Table of Contents

Chapter 4: Detailed investigation of a milk fat percentage QTL underpinned by the *MGST1* gene ... 99

Chapter 5: *DGAT1* K232A; A familiar milk fat mutation with a new mechanism ... 125

List of Tables

List of Figures

List of Abbreviations

A	adenine
A	alanine
ABCG2	ATP-binding cassette sub-family G member 2
AGPAT6	glycerol-3-phosphate acyltransferase
AGRF	Australian Genome Research Facility
AMP	ampicillin
AR^2	allelic r-squared
ATAC	assay for transposase-accessible chromatin
ATP	adenosine triphosphate
bp	base pair
BLAST	basic local alignment search tool
BW	breeding worth index
C	cytosine
Cas9	CRISPR associated protein 9
cDNA	complimentary deoxyribonucleic acid
CGPM	Centre for Genomics Proteomics and Metabolomics
ChIPseq	chromatin immunoprecipitation sequencing
chr	chromosome
cM	centimorgan
CMV	*Cytomegalovirus*
CNV	copy number variant
CRISPR	clustered regularly interspaces palindromic repeats
crRNA	CRISPR RNA
CSN3	kappa-casein
DBS	double-stranded break
DDT	dichlorodiphenyltrichloroethane
DGAT1	diacylglycerol O-acyltransferase 1
DMEM	Dulbecco's modified Eagle medium
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
DNaseseq	DNase-sequencing
dNTP	deoxynucleoside 5′-triphoshate
DSB	double-stranded break
EBV	estimated breeding value
EIF3K	eukaryotic translation initiation factor 3, subunit K
ENCODE	The Encyclopaedia of DNA Elements
eQTL	expression quantitative trait loci
ESE	exonic splice enhancer

ESS	exonic splice silencer
F	Friesian
F	phenylalanine
FACS	fluorescence-activated cell sorting
FASN	fatty acid synthase
FJXB	Friesian-Jersey cross-breed
FPKM	fragments per kilobase of exon model per million mapped
G	guanine
gDNA	genomic DNA
GEBV	genome enhanced breeding value
GFP	green fluorescent protein
GHR	growth hormone receptor
gRNA	guide RNA
GWAS	genome-wide association study
GS	genomic selection
GTF	gene transfer format
HDR	homology-directed repair
HF	Holstein-Friesian
IBD	identical by descent
IGV	integrated genomics viewer
indel	insertion/deletion
ISE	intronic splice enhancer
ISS	intronic splice silencer
J	Jersey
K	lysine
kb	kilobase
kDa	kilodalton
kg	kilogram
LB	Luria-Bertani
LD	linkage disequilibrium
LIC	Livestock Improvement Cooperation
MA	mixed ancestry
MAC-T	bovine mammary alveolar cell line
MAF	minor allele frequency
MAS	marker-assisted selection
Mbp	megabase
MFG	milk fat globule
MGST1	microsomal glutathione S-transferase 1
MNP	multiple nucleotide polymorphism
mRNA	messenger RNA
N	asparagine

NCBI	National Centre for Biotechnology Information
NGS	next generation sequencing
NHEJ	non-homologous end joining
NMD	nonsense-mediated decay
NTC	no template control
NZ	New Zealand
NZGL	New Zealand Genomics Limited
PAM	protospacer adjacent motif
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PLAG1	pleomorphic adenoma gene 1
pre-mRNA	precursor mRNA
PRL	prolactin
PRLR	prolactin receptor
qPCR	quantitative real time polymerase chain reaction
QTL	quantitative trait loci
QTN	quantitative trait nucleotide
RNA	ribonucleic acid
RNAseq	RNA sequencing
RNP	ribonucleoprotein
rRNA	ribosomal RNA
rpm	revolutions per minute
RPS15A	ribosome protein S15a
RT	reverse transcriptase
RT-PCR	reverse transcription polymerase chain reaction
S	serine
SD	standard deviation
SDS	sodium dodecyl sulphate
sgRNA	single guide RNA
SNP	single nucleotide polymorphism
snRNP	small nuclear ribonucleo protein
snRNA	small non-coding RNA
SP1	specificity protein 1
SR	serine/arginine-rich
ssODN	single strand oligonucleotide
STAT	signal transducer and activation of transcription
T	thymine
TALEN	transcriptional activator-like effector nucleases
TBE	Tris buffered ethylenediaminetetraacetic acid
tracrRNA	trans-activating CRISPR RNA
TSS	transcription start site

T7EI	T7 endonuclease I
U	units
UoA	University of Auckland
USD	United States dollar
UTR	untranslated region
VNTR	variable number tandem repeat
VST	variance stabilising transformation
WGBS	whole genome bisulfite sequencing
WGS	whole genome sequence
WT	wild-type
Y	tyrosine
ZFN	zinc finger nucleases

List of papers arising from work presented in this thesis

Littlejohn M. D., Tiplady K, Lopdell T, **Law T,** Scott A, Harland C, et al. (2014) Expression variants of the lipogenic *AGPAT6* gene affect diverse milk composition phenotypes in *Bos taurus*. *PLoS One*, 9(1), e85757.http://doi:10.1371/journal.pone.0085757

