sample. The reads were base-called using Guppy basecaller (v4.0.14) (46), with the samples then separated based on their barcodes by Guppy barcoder (v4.0.14), and subsequently aligned to the ARS-UCD1.2 reference genome (21) plus the *PMEL*-specific CRISPR-Cas9 expression plasmid sequence using minimap2 (v2.14) (24).

6.5.5 Investigation of the presence of the *PMEL*-specific CRISPR-Cas9 expression plasmid

The genomic DNA samples described above were also used for PCR. Two primer pairs were designed across the *PMEL*-specific CRISPR-Cas9 expression plasmid sequence,

where targeted regions were chosen based on mapped short-read WGS data from the CC14 cell clone (Table 6.S5). Each PCR reaction contained two sets of primer pairs at a concentration of 5μM per primer: one primer pair specific for the editing plasmid, and another primer pair targeted to amplify Chr2:110,817,757-110,818,275 (Table 6.S5). The PCR was conducted using the Kapa 2G Fast Hotstart PCR kit (KapaBiosystems) with the following cycling conditions: initial denaturation at 95°C for 3 minutes; denaturation at 95°C for 15 seconds, anneal at 60°C for 15 seconds, extend at 72°C for 15 seconds, for a total of 35 cycles; and final extension at 72°C for 1 minute.

6.5.6 Identification of *de novo* mutations

De novo SNVs and indels unique to each sample were identified using a filtering criteria similar to that described above for identifying off-target mutations. As described above, variants were initially filtered for monomorphic sites, sites missing in BEF2 and alleles already identified to segregate in a sequenced NZ dairy cattle population. From the remaining variants, *de novo* mutations were identified by the following criteria: (1) keeping heterozygous SNVs and indels specific to each sample; (2) filtering to remove reads with a map quality score less than 60; (3) classifying SNVs and indels as heterozygous or mosaic *de novo* mutations, where mosaic variants were defined as having an allele dosage significantly less than 0.5, as determined by the binomial probability function described previously; (4) filtering variants based on manual examination of sequence alignments in IGV to remove misaligned or miscalled SNVs and indels. Sequence alignments were also examined at each candidate mosaic SNVs and indel site for evidence of more than one bi-

allelic variant segregating on the sequence read, or read pair, that could indicate the presence of three haplotypes and support mosaicism.

Comparison of mutations in the control calves (1802, 1803, and 1804) and the gene edited samples (CC14, 1805, and B071) were conducted using the following filtering criteria: (1) (a) all variants identified to be heterozygous or homozygous alternate by GATK HaplotypeCaller (v4.0.2.1) (22) in the group of samples, but homozygous reference in BEF2 were kept, OR (b) all variants that were identified to be heterozygous or homozygous alternate by GATK HaplotypeCaller in BEF2, but homozygous reference in the group of samples were kept, (2) variants with a map quality of less than 60 were filtered out, and (3) variants called as homozygous reference by GATK HaplotypeCaller, but had an alternative allele depth greater than 0 were filtered out (i.e. likely called incorrectly as homozygous reference by GATK HaplotypeCaller).

The structural variant analysis was conducted using a case-control study design in DELLY (v0.8.1) (23). Each cloned calf was compared to the parental cell line BEF2 to identify candidate SVs, which were then filtered based on the DELLY quality control filter. All candidate SVs that passed the quality filter were manually examined by visualisation of sequence reads in IGV, and were either confirmed or rejected as legitimate polymorphic structural variants.

All pair-wise comparisons of *de novo* mutation rates reported in this study were conducted using a two-proportions Z-test, and comparisons of *de novo* mutation spectra were conducted using Fisher's exact test.

6.6 Declarations

6.6.1 Ethics approval

All experiments were conducted in strict accordance with the rules and guidelines outlined in the New Zealand Animal Welfare Act 1999, with approvals from New Zealand's Environmental Protection Authority (GMD100279) and the Ruakura Animal Ethics Committee (14236).

6.6.2 Consent for publication

Not applicable.

6.6.3 Availability of data and materials

Whole genome sequence data used in this study are available at NCBI-SRA under BioProject ID PRJNA701980.

6.6.4 Competing interests

CH and ML are employees of Livestock Improvement Corporation, a commercial provider of bovine germplasm. The remaining authors declare that they have no competing interests.

6.6.5 Funding

This work was supported by the Ministry of Business, Innovation and Employment Endeavor Funds CONT-62639-ENDRP-AGR and CONT-57639-ENDRP-LIC (Wellington,

New Zealand). External funders had no role in the design of the experiment, the analysis or interpretation of the data, or writing of the manuscript.

6.6.6 Authors' contributions

SJ performed most of the bioinformatic and statistical analysis with help from CH; SJ designed and performed most of the lab-based experimental work with help from RS and ML; GL, SC and BB conducted DNA extractions and generated the dataset; SJ, GL, ML and DG conceived the study and experiments; GL and ML secured funding for the project; GL, ML, RS, DG were involved in supervision of the project; SJ wrote the manuscript; GL, ML, DG and CH were involved in reviewing and editing the manuscript.

6.6.7 Acknowledgments

The authors would like to acknowledge Massey University and Livestock Improvement Corporation (LIC) for their support in this research, AgResearch for the generation of this dataset, The University of Auckland for access to laboratory resources, and the New Zealand eScience Infrastructure (NeSI) for providing the computational resources required for the analyses described here. This work was funded by AgResearch and the Ministry of Business, Innovation and Employment.

6.7 References

1. Laible G, Cole S-A, Brophy B, Wei, Leath S, Jivanji S, et al. Holstein Friesian dairy cattle edited for diluted coat color as a potential adaptation to climate change. BMC

Genomics. 2021; 22(856):1-12.

2. Bishop TF, Van Eenennaam AL. Genome editing approaches to augment livestock breeding programs. Vol. 223, Journal of Experimental Biology. Company of Biologists Ltd; 2020.
3. Cong L, Zhang F. Genome engineering using CRISPR-Cas9 system. Chromosom Mutagen Second Ed. 2014;8(11):197–217.
4. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816–22.
5. Pattanayak V, Lin S, Guilinger JP, Ma E, Doudna JA, Liu DR. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. Nat Biotechnol. 2013;31(9):839–43.
6. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. 2013;31(9):822–6.
7. Zhang X-H, Tee LY, Wang X-G, Huang Q-S, Yang S-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Mol Ther - Nucleic Acids. 2015;4:e264.
8. Wang X, Liu J, Niu Y, Li Y, Zhou S, Li C, et al. Low incidence of SNVs and indels in trio genomes of Cas9-mediated multiplex edited sheep. BMC Genomics. 2018;19(1):397.
9. Korablev A, Lukyanchikova V, Serova I, Battulin N. On-Target CRISPR/Cas9 Activity Can Cause Undesigned Large Deletion in Mouse Zygotes. Int J Mol Sci. 2020;21(10):3604.

10. Kosicki M, Tomberg K, Bradley A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol. 2018;36(8):765–71.

11. Young AE, Mansour TA, McNabb BR, Owen JR, Trott JF, Brown CT, et al. Genomic and phenotypic analyses of six offspring of a genome-edited hornless bull. Nat Biotechnol. 2020;38(2):225–32.