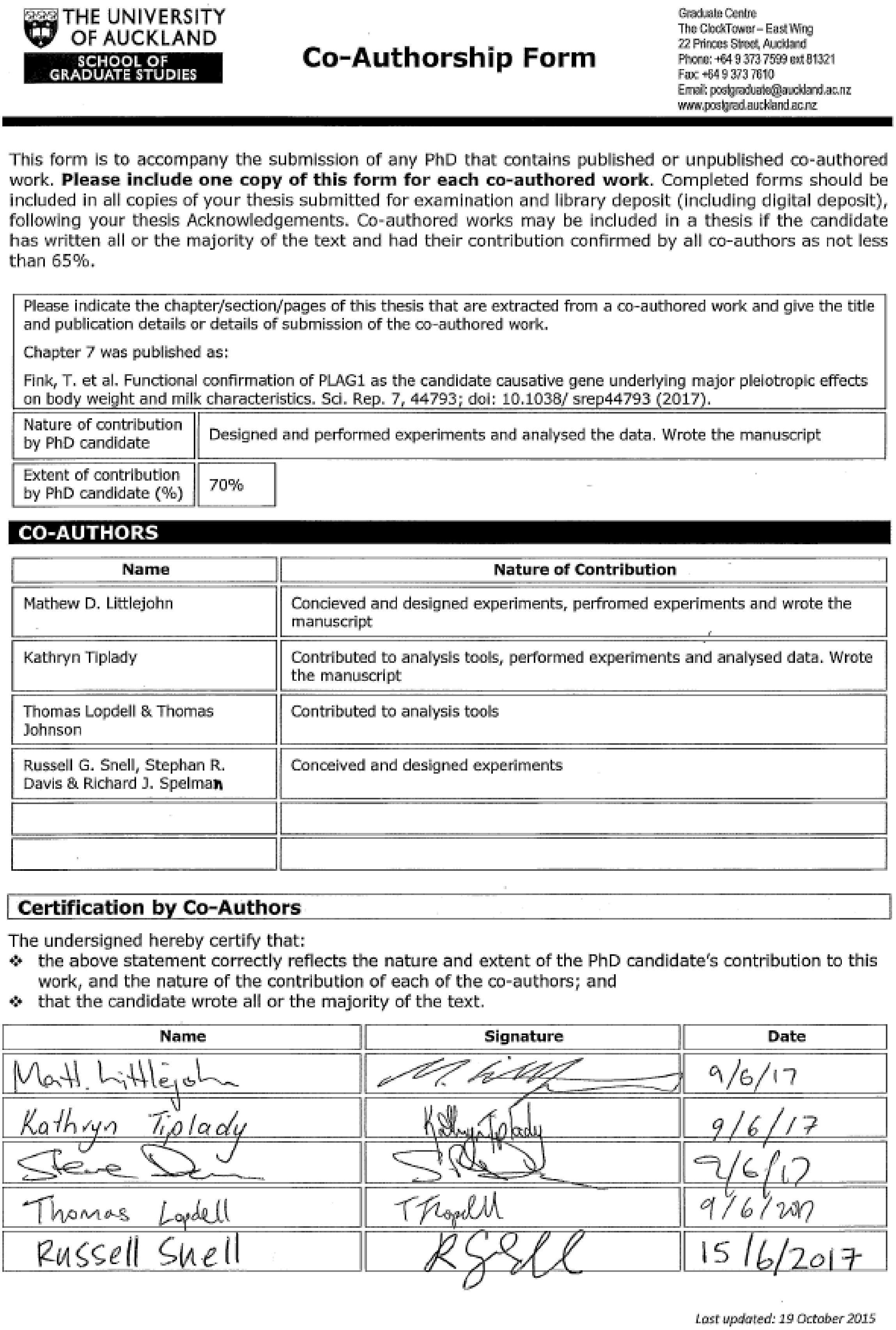
Littlejohn, M. D., Tiplady, K., **Fink, T. A**., Lehnert, K., Lopdell, T., Johnson, T., … Spelman, R. J. (2016). Sequence-based association analysis reveals an *MGST1* eQTL with pleiotropic effects on bovine milk composition. *Scientific Reports*, *6*(25376), 1–14. http:// doi:10.1038/srep44793

Fink, T. A., Tiplady, K., Lopdell, T., Johnson, T., Snell, R. G., Spelman, R. J., … Littlejohn, M. D. (2017). Functional confirmation of *PLAG1* as the candidate causative gene underlying major pleiotropic effects on body weight and milk characteristics. *Scientific Reports*. 7(44793), 1-8. http://doi: 10.1038/ srep44793

Table of Contributions

This thesis describes the use of existing datasets described in detail in General Methods (section 2.1.3 and summarised in Table 2.1). The below table outlines contributions made to the work presented in the experimental chapters of this thesis.

Contributor	**Nature of Contributions**
Kathryn Sanders	Contributions were made to the analysis tools and execution of existing scripts for genetic association analysis (included in Appendix VI). I developed a number of scripts that leveraged and modified these existing scripts to carry out the association analyses described in this thesis.
Thomas Lopdell	Contributions were made to the design of experiments and analysis tools used in Chapter 5. In conjunction with Thomas I developed a number of bioinformatic tools and scripts used to analyse the RNAseq dataset.