12. Chakraborty S. Unreported off-target integration of beta-lactamase from plasmid in gene-edited hornless cows. OSF Preprints. 2019.

13. Bae S, Park J, Kim J-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics. 2014;30(10):1473–5.

14. Xiao A, Cheng Z, Kong L, Zhu Z, Lin S, Gao G, et al. CasOT: A genome-wide Cas9/gRNA off-target searching tool. Bioinformatics. 2014;30(8):1180–2.

15. Zhu H, Misel L, Graham M, Robinson ML, Liang C. CT-Finder: A web service for CRISPR optimal target prediction and visualization. Sci Rep. 2016;6(1):1–8.

16. Li C, Zhou S, Li Y, Li G, Ding Y, Li L, et al. Trio-Based deep sequencing reveals a low incidence of off-target mutations in the offspring of genetically edited goats. Front Genet. 2018;9:449.

17. Schaefer K, Wu W, Colgan D, Tsang S, Bassuk A, Mahaja V. Unexpected mutations after CRISPR–Cas9 editing in vivo. Nat Methods. 2017;14(6):547.

18. Wu WH, Tsai YT, Justus S, Lee TT, Zhang L, Lin CS, et al. CRISPR repair reveals causative mutation in a preclinical model of retinitis pigmentosa. Mol Ther. 2016;24(8):1388–94.

19. Schaefer KA, Darbro BW, Colgan DF, Tsang SH, Bassuk AG, Mahajan VB.

Corrigendum and follow-up: Whole genome sequencing of multiple CRISPR-edited mouse lines suggests no excess mutations. bioRxiv. 2017;154450.

20. Schmutz SM, Dreger DL. Interaction of *MC1R* and *PMEL* alleles on solid coat colors in Highland cattle. Anim Genet. 2013;44(1):9–13.

21. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9(3):giaa021.

22. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GA Van der, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017;201178.

23. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28(18):i333–9.

24. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100.

25. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

26. Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. Nucleic Acids Res. 2013;41(W1):W29–33.

27. Harland C, Charlier C, Karim L, Cambisano N, Deckers M, Mni M, et al. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. bioRxiv. 2016;079863.

28. Harland C, Durkin K, Artesi M, Karim L, Cambisano N, Deckers M, et al. Rate of de

novo mutation in dairy cattle and potential impact of reproductive technologies. Proc World Congr Genet Appl to Livest Prod. 2018.

29. Würtele H, Little KCE, Chartrand P. Illegitimate DNA integration in mammalian cells. Vol. 10, Gene Therapy. Nature Publishing Group; 2003. p. 1791–9.

30. Graham C, Cole S, Laible G. Site-specific modification of the bovine genome using Cre recombinase-mediated gene targeting. Biotechnol J. 2009;4(1):108–18.

31. Kim S, Kim D, Cho SW, Kim J, Kim JS. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. Genome Res. 2014;24(6):1012–9.

32. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. PLoS One. 2014;9(5):e97876.

33. Gundry M, Li W, Maqbool SB, Vijg J. Direct, genome-wide assessment of DNA mutations in single cells. Nucleic Acids Res. 2012;40(5):2032–40.

34. Kim M, Rhee JK, Choi H, Kwon A, Kim J, Lee GD, et al. Passage-dependent accumulation of somatic mutations in mesenchymal stromal cells during in vitro culture revealed by whole genome sequencing. Sci Rep. 2017;7(1):1–10.

35. Koh G, Zou X, Nik-Zainal S. Mutational signatures: Experimental design and analytical framework. Vol. 21, Genome Biology. BioMed Central Ltd.; 2020.

36. Behjati S, Huch M, Van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. Nature. 2014;513.

37. Kuijk E, Jager M, van der Roest B, Locati MD, Van Hoeck A, Korzelius J, et al. The mutational impact of culturing human pluripotent and adult stem cells. Nat Commun.

2020;11(1):1–12.

38. Harland CS. Germline mutations in Bos taurus. Universite de Liege; 2018.

39. Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. Nat Commun. 2017;8(1):1–8.

40. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.

41. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.

42. Van der Auwera G, O'Connor B. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition). O'Reilly Media; 2020.

43. Jivanji S, Worth G, Lopdell TJ, Yeates A, Couldrey C, Reynolds E, et al. Genome-wide association analysis reveals QTL and candidate mutations involved in white spotting in cattle. Genet Sel Evol. 2019;51(1):1–18.

44. Lopdell TJ, Tiplady K, Struchalin M, Johnson TJJ, Keehan M, Sherlock R, et al. DNA and RNA-sequence based GWAS highlights membrane-transport genes as key modulators of milk lactose content. BMC Genomics. 2017;18(1):1–18.

45. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.

46. Wick RR, Judd LM, Holt KE. Performance of neural network basecalling tools for Oxford Nanopore sequencing. Genome Biol. 2019;20(1):129.

6.8 Appendix

Additional File 1 Table 6.S1

Format: .csv

Title: Predicted and candidate off-target mutations. All predicted and candidate off-target mutations identified by our filtering criteria, with additional information about mutation type, genes that the mutations may map within, and the predicted variant effect.

Available at: https://static-content.springer.com/esm/art%3A10.1186%2Fs12864-021-07804-x/MediaObjects/12864_2021_7804_MOESM1_ESM.csv

Table 6.S2 Structural variants (SVs) identified in the gene-edited cell line (CC14) and gene-edited cloned calves (1805 and B071) using DELLY with the parental cell line (BEF2) and non-edited cloned calves (1802, 1803 and 1804) as reference samples.

	CC14	**1805**	**B071**
BEF2 as reference sample	27	38	32
1802, 1803 and 1804 added as reference samples	1	5	5
SVs common between gene-edited samples	0	0	0

Table 6.S3 Number of variants remaining after each filter to determine their presence in control calves (1802/1803/1804), but absence in parental cell line BEF2, and presence in gene edited samples (CC14/1805/B071), but absence in BEF2. Variants were kept if they (1) were called as heterozygous or homozygous alternate by GATK HaplotypeCaller in group, but homozygous reference in BEF2, (2) had a map quality of 60, and (3) called as homozygous reference in BEF2 and had an alternate allele depth of 0.

Group	Heterozygous or homozygous alternate in group samples but homozygous reference in BEF2	Quality-filtered variants	Alternate allele depth = 0 in BEF2
1802/1803/1804	972	256	60
CC14/1805/B071	1338	452	253

Table 6.S4 Number of variants remaining after each filter to determine their presence in parental cell line BEF2, but absence in control calves (1802/1803/1804), and presence in BEF2, but absence in gene edited samples (CC14/1805/B071). Variants were kept if they (1) were called as heterozygous or homozygous alternate by GATK HaplotypeCaller in BEF2, but homozygous reference in group samples, (2) had a map quality of 60, and (3) called as homozygous reference in group samples and had an alternate allele depth of 0.