THE UNIVERSITY OF AUCKLAND
SCHOOL OF GRADUATE STUDIES

Co-Authorship Form

Graduate Centre
The ClockTower – East Wing
22 Princes Street, Auckland
Phone: +64 9 373 7599 ext 81321
Fax: +64 9 373 7610
Email: postgraduate@auckland.ac.nz
www.postgrad.auckland.ac.nz

This form is to accompany the submission of any PhD that contains published or unpublished co-authored work. **Please include one copy of this form for each co-authored work**. Completed forms should be included in all copies of your thesis submitted for examination and library deposit (including digital deposit), following your thesis Acknowledgements. Co-authored works may be included in a thesis if the candidate has written all or the majority of the text and had their contribution confirmed by all co-authors as not less than 65%.

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 7 was published as:

Fink, T. et al. Functional confirmation of PLAG1 as the candidate causative gene underlying major pleiotropic effects on body weight and milk characteristics. Sci. Rep. 7, 44793; doi: 10.1038/ srep44793 (2017).

Nature of contribution by PhD candidate	Designed and performed experiments and analysed the data. Wrote the manuscript
Extent of contribution by PhD candidate (%)	70%

CO-AUTHORS

Name	Nature of Contribution
Mathew D. Littlejohn	Concieved and designed experiments, perfromed experiments and wrote the manuscript
Kathryn Tiplady	Contributed to analysis tools, performed experiments and analysed data. Wrote the manuscript
Thomas Lopdell & Thomas Johnson	Contributed to analysis tools
Russell G. Snell, Stephan R. Davis & Richard J. Spelman	Conceived and designed experiments

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and
- that the candidate wrote all or the majority of the text.

Name	Signature	Date
Matt. Littlejohn		9/6/17
Kathryn Tiplady	Kathryn Tiplady	9/6/17
Steve [illegible]		9/6/17
Thomas Lopdell	T Lopdell	9/6/2017
Russell Snell		15/6/2017

Last updated: 19 October 2015

Thomas JJ Johnson		14/7/17
Richard Spelman		14/6/17

Chapter 1: Introduction

1.1. Overview

Over the last two decades there has been a remarkable increase in genetic mapping studies, where genotyping and statistical association of DNA sequence variants has been used to identify loci underpinning variation for a variety of diseases and complex traits. In particular, technological advancements in genotyping and sequencing technology have led to the routine implementation of large-scale genome-wide association studies (GWAS) for many different species, including humans, model organisms, and agricultural species such as *Bos taurus.* These GWAS leverage a genome-wide coverage of genetic markers to identify chromosomal intervals harbouring variants that impact quantitative traits (called quantitative trait loci, QTL). The aim of many of these studies is to act as a discovery platform that can then be used to identify the causative variants responsible for QTL of interest, however, in many cases, identification of these variants remains a major challenge. In the case of bovine milk composition and production traits, of key interest to dairy industries in New Zealand (NZ) and abroad, relatively few causative mutations have been identified, despite a wealth of association data.

It is difficult to determine which genetic variant is causative for a given genetic signal as the associated loci often encompass several megabase pairs (Mbp) of DNA sequence, and the most highly associated variants may reside in non-coding regions of the genome, which are difficult to interpret. Identifying the causative variant for the genetic signal at a given locus requires understanding its mechanism of action and target gene(s). While this can be relatively straightforward for variants predicted to impact the protein-coding sequence of a gene, it is much more difficult to attribute a functional consequence to non-coding variants, given the diverse functions of non-coding DNA. Further, the incomplete annotation of regulatory elements, particularly relevant to non-model species for which scant functional information exists, makes prediction of non-coding variant effects especially challenging. One approach to address these challenges is to build the genomic and functional annotation resources required for interpretation, for example conducting gene expression analyses to identify which genes may be actively expressed/regulated at a QTL of interest. Another, albeit more difficult and time consuming approach, is to test the function of variants directly, where engineering *in vitro* or model organism systems can be used to assess the consequences of genetically isolated candidate variants.

This thesis describes the detailed investigation of four major milk production loci in dairy cattle, where the mechanisms of these QTL incorporate gene expression-based effects. The work detailed focuses on exploration of these mechanisms, and employs applications of the annotation and functional testing approaches described above to identify the causative gene and/or variants responsible. In this first chapter, I will review work relevant to these discoveries, beginning by discussing concepts of quantitative genetics and causal variant identification, within a broader context of bovine milk production traits and lactation biology. The chapter will conclude with a discussion of genome editing technologies, and a summary of the scope and outline of this thesis.

1.2. Mammary gland and lactation biology

The mammary gland is the defining feature of the class Mammalia, and possesses the unique ability to synthesise and secrete substantial amounts of milk to support the growth of the developing neonate (Akers, 2002). Although there is some contention around its evolutionary origins, the mammary gland is proposed to have evolved over 300 million years ago and regardless of the differing location and external form among species, shows parallel phases of development and mechanisms of milk production (M C Neville, 2001; Oftedal, 2002).

The mammary gland is a highly organised and complex organ, located on the ventral surface of mammals. The gland is characterised by two distinct tissue types: the highly specialised mammary epithelium, which consist of ducts and milk-producing alveolar cells, and the stroma, or connective tissue, which is also called the mammary fat pad (Hennighausen & Robinson, 2005; M C Neville, 1990). Key secretory functions are fulfilled by the epithelial cells, the cytoplasms of which are populated with numerous mitochondria and rough endoplasmic reticulum, a common feature of highly active secretory cells across other tissues (McManaman & Neville, 2003).

The secretory epithelial cells come together in a single layer to form ducts and alveoli sharing a central lumen into which milk is secreted. Each of these alveoli are surrounded by contractile myoepithelial cells, as well as supporting tissue and ligaments which form a mesh-like system to aid the excretion of milk into the lumen (McManaman & Neville, 2003). Several alveoli form grape-like structures called lobules, which are connected through a complex network of ducts that opens to the body surface through the teat (Margaret C Neville, McFadden, & Forsyth, 2002).

The secretory epithelial cells of the mammary gland synthesise and secrete the individual milk constituents in a highly organised and synchronised manner in order to meet the specific nutritional requirements of the developing young. Both local and global mechanisms control the mammary epithelial cells, the most important of which include the lactogenic hormones, local growth factors and cell-to-cell interactions (McManaman & Neville, 2003; Margaret C Neville et al., 2013).

Milk is an opaque white fluid highly regarded as one of the best nutritional sources of protein, lipid, lactose and calcium. In addition to these macronutrients, the mammary gland also contributes many micronutrients and other bioactive factors including vitamins, minerals, oligosaccharides, immunoglobulins, cytokines, antibodies, enzymes, enzyme inhibitors, growth factors, hormones and antibacterial agents to milk. The components all function to aid the survival and growth of the neonate (Wickramasinghe, Rincon, Islas-trejo, & Medrano, 2012). Milk composition varies throughout lactation, with the most significant changes occurring around the first 36 hours after parturition. During this time, the mammary epithelial cells are transitioning from a quiescent state to a fully active state, and are insufficiently developed to synthesise all of the individual milk constituents (Rudolph et al., 2007). As a consequence, the first milk produced upon parturition is referred to as colostrum and has high concentrations of immunoglobulins and lactoferrin important to the early phase of neonatal life. Once fully active, the mammary epithelial cells are able to produce casein and lactose, which characterises the onset of copious milk secretion (Anderson, Rudolph, McManaman, & Neville, 2007).

The composition of milk varies both between and within mammalian species, which in *Bos taurus* is the result of a number of factors, including genetics, nutritional state, food composition, season, breed, milking frequency, and environmental stresses (Holmes et al., 2002). Overall, the inter-species variation in gross milk composition is due to the differential reproductive strategies and nutritional requirements of the neonate, namely its maturity at birth, growth rate and energy requirements (Oftedal, 2009). For example, the fat content of seal milk can be as high as 60%, whereas it is negligible in rhinoceros (0.2%) (Brennan et al., 2007; Oftedal & Iverson, 1995). Several studies have investigated intra-species variability in milk composition, particularly in different breeds of *Bos taurus*, where milk fat percentage varies between 5.5% in Jersey and 3.5% in Holsteins (Jensen, 2002). Importantly, this variation in milk composition has been shown to be partly attributable to genetics, with milk protein and lipid composition demonstrating high levels of heritability (Bovenhuis, Visker, & Lundén, 2013).

1.3. Genetics

Genetics involves the study of genes, gene function, genetic variation, and heritability, where genetic variation is used to describe the DNA sequence variation in the genome of an individual or population. Heritability is the proportion of a trait that can be explained by genes, and was first dissected by Gregor Mendel using pea plant breeding. Mendel conducted extensive experiments on these plants to develop three principles to explain the patterns of inheritance he observed for different characteristics in pea plants, including pod shape, seed shape and pea colour (Mendel, 1865).

As DNA was not identified as the carrier of genetic material until 1944, Mendel's experiments on heritability were conducted prior to the discovery of the molecular basis of inheritance (Avery, MacLeod, & McCarthy, 1944). The discovery of DNA, and the advent of molecular techniques, in particular Sanger sequencing, enabled the characterisation of both the structure and sequence of an individual's DNA (Sanger, Nicklen, & Coulson, 1977; Watson & Crick, 1953). Subsequently, we know that the sum of hereditary material transmitted from parent to offspring is an individual's genome, which is DNA arranged into chromosomes. Mammalian genomes contain a few billion of bases of DNA and individuals within a population (and species) differ from each other at many positions, from tens of thousands in yeast to millions in human populations (Steinmetz et al., 2002; Visscher, McEvoy, & Yang, 2010).

1.4. Quantitative genetics

Quantitative genetics involves the study of traits that show a continuous phenotypic distribution within a segregating species (or population). This continuous variation is the result of the complex interplay between many genetic and environmental factors, and often results in these traits exhibiting a moderate-to-low heritability (Doerge, 2002). The majority of mammalian phenotypes, including bovine milk production and composition, are quantitative traits, exhibiting a complex genotype-to-phenotype relationship (Georges et al., 1995). This makes it difficult to understand the genetic architecture of these traits, and as a result, modest numbers of the DNA sequence variants that cause variation in quantitative traits have been identified.

In attempts to explain the heritability of the phenotypic variation present in complex traits, early quantitative geneticists developed statistical methods to partition the phenotypic variance into genetic and non-genetic variance (Fisher, 1919), and describe the relationships between related individuals (S. Wright, 1921). Fisher (1919) illustrated that the variation of quantitative traits is consistent with Mendelian inheritance, such that even though each gene is inherited according to Mendel's Laws, the trait approximates a statistically normal distribution if there are three or more genes influencing the trait. Consequently, it is assumed that the genetic architecture of quantitative traits follows Fisher's infinitesimal model, whereby the phenotype is determined by an infinite number of genes, each with a small effect (Dekkers & Hospital, 2002).