Group	Homozygous reference in group samples but heterozygous or homozygous alternate in BEF2	Quality-filtered variants	Alternate allele depth = 0 in group samples
1802/1803/1804	2256	759	80
CC14/1805/B071	2073	622	82

Table 6.S5 Description of PCR primer pairs designed to investigate the on-target site and plasmid integration

	Sequence	Melting Temp (˚C)	PCR product size	Position
Long-range PCR primers				
Forward primer	GTGCCACTGACATGTAGCAAAG	60.8	8,860bp	Chr5:57,340,856-57,349,715bp
Reverse primer	CCCTCCTCAGTCCTTACCAGTA	59.6		
Vector integration PCR primers				
Set 1				
Forward primer	TGACGTTGGAGTCCACGTTC	62.1	757bp	gRNA/Cas9 plasmid:6,263-7,019bp
Reverse primer	TCTTCGGGGCGAAAACTCTC	63.9		
Set 2				
Forward primer	AGATCAGTTGGGTGCACGAG	61.3	690bp	gRNA/Cas9 plasmid:6,939-7,628bp
Reverse primer	TGACTCCCCGTCGTGTAGAT	60.5		
Internal control PCR primers				
Forward primer	ATGTTAGGTGCAGGTGGAGC	60.1	519bp	Chr2:110,817,757-110,818,275bp
Reverse primer	GCTTCCCACCTTGACCTCTC	61.2		

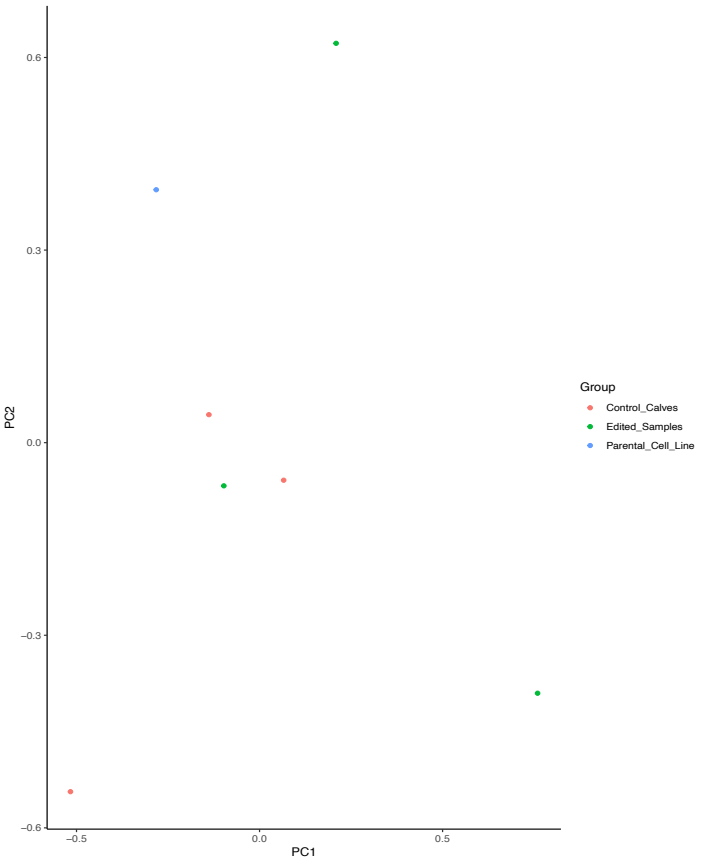


Figure 6.S1 First two principal components (PC) for non-edited control calves (1802/1803/1804), gene edited samples (CC14/1805/B071), and the parental cell line BEF2 plotted against each other revealed no clustering by treatment group.

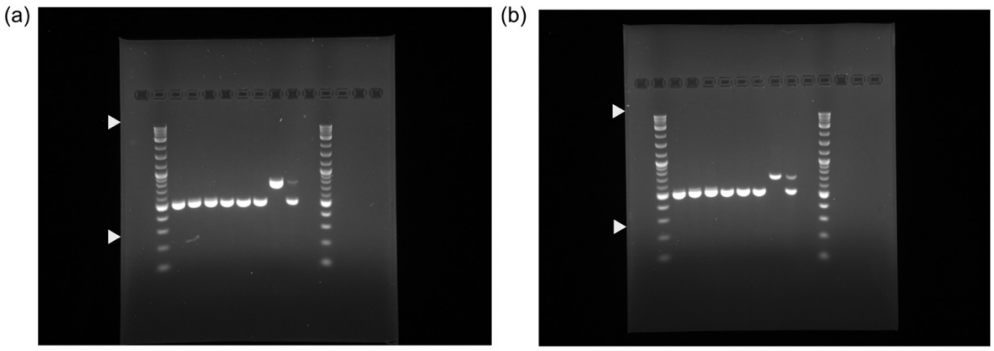


Figure 6.S2 Uncropped, full-length versions of the gels presented in 'Fig6.3'. (a) and (b) correspond to Fig 3(a) and 3(b), where the white arrows indicate where the gel image was horizontally cropped. Images were also cropped vertically to remove the background.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

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:	Swati Jivanji
Name/title of Primary Supervisor:	Dorian Garrick
In which chapter is the manuscript /published work:	Chapter Six

Please select one of the following three options:

(●) The manuscript/published work is published or in press

- Please provide the full reference of the Research Output:

Jivanji S, Harland C, Cole S, Brophy B, Garrick D, Snell R, Littlejohn M, Laible G. The genomes of precision edited cloned calves show no evidence for off-target events or increased de novo mutagenesis. BMC Genomics 2021 221. 2021;22(1):1–14.

( ) The manuscript is currently under review for publication – please indicate:

- The name of the journal:

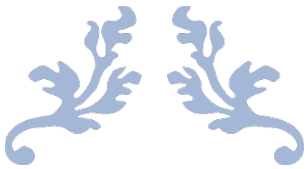
- The percentage of the manuscript/published work that was contributed by the candidate:

- Describe the contribution that the candidate has made to the manuscript/published work:

( ) It is intended that the manuscript will be published, but it has not yet been submitted to a journal

Candidate's Signature:	Swati Jivanji
Date:	25/11/2021
Primary Supervisor's Signature:	Dorian Garrick
Date:	25 Nov 2021

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.



CHAPTER SEVEN

General Discussion

7.1 General overview

In this thesis we implicated the involvement of *PAX3*, *MITF*, and *KIT* in the proportion of white spotting and for the first time, presented candidate causal mutations that likely underlie these effects. We provide evidence to suggest causality of a protein-coding mutation in the *PAX3* gene, an expression-modifying mutation upstream of the *MITF* gene, and a novel, expression-modifying 6.9 kb deletion upstream of the *KIT* gene as modulators of the proportion of white spotting in cattle. Further, we demonstrated multiple complex, epistatic interactions between variants at the *KIT* and *MITF* loci. These include an interaction between two variants at the *KIT* locus that result in pigmentation around the eyes in otherwise white-faced cattle, and an interaction between a serial duplication upstream of *KIT* and the previously proposed expression-modifying *MITF* mutation, which causes adulteration of the same white-face trait characteristic of Hereford cattle. These findings provide insight into the biology of pigmentation, and the evolutionary history of white spotting in modern cattle breeds. We also explored the application of gene-editing as a technology to introgress a coat-colour dilution mutation typically observed in Galloway and Highland cattle into a Holstein-Friesian background. The study describing the generation of the gene-edited calves was not presented in this thesis, however our work provides proof of concept that the gene-editing system clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 did not generate perceptible off-target effects, and thus could be applied to accelerate genetic gain if the desired trait is mono- or oligogenic, and if the causal mutation(s) is known. Although unbiased interrogation of whole-genome sequence data did not detect differences in mutation load between treatment groups, the mutational signature observed in all calves generated by cloning differed from what would be expected if the calves were born from natural matings, or other reproductive

technologies such as artificial insemination (AI), *in vitro* fertilisation (IVF), or multiple ovulation and embryo transfer (MOET). This study, presented in Chapter Six, highlighted considerations for future applications of gene-editing, cloning, and cell culture experiments in cattle, and findings regarding assisted reproductive technologies that may have broader implications in human reproductive medicine.

The chapters in this thesis include chapter-specific discussion sections, so in this chapter I will discuss the overarching themes relevant across the results presented throughout this thesis more broadly. I will touch on reference genome considerations when investigating causal mutations, the contribution of structural variation and epistasis to coat patterning traits, the selection utility of coat patterning traits, and finally theorise on the future of artificial selection.

7.2 Reference genome considerations for causal mutation discovery

White spotting is a common coat patterning trait observed in many cattle breeds such as Holstein, Friesian, Hereford, Normande, Maine-Anjou, and Montbeliarde cattle. In our investigation of the proportion of white spotting in Holstein-Friesian, Jersey, and crossbreed cattle, we first identified candidate causal mutations for the genes *PAX3* and *MITF* (Chapter Three), and then later for the *KIT* gene (Chapter Four). In Chapter Three, our analyses were based on whole-genome sequence data mapped to the UMD3 bovine reference genome. We observed a highly dispersed association signal at the chromosome 6 locus and hypothesised that efforts to uncover the causal mutation at this locus were confounded by errors in the genome assembly around the *KIT* gene. This

hypothesis was supported by previous observations made by Whitacre (1) who suggested that the region around the *KIT* gene was misassembled in the UMD3 reference genome, due to the presence of misrepresented repeat regions. It is possible that genome assembly through this region was made difficult by a large number of co-locating genomic repeats, which are difficult to resolve using the clone-by-clone sequencing of bacterial artificial chromosome clones, and whole-genome shotgun sequencing applied to construct the UMD3 reference genome (2). A new bovine reference genome, ARS-UCD1.2, was compiled and released in 2018 (3). Construction of the new reference genome took advantage of long-read sequencing technologies that were sufficient to span large genomic repeats. The new assembly represented a 200-fold improvement in sequence continuity compared to the UMD3 reference assembly, and a 10-fold improvement in per-base accuracy across the genome (3). Despite these improvements in sequence continuity and per-base accuracy, we were still unable to identify any compelling candidate causal mutations for the proportion of white spotting at the chromosome 6 locus using association analyses and short-read sequence data alone.

Notably, all of the 'white-increasing' alleles identified in this thesis were the reference alleles, including the 6.9 kb deletion allele upstream of the *KIT* gene. Given the phenotype, these alleles can be assumed to be the derived (versus ancestral) alleles, since pigmented skin and hair are the wild-type states in most breeds (4). In the case of the 6.9 kb structural variant, the absence of the long (i.e., ancestral) form likely complicated its detection, since large insertions are more difficult to visualise than deletions (5). While we might not have contemplated the fact at the start of this project, the widespread use of a Hereford reference genome may be the reason why the white-

spotting causal mutations have remained elusive to the bovine research community for so long. Had the reference genome been based on a solid-coloured animal (i.e., a more representative wild-type), the candidate causal mutation upstream of the *KIT* gene would have likely presented as a straightforward deletion in Holstein cattle, and may have been easier to detect by association analysis and short-read sequence data, even with errors in the genome assembly around the *KIT* gene. Indeed, remapping of short-read sequence data to our bespoke reference genome that incorporated the 6.9 kb sequence upstream of the *KIT* gene, enabled genotyping cattle for the structural variant with relative ease. Incorporation of the structural variant genotypes into an association analysis supported strong association between the 6.9 kb deletion and the proportion of white spotting on the coat, with the 6.9 kb deletion variant appearing close to the top of the association peak (Chapter Four). Furthermore, Hereford cattle have the mutant phenotype (i.e., white spotting), but where these cattle typically have a white face, and a white belly, the white-spotting pattern characteristic of Holstein cattle typically manifests as white on the legs and belly, accompanied by large well-defined white spots distributed somewhat randomly across body (Fig 7.1a & b). Due to these differences in phenotype presentation, we expected that Holstein cattle would carry different alleles at some, if not all, of the causal mutation sites relative to the Hereford reference genome, so the discovery that all 'white-increasing' alleles were indeed reference alleles was surprising. As demonstrated throughout this thesis, and discussed below in 'Epistasis between coat colour loci', the genetic background on which a variant is inherited may alter how these variants influence coat patterning traits. It is possible that although Hereford cattle carry the same variants that cause white spotting in Holstein cattle, interaction between these candidate causal mutations with other as yet unknown genetic variants, may alter the way white spotting manifests in Hereford cattle (i.e., white face

and white belly rather than the white-spotting pattern typical of Holsteins). Further work is required to better understand how the inheritance of candidate causal mutations at all three major white spotting QTL may result in vastly different coat patterning traits between Holstein and Hereford cattle.