1.4.1. Quantitative trait loci mapping

Quantitative trait loci (QTL) mapping is the process of associating DNA sequence variants within a particular genomic interval with the variation in a quantitative trait. QTLs can be mapped to genomic intervals through their linkage with polymorphic markers, such as single nucleotide polymorphisms (SNPs), small insertions or deletions (indels), or copy number variants (CNVs), which demonstrate Mendelian segregation and have known positions in the genome (Mackay, 2001). A QTL is identified if individuals of different marker genotypes demonstrate different mean values of the quantitative trait (Lander & Botstein, 1989).

Based on Fisher's model and since the landmark paper of Lander & Botstein in 1989, considerable effort has been invested in associating DNA sequence variants with quantitative trait variation. There are two major approaches to mapping QTL, exploiting either linkage or linkage disequilibrium (LD) between markers. Traditionally, QTL mapping was pursued using linkage-based analyses largely due to the absence of dense marker panels; however, this has now been superseded by LD-based mapping given the advent of large-scale genotyping and sequencing resources. LD mapping is discussed in more detail in the following section (1.4.2), while linkage-based QTL mapping and how this has been applied to dairy cattle is described below.

In early linkage-based studies, QTLs were identified based on the inheritance pattern of phenotypes and genotypes observed in pedigrees and experimental crosses (Lander & Schork, 1994). Tracking chromosomes and mapping the phase between markers from one generation to the next enabled these studies to identify QTL that segregate with phenotype more often than expected by chance. The strength of linkage-based studies to detect QTL was maximised by the use of crosses between phenotypically divergent animals, such as Jersey and Holstein-Friesians, as the crossbreed progeny often segregate for the alleles underpinning their phenotypic differences. As recombination events are rare per meiosis, tagging a QTL requires only a few genetic markers per chromosome. However, the resolution of a particular QTL is also limited by the lack of recombination, resulting in large confidence intervals (typically 20 centimorgans (cM) or more) with QTL identified through linkage-based analyses (Goddard & Hayes, 2009; Ron & Weller, 2007).

Early QTL studies in dairy cattle conducted linkage-based analyses within pedigrees with the objective of identifying genes and markers that could be included in breeding programs via marker-assisted selection (MAS; Khatkar, Thomson, Tammen, & Raadsma, 2004). The DNA markers identified in these studies, most commonly microsatellites and variable number tandem repeats (VNTR), were used to predict animal performance and improve milk production (Beuzen, Stear, & Chang, 2000).

The first QTL affecting milk production was detected in dairy cattle by exploiting progeny testing (Georges et al., 1995). In this study, 1,518 progeny tested sires (each with between 50 to several thousand daughters) were genotyped for 159 microsatellite markers covering two-thirds of the bovine genome. Their genotypes for these markers were used in conjunction with the lactation performances of their daughters to reveal five QTLs across the genome. Specifically, a QTL on chromosome 9 was associated with milk, fat and protein yield, while two QTL on chromosome 6 and 20 only influenced milk yield. Additionally, a QTL on chromosome 10 was associated with milk and fat yield, while a QTL on chromosome 1 influenced milk and protein yield (Georges et al., 1995).

1.4.2. Genome-wide association studies

Due to the decreasing costs of high-throughput genotyping and sequencing, genome-wide association studies (GWAS) are now routinely implemented, where for a population, phenotypes are recorded and these individuals are assayed for a genome-wide SNP panel to detect statistical associations between the trait and the SNP markers (Goddard & Hayes, 2009). GWAS differ from the more traditional linkage-based association studies as they do not rely on observing familial inheritance patterns, but rather the LD between markers, where the correlation between alleles on ancestral chromosomes has been eroded by recombination over time (Lander & Schork, 1994). By exploiting the LD between markers on a high density SNP panel, GWAS have a higher power to detect common QTL and a more precise estimate of QTL locations (Mackay, 2001).

Similar to linkage-based analyses, for GWAS, individuals within a population are grouped by their genotype class for each marker, and the phenotype means of the different classes are compared for each of the markers (Georges, 2007). This process is repeated for all markers across the genome, and usually only a handful of polymorphisms will associate with differences in phenotypic means within the population (Mackay, Stone, & Ayroles, 2009). Many large-scale GWAS have been conducted in humans and other organisms to identify hundreds of associations with complex traits (Flint & Mackay, 2009; Mackay et al., 2009; Stranger, Stahl, & Raj, 2011). In these studies, the effect of each marker on the quantitative trait is tested based on a simple linear model that includes the effect of the SNP, a fixed effect such as the cohort, and a polygenic breeding value of each animal, which is a measure of the influence of all other genes affecting the trait (described in detail in Goddard & Hayes 2009).

As the underlying assumption of these studies is that any association is the result of LD between the marker and the underlying causative variant(s) for the QTL, GWAS are not without their limitations. Markers included in genome-wide SNP panels are selected because both their alleles are common and they are amenable to high-throughput genotyping, therefore it is unlikely they are in complete LD with causal variants, which may have small minor allele frequencies (MAF) (Wray et al., 2013). As a result, GWAS do not tend to capture all the genetic effects and given the effect sizes of individual loci are small, sample numbers must be sufficiently large to detect association signals. This is not always achievable given the cost in conducting these studies (Raychaudhuri, 2011).