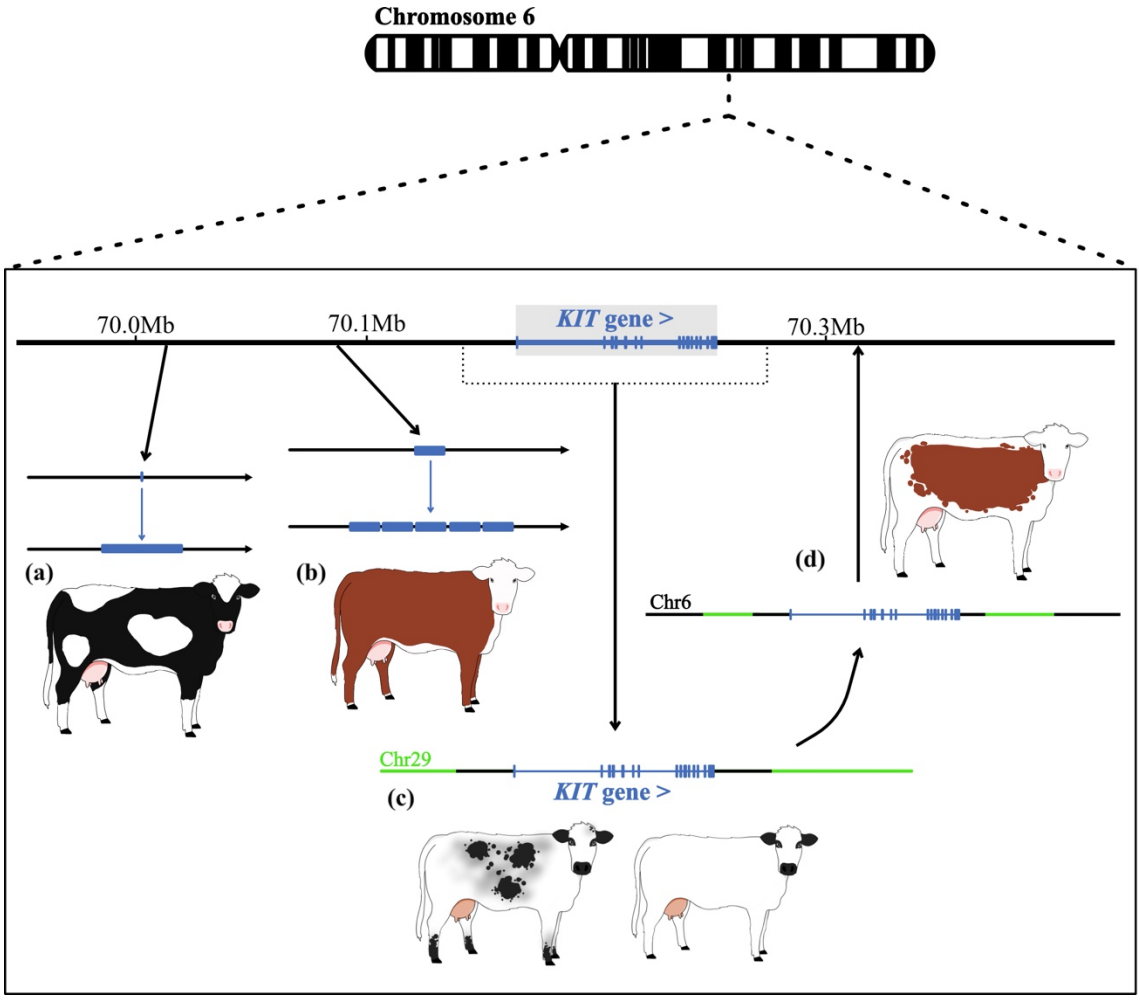


Figure 7.1 Zoomed in representation of the KIT locus on chromosome 6 annotated with the KIT gene and structural variants implicated in coat patterning traits. The KIT gene is depicted in blue, and highlighted in grey. The exons are represented by vertical lines, introns by a horizontal line, and the direction of transcription indicated by the symbol adjacent the gene name. (a) An insertion/deletion at 70.03Mb is associated with white spotting on the coat, (b) a serial duplication at 70.09Mb is associated with the white-face trait in Hereford cattle, (c) translocation of a segment from chromosome 6 encompassing the KIT gene to chromosome 29 is associated with the speckling (heterozygous) and white with pigmented points (homozygous) traits in White Galloway and White Park cattle, and (d) the subsequence translocation of this region, along with some adjacent sequence from chromosome 29, back into a

region downstream of the KIT gene at ~70.33Mb is associated with the colour-sidedness trait. All genome coordinates depicted here are relative to the ARS-UCD1.2 reference genome.

Limitations to causal mutation discovery imposed by the reference genome may not be restricted to the investigation of coat colour and could be relevant for the investigation of a variety of traits. To most effectively discuss the variation in any trait, it would perhaps be more informative to explain these effects relative to a more appropriate reference genome (i.e., a more representative wild-type). However, the Hereford cow Dominette was chosen for the reference assembly due to limitations in sequencing technologies, namely: the inability to represent structural polymorphic haplotypes, repeat regions, and the computational resources demanded at the time of genome assembly. The Hereford cow utilised was highly inbred, and therefore had limited contrast between the two parental alleles, enabling the creation of a somewhat representative haploid reference genome for a diploid organism (2). Later, Dominette was again chosen for the construction of the ARS-UCD1.2 reference genome so that existing genomic studies could be easily translated between genome builds, and to minimise disruption of ongoing studies (3). Future research could aim to produce a pan-genome reference that represents DNA segments from a variety of commonly used cattle breeds, as has been done in humans (6). The pan-genome approach uses a graph format that can easily be amended with minimal disruption to the "main" genome, and can therefore be improved by adding population-specific variant sites or haplotypes as they are discovered. The use of such a reference genome would be translatable across projects, research groups, and breeds, and, in theory, better facilitate the discovery of breed-specific causal variants.

7. 3 The *KIT* gene, coat patterning, and structural variation

Structural variation at the *KIT* locus accounts for a large proportion of mutations associated with coat patterning traits in livestock. The KIT receptor plays a central role in mast, germ, haematological, and melanocytic cell systems during embryonic development and in maintaining homeostasis (7). Mutations in the coding regions of *KIT* have been associated with gastrointestinal stromal tumours, mast cell disease, and piebaldism in humans, and homozygous loss-of-function mutations at the mouse *kit* locus are often lethal or sublethal due to their detrimental effects on haematosis (8,9). The majority of the *KIT* locus structural variants observed in livestock do not appear to disrupt coding regions of the *KIT* gene, and give rise to spectacular and visually striking coat patterning traits without causing negative pleiotropic effects (1,10). At least four such structural variations have been reported at the bovine *KIT* locus. A serial duplication upstream of the *KIT* gene was proposed to cause a white face in Hereford cattle (1), and one copy of a duplication and subsequent translocation of a large sequence encompassing the *KIT* locus to chromosome 29 in Galloway cattle causes a speckling pattern, while two copies cause a white coat with black pigmented nose, ears, and feet (11). A reciprocal translocation of this sequence back to chromosome 6 causes the colour-sidedness trait in Brown Swiss and Belgain Blue cattle (10), and we now propose that deletion of a 6.9 kb segment approximately 114 kb upstream of the *KIT* gene causes the Holstein white spotting pattern (Fig 7.1), and contributes to the white markings of several other breeds, including Hereford cattle. A similar phenomenon has been observed in pigs (12,13), goats (14), and horses (15,16), where a diverse range of coat patterning traits appear to be impacted by an over-representation of *KIT* structural variations. Non-coding regulatory elements that influence *KIT* gene expression have been reported to map hundreds of kb both upstream and downstream of the coding

sequence, however the exact number of regulatory elements that influence *KIT* expression is unknown, and none have been annotated on the bovine genome (17). These distal regulatory elements may explain how a range of structural variants at a single locus can create a plethora of patterning traits without negative pleiotropic effects. However, the molecular mechanisms through which such mutations may create a diverse range of coat patterning traits are poorly understood. Future studies could aim to conduct cell imaging experiments of single melanoblast clones edited for these mutations, as previously done by Mort et al. (18), to assess the influence of these mutations on melanoblast physiology during development.

Many of the structural variants characterised to-date have been associated with low-copy repeat sequences, also referred to as segmental duplications (19). It is hypothesised that two lengths of DNA that have a high degree of sequence similarity, but are not alleles of one another (i.e., non-allelic), sometimes become aligned during mitosis, or meiosis. Misalignment of these regions cause non-allelic homologous recombination (NAHR), and often results in duplication and/or deletion events (19). Furthermore, sequence analysis of NAHR events suggest that low-copy repeats in proximity to palindromes, minisatellites, or transposons, where double-stranded breaks in the DNA could occur, are more likely to be NAHR hotspots (20). Olfactory and innate immunity gene family loci are examples of NAHR hotspots, and although others may exist, they may be going undetected due to having deleterious effects and not persisting at high frequencies in the population, or due to limitations in our ability to detect them (21). We observed the presence of a low-copy number repeat flanking the 6.9 kb deletion discovered in Chapter Four, and a similar observation was reported for the structural variant that causes the colour-sidedness and speckled traits in cattle (10). Although we

did not study this phenomenon directly, previous genome assembly difficulties seem to imply an abundance of repetitive elements at the *KIT* locus. This could make the *KIT* locus a hotspot for structural variation by NAHR. The enrichment in structural variation at the *KIT* locus compared to other well-studied loci may support our hypothesis (1,10,11). Indeed, the number of causal structural variants implicated in more commonly studied phenotypes, such as lactation and fertility traits, appear to be limited (22). We concede that there may be some level of ascertainment bias operating here since major structural effects may be more easily identified in Mendelian contexts (such as coat colour and patterning traits) as there is inherently more power to classify animals on this basis and identify candidate mutations. Of note, some NAHR derived structural variants have been observed in unrelated individuals, and are thought to have a recurrent nature (23). Although it seems likely that the 6.9 kb deletion we discovered in Chapter Four occurred in a common ancestor of many modern-day spotted breeds, the breakpoints of the deletion were not carefully curated in breeds other than Holstein, Jersey, and Hereford, so it would be interesting to investigate the possibility of recurrent, otherwise independent, NAHR events in the breeds that presented with the deletion.

7.4 Epistasis between coat colour loci

The phenomenon of epistasis can be described as an interaction between different genetic loci that influence a trait in a way that deviates from a simple Mendelian or additive mode of inheritance (24). Epistasis has been of great interest in the field of quantitative genetics, but statistical epistasis is difficult to detect and measure. We hypothesise that all of the coat patterning traits investigated in this thesis result from some degree of epistatic interaction between coat colour loci: white spotting in Holstein

cattle, splotchy face in Hereford crossbred calves, ambilateral circumocular pigmentation in Hereford cattle, and the grey and white spotting in *PMEL* gene-edited Holstein calves.

It is noteworthy that the proportion of white spotting trait investigated in Chapters Three and Four cannot be predicted by the sum of effect sizes estimated across each implicated quantitative trait locus (QTL). The effect sizes estimated by additive mixed linear models from across the candidate causal mutations, rs209784468 at the *MITF* locus, p.Thr424Met at the *PAX3* locus, and the 6.9 kb deletion at the *KIT* locus, suggest that having the maximum number of white-increasing alleles would make an animal approximately 56.5% white. In Chapter Three the average proportion of white spotting observed in F_2 Holstein-Friesian × Jersey cows with the maximum number of white-increasing alleles was 32.6%. Although we know that white spotting patterns in mammals is in part stochastic, this does not explain the large difference between the expected and observed proportion of white spotting in these cattle (18). Indeed, Feldmann et al. (25) also commented on the overestimation of our reported effect sizes derived from additive mixed-linear association models. However, we also observed that at least three white-increasing alleles from across the three identified quantitative trait loci (QTL) were required to see any white spotting in the F_2 crossbreed cows (i.e., an apparent, non-additive threshold effect). It is thus likely that the three white spotting loci act via an epistatic mechanism to influence the proportion of white spotting on the coat. Epistatic interactions are difficult to detect and measure using statistical methods, so this non-additive mode of inheritance likely caused overestimation of QTL effect sizes in our association models. This observation is not entirely surprising when we consider the interdependent roles PAX3, MITF and KIT play in pigmentation. The

PAX3 protein is a transcription factor necessary for *MITF* transcription, and MITF is a transcription factor that modulates *KIT* expression. Feedback regulation between these molecules has also been demonstrated (26). The p.Thr424Met missense mutation in the *PAX3* gene likely alters PAX3 binding efficiency to its transcription factor binding site, modulating *MITF* transcription (27). The A allele at rs209784468 is also hypothesised to modulate *MITF* transcription efficiency, based on its location immediately upstream of the annotated transcription start site, and its high site-wise conservation score (Chapter Three). Given the sign of effect (i.e., more spotting), this allele could be assumed to reduce MITF expression. The MITF transcription factor, also referred to as the 'master transcription regulator' of the melanocyte lineage, regulates genes for pigmentation enzymes and lineage survival factors, including the *KIT* gene (28). Although we hypothesised that the 6.9 kb deletion upstream of the *KIT* gene knocks out a MITF transcription factor binding site, our observation that a homozygous genotype at the structural variation site is not sufficient to cause the manifestation of white spotting on its own, suggests that there is likely more than one MITF transcription factor binding site that regulates *KIT* transcription. Therefore, reduced *MITF* expression likely has flow-on effects to *KIT* expression, which ultimately alters melanoblast physiology and results in white spotting of the coat.

The Hereford white-face mutation has a dominant effect, where the inheritance of one copy of the candidate causal serial duplication upstream of the *KIT* gene is sufficient to cause the white-face trait (1). This mutation is proposed to modulate *KIT* expression and result in a complete lack of pigmentation in the face. We observed an association between the *MITF* rs209784468 variant (highlighted above as the candidate mutation impacting the proportion of white spotting at the chromosome 22 locus) and a broken

white face in Hereford crossbreed cattle (Chapter Four). The inheritance of one copy of the G allele at rs209784468 appeared sufficient to partially rescue pigmentation of the face in these cattle. Thus the interaction between the rs209784468 variant and the white-face mutation is another example of an epistatic interaction. We did not have any individuals in our study that had two copies of the rs209784468 G allele, but future studies could aim to investigate how different genotype combinations at these loci interact to alter face colour. We also investigated the penetrance of pigmentation around the eyes (i.e., ambilateral circumocular pigmentation; ACOP) in white-faced American Herefords, where we found an association signal that co-located to the *KIT* locus (Chapter Five). Although we were unable to fine map candidate causal mutations at this locus, these results implied that there was a mutation at the *KIT* locus that interacts with the white-face mutation and rescues pigmentation, but only around the eyes. These results suggest complex epistatic interactions at the *KIT* locus that create varied coat patterning traits through molecular mechanisms that are poorly understood. These observations further emphasise the relevance of the reference genome in interpretation of 'mutant' versus 'wild-type' effects. While the Hereford white face is often referred to as a dominantly inherited trait, this only appears to be true in crosses of white-spotted breeds, since it is the ancestral allele at other loci that are responsible for these epistatic effects. In this way, the white face trait is not technically dominant, as Hereford crosses of 'wild-type' (i.e., solid-coloured) breeds would not be expected to enable expression of the trait.

The coat colour and patterning traits observed in Holstein-Friesian calves edited for a coat colour dilution mutation in the *PMEL* gene (p.Leu18del), deviated from what we would have expected. The *PMEL* dilution mutation has a codominant mode of

inheritance, where one copy of the mutation causes some dilution of the base coat colour, and two copies cause complete coat-colour dilution (29). In black Highland cattle, two copies of the dilution mutation result in a homogenous, off-white coat colour (also referred to as 'dun'), and so we would have expected Holstein-Friesian edited for two copies of this mutation to be a similar colour (i.e., nearly purely white). Instead, these calves were grey with white spots (30). Furthermore, although the gene-edited calves were only edited for the *PMEL* mutation, they had a vastly different proportion of white spotting on their coats compared to their cloned, non-edited counterparts. Genotyping of the white spotted tag variants from Chapter Three confirmed that all calves had the same white spotting genotypes (30). These results suggest complex epistatic interactions between the white spotting loci and the *PMEL* locus that alter the proportion of white spotting when co-inherited.