The genomic architecture and selection history of cattle also carries distinct advantages for the discovery of QTL in GWAS. During the domestication of cattle, the effective population size for some breeds was estimated to fall to just 100, which has resulted in long range LD that extends to similar degrees as that seen in domesticated dogs (Boyko et al., 2010; Farnir et al., 2000). This means that large chromosomal segments are identical by descent (IBD) which results in fewer markers being required to identify loci associated with milk production traits (Farnir et al., 2000; Goddard & Hayes, 2009).

1.5. Genetics of dairy cows and milk production

Since the advent of agriculture, dairy cows have been undergoing selection. Initially, as part of the domestication process, animals were selected based on their docility and amenability to farming. However, over the last 65 years, there has been a strong focus on breeding animals for high milk production, which in the US resulted in the doubling of the average milk yield of an individual cow in just 40 years (Georges, 2007). The strong selective breeding of dairy cattle has led to marked phenotypic diversity and genetic adaptation to the various environmental and farming conditions, demonstrated by the large changes in milk production across breeds and populations (Andersson & Georges, 2004). Selection has predominately been conducted by breeding animals with favourable milk production characteristics in the hope their offspring will also exhibit the phenotype (and culling inferior individuals; Holmes et al., 2002). However, more recently, the selection of animals has been based on their estimated breeding values (EBVs), calculated from phenotypic records and pedigree information, and the knowledge of the heritability of each trait (described later in section 1.5.1; Goddard & Hayes, 2009).

Milk production phenotypes are measured through herd testing, which provides data on milk volume, protein, fat, and lactose percentages and yields, and somatic cell count. The latter is an indicator for mastitis, where the numbers of cells in milk are assumed as leucocytes and taken as a proxy of response to infection. Additional measuring of the animal's weight and recording of calving and mating events also provides essential information about the growth, efficiency and fertility of individuals. All but lactose percentage and yield make up components of an animal's Breeding Worth Index (BW) in NZ, with those sires with a superior BW used prolifically in artificial insemination. These BWs are expressed at $ net farm income per 5,000 kg dry matter feed intake and is compared to a genetic base of cows born in the year 2000 (DairyNZ, 2014b).

1.5.1. Genomic selection in dairy cattle

The results of GWAS can serve to identify genes important to physiological or disease processes, and have intrinsic academic value for that purpose. The identity of trait-associated markers may also be used for animal breeding, and may be incorporated in selection schemes such as genomic selection (GS). GS is a form of marker assisted selection (MAS), and uses a genome-wide panels of dense SNP markers, under the assumption that all QTLs are captured by at least one marker, to predict the genetic value of an individual (B. J. Hayes, Bowman, Chamberlain, & Goddard, 2008). Using this principle of GS, a sample of animals with both genotypic and phenotypic data available (the reference population) is used to generate genome enhanced breeding values (GEBVs) of other animals, such as young bulls, in the absence of phenotypic information (B. J. Hayes et al., 2008; Meuwissen, Hayes, & Goddard, 2001).

The use of a dense genome-wide panel of markers enables the prediction of performance at a higher accuracy than before, with simulations and early experiments demonstrating that EBVs could be predicted with accuracy up to 0.85, where accuracy is the correlation between an animal's true breeding value and EBV (Goddard & Hayes, 2009; Meuwissen et al., 2001). The demonstration that it was possible to make accurate selection decisions when breeding values were predicted from dense genetic markers alone, led to the widespread adoption of GS in dairy cattle breeding programs in many countries (van Marle-Köster, Visser, & Berry, 2013). Indeed, GS was predicted to double the rate of genetic gain through selection and breeding from bulls at 2 years of age (i.e. in the absence of phenotypic information from its progeny; Schaeffer, 2006). However, GS is not able to capture new or rare variants and haplotypes, so while progress has been made in the genetic improvement of dairy cattle, further understanding of the causative genes and DNA sequence variants that influence bovine milk production may help augment GS methods and contribute to greater genetic gain in dairy cattle.

1.5.2. QTL mapping in dairy cattle

Since the initial study conducted by Georges et al., (1995) (refer to section 1.4.1), many QTL have been mapped for bovine milk phenotypes, primarily focused on the milk production traits: milk, protein and fat yield, and milk composition traits: milk fat and protein percentage. Milk production and composition QTLs are present on all bovine autosomes, with the most QTLs reported for chromosome 6, 14, 20 and 27. In a review of 59 studies, 238 milk production QTL were identified for these traits, spanning only 63 unique genome locations (Lemay et al., 2009). In particular, milk, fat and protein yields share common loci in many GWAS (Cole et al., 2011). According to the Cattle QTL database (http://www.animalgenome.org/cgi-bin/QTLdb/BT/summary) there are 1,895 QTL for milk yield, 5,321 QTL for milk fat yield, 3,200 QTL for milk fat percentage, and 3,012 QTL for milk protein percentage, as at December 2016. The observation that a number of these QTL have pleiotropic effects on the different milk production phenotypes highlights the effect of strong selection pressure on one or more of these traits, which has correspondingly also affected the mean(s) of related traits (Visscher & Yang, 2016).

In addition to the observation that many QTL have pleiotropic effects, it is also important to note that many of these QTL have moderate-to-large effects on milk composition (Georges, 2007). This is not in line with Fisher's infinitesimal model (described in section 1.4), which assumes loci have small effects, highlighting the unique genetic architecture of bovine milk production, presumably as a consequence of strong selection pressure in cattle.

1.5.3. Major causative variants for bovine production traits

To gain insight into mammary biology and increase the rate of genetic gain in dairy cattle, numerous studies have been conducted towards identifying the genes and polymorphisms that are responsible for bovine milk production QTLs (Table 1.1). The genes discovered so far represent diverse functions including genes that encode the major milk constituents such as the casein gene cluster on bovine chromosome 6 (Parma, Curik, Greppi, & Enne, 2005), or genes that encode lactogenic hormones and their ligands, such as insulin and growth hormone receptor (Blott et al., 2003; Viitala et al., 2006).

Two of the best characterised genes with an influence on bovine production traits are *DGAT1* and *PLAG1,* which are responsible for a significant amount of the phenotypic variation in milk composition and liveweight, respectively (described below; Grisart, Coppieters, Riquet, et al., 2002; Karim et al., 2011). In general, a quantitative trait nucleotide (QTN) with a large effect on a trait under intense selection, such as milk production or animal growth in dairy cows, will tend towards fixation faster than a QTN with little effect on a trait due to the subsequent selection pressure (Falconer & Mackay, 1996). As such, opposing alleles of the *PLAG1* QTL (that has profound effects on liveweight) have become almost fixed in Holstein-Friesians and Jerseys (Fortes, Reverter, Kelly, Mcculloch, & Lehnert, 2013; Kathryn E Kemper, Visscher, & Goddard, 2012).

DGAT1, *PLAG1*, and several other QTNs with notable effects on production traits are briefly summarised below.

Diacylglycerol O-acyltransferase 1

The variant with the most widely validated influence on bovine milk production is a mutation in *diacylglycerol O-acyltransferase 1 (DGAT1).* Specifically, an AA to GC dinucleotide substitution causes a lysine to alanine amino acid substitution (K232A), which is responsible for a large pleiotropic QTL on the centromeric end of bovine chromosome 14. This QTL has a moderate-to-large effect on a broad range of milk production phenotypes, the most significant of which is milk fat percentage (Grisart, Coppieters, & Farnir, 2002).

DGAT1 encodes an enzyme that catalyses the terminal reaction in the mammary triglyceride synthesis pathway and the K232A mutation has been shown to increase triacylceride synthesis *in vitro* (Grisart et al., 2004). The allele that increases milk fat yield also decreases both protein and milk yield (Grisart, Coppieters, & Farnir, 2002) in both *Bos indicus* and *Bos taurus* breeds (Kaupe, Winter, Fries, & Erhardt, 2004). The magnitude and prevalence of this QTL is reflected in part by the fact that chromosome 14 has the highest number of reported QTLs (3,199 QTL; http://www.animalgenome.org/cgi-bin/QTLdb/BT/summary).

Pleomorphic adenoma gene 1

A major pleiotropic QTL located at approximately 25 Mbp on chromosome 14 affects stature and growth traits in both *Bos taurus* and *Bos indicus* cattle breeds. Specifically, two variants; rs209821678, a (CCG) repeat of 9 or 11 copies and rs210030313 an A to G nucleotide substitution in the bidirectional promoter of *pleomorphic adenoma gene 1 (PLAG1)* and *CHCHD7* have been identified as likely causal for these QTL (Karim et al., 2011; M. Littlejohn et al., 2012; Utsunomiya et al., 2013). These mutations were associated with the foetal expression of seven of nine genes within a fine-mapped, ~ 780 kilobase (kb) interval, with three of these; *PLAG1*, *RPS20* and *SDR16C5* plausible candidates given their roles in growth and oncogenesis (De Vos et al., 2008; Lettre et al., 2008; McGowan et al., 2008). Of these, *PLAG1* is the obvious candidate given *plag1* knockout mice suffer from slow growth and dwarfism (Hensen et al., 2004), with the PLAG1 transcription factor known to regulate several growth factors including IGF2, a key modulator of growth in dogs and humans (Voz, Agten, Van de Ven, & Kas, 2000).

Growth Hormone Receptor

A major QTL located on chromosome 20 affects milk yield and composition in many dairy cattle populations. This QTL has been attributed to *growth hormone receptor* (*GHR*), based on the fine mapping of this locus using microsatellite markers (Blott et al., 2003). *GHR* largely determines the action of growth hormone which plays an important role in the initiation and maintenance of lactation (Hennighausen & Robinson, 2005). Specifically, the QTL has been attributed to a non-conservative phenylalanine to tyrosine amino acid substitution at position 279 (F279Y) in GHR (Blott et al., 2003). The phenylalanine (F) residue is highly conserved among mammals, and is associated with decreased milk yield, increased protein percentage and increased fat percentage (Viitala et al., 2006). However, in many populations the MAF is so low that the effect of GHR is difficult to establish (Jiang et al., 2010; Komisarek, Michalak, & Walendowska, 2011; Viitala et al., 2006).