The results presented here demonstrate how mutations in coding and regulatory regions in a small number of genes can interact to generate substantial phenotypic diversity. The interactions between the *MITF* and *KIT* loci to create a splotchy faced animal, and interactions between the *PMEL, KIT, MITF,* and *PAX3* gene loci to create cattle with grey and white coat colouring and increased proportion of white on the coat, were only discoverable by investigating coat colour and patterning traits outside their typical breed context. Given the overall lack of real-world examples of epistasis in any species (at least at a gene × gene level), further exploration of these interactions would be of great academic value. It is therefore interesting to contemplate crossbreeding experiments that could be utilised to investigate interactions between coat colour and patterning loci, that should enhance our understanding of the molecular mechanisms that influence these traits, and modes of epistatic behaviour more broadly. For example, would crosses of

red-and-white Holsteins bearing the recessive *MC1R* mutation be expected to show similar patterning modifications when introgressed with the *PMEL* variant? Or would inheritance of the *MITF* rs209784468 G allele cause increased pigmentation in crosses of cattle that are typically white, such as Charolais or White Galloways?

7.5 Implications and selection utility of coat colour and patterning

Pasture-based farming systems are favoured for both beef and dairy production in New Zealand. With rising global temperatures, and harsh ultra-violet (UV) radiation, pasture-based farming systems may make cattle vulnerable to welfare issues such as heat stress, sunburn, and skin cancer. A major predisposing factor to these events is coat colour. Dairy cattle in major farming regions across New Zealand are thought to experience heat stress for slightly less than 20% of total lactation days (31), and the black hairs on the coat are estimated to absorb twice as much solar radiation in comparison to white hairs, contributing towards heat stress in darker coloured cattle (32). Our initial interest in the proportion of white spotting on the coat was in part driven by studies that suggested that increased white on the coat was negatively correlated with heat stress in cattle (33,34). We hypothesised that the discovery of genetic variants that contribute towards white spotting might facilitate breeding cattle with a higher proportion of white on their coat, and reduce the incidence of heat stress. A shortcoming of this strategy, however, is that it does not account for an increased incidence of sunburn associated with white coat colouration. Laible et al. (30) suggested dilution of the coat colour would be more beneficial. In theory, dilution of the black coat could reduce heat stress by decreasing the amount of UV radiation absorbed, without increasing the risk of sunburn. The mutation responsible for coat colour dilution (p.Leu18del) segregates in Highland and Galloway cattle - breeds that have been selected for beef production.

Rather than using traditional crossbreeding to introgress the coat colour dilution trait into the dairy herd, Laible et al. (30) directly introduced the mutation into a Holstein-Friesian background via gene-editing with CRISPR-Cas9. This method mitigated the risk of genetic drag, where unfavourable beef-associated genetics would not be inherited along with the favourable coat colour dilution mutation. Unexpectedly, the gene-edited calves had more white spotting on their coat, compared to their unedited counterparts, as discussed above in "Epistasis between coat colour loci". Unbiased whole-genome sequence analysis of precursor cell-lines and resultant calves cloned from those edited and non-edited cell lines revealed that regions with sequence similarities to the on-target site (i.e., regions most likely to have off-target mutation events), did not differ across samples (Chapter Six). Furthermore, the number of *de novo* mutations observed in the calves did not significantly differ based on their edited status. However, all of the cloned calves shared a distinct mutation signature that deviated from the expected excess of C>T mutations observed in cattle born from natural mating, AI, IVF, or MOET (35). The *de novo* mutations observed in the cloned calves were instead predominantly T>G mutations, and the *de novo* mutations observed in cells in culture were predominantly C>A mutations. Although these observations were limited by sample size, they highlight the requirement for studies that investigate the genetic consequences of applying reproductive technologies beyond measuring differences in mutation load. Together these studies demonstrated the utility of gene-editing to introgress favourable genetic variation in a single generation, without losses in genetic gain that would normally be associated with crossbreeding.

Public demand for higher animal welfare standards is rising across the globe, and in particular, there have been rising concerns about bobby calves (36). Bobby calves are

surplus calves that are not required by farmers to replace cows in their current herd and are exported for slaughter when they are less than 30 days old. Coat colour and patterning traits can be utilised to reduce the number of bobby calves produced annually by serving as a marker of beef parentage. Cows in the herd that will not be bred for replacement cows can be mated to a beef bull so that the progeny will grow and perform like a beef animal, thus giving an advantage over 'pure' dairy calves. For this reason, calves that are born from dairy dams and beef sires receive a higher premium when sold if they have visible markers that can prove beef parentage (36). Hereford bulls are often used for this purpose due to their characteristically white face, however some dairy × Hereford calves have splotchy faces (as previously discussed in "Epistasis between coat colour loci"). Although favourable for parentage identification, the white-face trait makes cattle more vulnerable to bovine infectious keratoconjunctivitis and bovine ocular squamous-cell carcinoma, especially when these cattle are raised in pasture-based farming systems where animals spend most of their time outdoors in the sun (28,37,38). Both of these diseases cause discomfort to the animal and sometimes cause their carcases to be rejected at meat processing plants due to being unfit for human consumption, resulting in economic loss to the farmer. Pigmentation around the eyes, or ACOP, caused by two co-locating mutations at the *KIT* locus (discussed previously in "Epistasis between coat colour loci"), has been found to reduce the risk of both bovine infectious keratoconjunctivitis and bovine ocular squamous-cell carcinoma in cattle with white faces (39,40). Knowledge of how face colour is modulated could thus enable breeders to produce dairy × Hereford calves with white faces and ACOP. This combination of coat patterning traits could help enhance animal welfare, minimise the number of calves that need to go on the bobby truck per season, and maximise profit on farm.

7.6 The future of artificial selection

Artificial selection has been a feature of agriculture for millennia, and likely began by focussing on the selection of obvious morphological traits such as coat colour and horned status (i.e., polled). It therefore seems fitting that these traits are amongst the first to be targeted with gene-editing technologies. The advent of genotyping platforms that facilitate the genotyping of thousands of loci simultaneously have also accelerated genetic gain by allowing breeders to assess the competence of bulls at earlier ages, thereby reducing generational intervals for cattle breeding. The utilisation of gene-editing has the potential to provide further step-wise changes in genetic gain by shortening generational intervals through direct introduction of favourable genetic variation in a single generation. However, before these technologies can be utilised for artificial breeding, the causal mutations targeted for introgression need to be known, so continued efforts to discover and characterise large to moderate effect mutations have greater justification now more than ever. Better consumer awareness and acceptance of these approaches will also be required, which means more research will be needed to understand these technologies – beyond the way they modify the genome and their propensity to cause off-target mutations. Furthermore, the epistatic interactions identified in this thesis also pose questions about whether these effects also need careful characterisation within the context of animals edited on different backgrounds. Although more work is required, current research in the field of gene-editing provides a glimpse into the future of artificial selection, and a tangible solution for how the agricultural industries might keep up with the consumer demands for enhanced animal welfare, the rapidly changing climate, and government regulations for more climate conscious farming systems.

7.7 References

1. Whitacre L. Structural variation at the KIT locus is responsible for the piebald phenotype in Hereford and Simmental cattle. University of Missouri; 2014.
2. Bickhart DM, McClure JC, Schnabel RD, Rosen BD, Medrano JF, Smith TPL. Symposium review: Advances in sequencing technology herald a new frontier in cattle genomics and genome-enabled selection. J Dairy Sci. 2020;103(6):5278–90.
3. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9(3):giaa021.
4. Olson TA. Genetics of colour variation. In: The genetics of cattle. 1999. p. 33–53.
5. Jakubosky D, Smith EN, D'Antonio M, Jan Bonder M, Young Greenwald WW, D'Antonio-Chronowska A, et al. Discovery and quality analysis of a comprehensive set of structural variants and short tandem repeats. Nat Commun 2020 111. 2020;11(1):1–15.
6. Eggertsson HP, Jonsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G, et al. Graphtyper enables population-scale genotyping using pangenome graphs. Nat Genet. 2017;49(11):1654–60.
7. Mort RL, Jackson IJ, Patton EE. The melanocyte lineage in development and disease. Development. 2015;142(7):1387–1387.
8. Jackson IJ. Molecular and developmental genetics of mouse coat colour. Annu Rev Genet. 1994;28(1):189–217.
9. Ke H, Kazi JU, Zhao H, Sun J. Germline mutations of KIT in gastrointestinal stromal tumor (GIST) and mastocytosis. Cell Biosci 2016 61. 2016;6(1):1–10.

10. Durkin K, Coppieters W, Drögüller C, Ahariz N, Cambisano N, Druet T, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature. 2012;482(7383):81–4.

11. Brenig B, Beck J, Floren C, Bornemann-Kolatzki K, Wiedemann I, Hennecke S, et al. Molecular genetics of coat colour variations in White Galloway and White Park cattle. Anim Genet. 2013;44(4):450–3.

12. Rubin C-J, Megens H-J, Barrio AM, Maqbool K, Sayyab S, Schwochow D, et al. Strong signatures of selection in the domestic pig genome. Proc Natl Acad Sci. 2012;109(48):19529–36.

13. Pielberg G, Olsson C, Syvänen A, Andersson L. Unexpectedly high allelic diversity at the KIT locus causing dominant white color in the domestic pig. Genet Soc Am. 2002;160:305–11.

14. Henkel J, Saif R, Jagannathan V, Schmocker C, Zeindler F, Bangerter E, et al. Selection signatures in goats reveal copy number variants underlying breed-defining coat color phenotypes. PLOS Genet. 2019;15(12):e1008536.

15. Brooks SA, Lear TL, Adelson DL, Bailey E. A chromosome inversion near the KIT gene and the Tobiano spotting pattern in horses. Cytogenet Genome Res. 2007;(119):225–30.

16. Dürig N, Jude R, Holl H, Brooks SA, Lafayette C, Jagannathan V, et al. Whole genome sequencing reveals a novel deletion variant in the KIT gene in horses with white spotted coat colour phenotypes. Anim Genet. 2017;48(4):483–5.

17. Andersson L. Mutations in domestic animals disrupting or creating pigmentation patterns. Front Ecol Evol. 2020;13(8):116.

18. Mort RL, Ross RJH, Hainey KJ, Harrison OJ, Keighren MA, Landini G, et al. Reconciling diverse mammalian pigmentation patterns with a fundamental

mathematical model. Nat Commun. 2016;7(1):1–13.

19. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. Pathog;1(1):1–17.

20. Lupski JR. Hotspots of homologous recombination in the human genome: not all homologous sequences are equal. Genome Biol. 2004;5(10):1–4.

21. Bickhart DM, Liu GE. The challenges and importance of structural variation detection in livestock. Front Genet. 2014;5(37):1–14.

22. Chen L, Pryce JE, Hayes BJ, Daetwyler HD. Investigating the effect of imputed structural variants from whole-genome sequence on genome-wide association and genomic prediction in dairy cattle. Animals. 2021;11(2):541.

23. Lupski JR, Stankiewicz P. Genomic Disorders: Molecular Mechanisms for Rearrangements and Conveyed Phenotypes. PLOS Genet. 2005;1(6):e49.

24. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008;9(11):867.

25. Feldmann MJ, Piepho HP, Bridges WC, Knapp SJ. Average semivariance yields accurate estimates of the fraction of marker-associated genetic variance and heritability in complex trait analyses. PLOS Genet. 2021;17(8):e1009762.

26. Hou L, Pavan WJ. Transcriptional and signaling regulation in neural crest stem cell-derived melanocyte development: do all roads lead to Mitf? Cell Res. 2008;18:1163–76.

27. Bondurand N, Pingault V, Goerich DE, Lemort N, Sock E, Caignec C Le, et al. Interaction among SOX10, PAX3 and MITF, three genes altered in Waardenburg syndrome. Hum Mol Genet. 2000;9(13):1907–17.

28. D'Mello S, Finlay G, Baguley B, Askarian-Amiri M. Signaling pathways in melanogenesis. Int J Mol Sci. 2016;17(7):1144.

29. Schmutz SM, Dreger DL. Interaction of MC1R and PMEL alleles on solid coat colors in Highland cattle. Anim Genet. 2013;44(1):9–13.

30. Laible G, Cole S-A, Brophy B, Wei, Leath S, Jivanji S, et al. Holstein Friesian dairy cattle edited for diluted coat color as a potential adaptation to climate change. BMC Genomics. 2021;22(856):1-12.

31. Bryant JR, López-Villalobos N, Pryce JE, Holmes CW, Johnson DL. Quantifying the effect of thermal environment on production traits in three breeds of dairy cattle in New Zealand. New Zeal J Agric Res. 2007;50(3):327–38.

32. Stewart RE. Absorption of solar radiation by the hair of cattle. Agric Eng. 1953;34:235–8.

33. Hansen PJ. Effects of coat colour on physiological responses to solar radiation in Holsteins. Vet Rec. 1990;127(13):333–4.

34. Bercerril CM, Wilcox CJ. Determination of percentage of white coat color from registry certificates in Holsteins. J Dairy Sci. 1992;75(12):3582–6.

35. Harland C, Durkin K, Artesi M, Karim L, Cambisano N, Deckers M, et al. Rate of de novo mutation in dairy cattle and potential impact of reproductive technologies. Proc World Congr Genet Appl to Livest Prod. 2018;(January 2020).

36. Coleman LW. The use of high genetic merit Angus and Hereford bulls in a New Zealand dairy herd. Massey University; 2020.

37. Heeney JL, Valli VEO. Bovine ocular squamous cell carcinoma: An epidemiological perspective. Can J Comp Med. 1985;49(1):21–6.

38. Seid A. Review on infectious bovine keratoconjunctivitis and its economic impacts in cattle. J Dairy Vet Sci. 2019;9(5):555774.

39. Ward JK, Nielson MK. Pinkeye (bovine infectious keratoconjunctivitis) in beef

cattle. J Anim Sci. 1976;49(2):361–6.

40. Davis KM. Digital analysis of eye pigmentation of Hereford, Hereford x Bos indicus or Hereford x Bos taurus cattle. Texas A&M University; 2013.