ATP-binding cassette sub-family G member 2

A QTL affecting fat and protein percentages is located on chromosome 6, and has been fine mapped to a 420 kb region between *ATP-binding cassette sub-family G member 2 (ABCG2)* and *LAP3* (Olsen et al., 2007). Of these genes, *ABCG2* represents a biological candidate as it translocates clinically and toxicologically important substrates into the milk of cows, mice and humans in an ATP dependent process, and is preferentially expressed in the bovine mammary gland at the onset of lactation (Cohen-Zinder et al., 2005; Litman et al., 2000). Cohen-Zinder et al., (2005) sequenced this locus in sires segregating for the QTL to reveal an A to C substitution which causes a tyrosine to serine amino acid substitution at position 581 (Y581S) in *ABCG2* which segregated with the QTL. The tyrosine coding (Y) allele is associated with decreased milk yield and increased fat and protein percentage as reported in this study (Cohen-Zinder et al., 2005).

Prolactin receptor

Prolactin (PRL) is a lactogenic hormone that is essential for the initiation and maintenance of milk production, as well as the stimulation of transcription of the milk protein genes (Holmes et al., 2002). In the bovine genome, *prolactin receptor* (*PRLR*) is located approximately 7 Mbp from *GHR*, and like GHR, PRLR has a major role in the regulation of growth hormone and prolactin in the mammary gland (C Brisken et al., 1999; Viitala et al., 2006). A serine to asparagine amino acid substitution (S18N) in the prolactin receptor is associated with milk, protein and fat yield in Finnish Ayrshire dairy cattle (Viitala et al., 2006). Similarly, this polymorphism is associated with milk yield and fat percentage in a Chinese population of Holstein dairy cattle (Zhang et al., 2008).

Caseins

The main components of milk protein by mass are αs1, αs2, β and κ-casein. Each of these genes have two or more variants (Dove, 2002). The casein genes reside in a ~250 kb cluster on chromosome 6 (Rijnkels, Kooiman, de Boer, & Pieper, 1997), and the locus has been attributed to harbour many milk production QTLs. The allele frequencies of these variants vary considerably between different cattle breeds. This has resulted in conflicting reports with respect to the significance and size of casein genotype effects on milk production traits (Dove, 2002). Overall, the αs1-casein B and β-casein A alleles are associated with increases in milk, fat and protein yields and the κ -casein B allele is associated with increase protein yield and fat percentage (Bovenhuis, Van Arendonk, & Korver, 1992; Deb et al., 2014; Rachagani & Gupta, 2008).

Table 1.1 Summary of known genes that influence *Bos taurus* milk production and composition

Gene		Chr	Affected Phenotype	Reference
LEPR	Leptin receptor	3	Milk yield and fat, liveweight measured at feedlot exit (kg)	(Bolormaa et al., 2011; Strucken, Laurenson, & Brockmann, 2015)
LEP	Leptin	4	Milk yield, energy balance and fertility, fat yield, protein yield, somatic cell score, milk production	(Banos, Woolliams, Woodward, Forbes, & Coffey, 2008; Fontanesi et al., 2014; Szyda, Morek-Kopeć, Komisarek, & Zarnecki, 2011)
IGF1	insulin-like growth factor 1	5	Milk yield and fat	(Strucken et al., 2015)
OLR1	Oxidised low-density lipoprotein receptor 1	5	Milk fat	(Khatib, Leonard, Schutzkus, Luo, & Chang, 2017)
ABCG2	ATP-binding cassette sub-family G member 2	6	Protein percentage, milk yield and composition, milk protein and fat	(Cohen-Zinder et al., 2005; Olsen et al., 2007)
CSN1S1	α_{s1}-casein	6	Milk protein, somatic cell score	(Deb et al., 2014; Fontanesi et al., 2014)
CSN1S2	α_{s2}-casein	6	Milk yield, protein yield and percentage	(Deb et al., 2014; Fontanesi et al., 2014; Molee, Poompramun, & Mernkrathoke, 2015)
CSN2	β-casein	6	Milk yield, protein yield and percentage, somatic cell score	(Raven, Cocks, Goddard, Pryce, & Hayes, 2014; Strucken et al., 2015)
CSN3	κ-casein	6	Milk protein and percentage, milk production	(Ilie, Magdin, Sălăjeanu, Neamţ, & Vintilă, 2009; Raven, Cocks, Goddard, et al., 2014; Strucken et al., 2015)
PAEP	progestogen-associated endometrial protein*	11	Milk protein	(Kuss, Gogol, & Geldermann, 2003)
DGAT1	diglyceride O-acyltransferase 1	14	Milk production and composition	(Bennewitz et al., 2004; Grisart et al., 2004; Grisart, Coppieters, & Farnir, 2002; Kühn et al., 2004)
PLAG1	pleiomorphic adenoma gene 1	14	Stature, milk yield	(Karim et al., 2011; M. Littlejohn et al., 2012)
BCO2	Beta-carotene oxygenase	15	Milk colour	(S. D. Berry et al., 2009)
PIGR	Polymeric immunoglobulin receptor		IgA content of colostrum	(S. Berry et al., 2013)

STAT5A	Signal transducer and activator of transcription 5A	19	Milk composition	(Brym, Kamiński, & Wójcik, 2005; Cobanoglu, Zaitoun, Chang, Shook, & Khatib, 2006; Schennink, Bovenhuis, Léon-Kloosterziel, Van Arendonk, & Visker, 2009)
FASN	Fatty acid synthase	19	Milk yield & composition, milk fat yield, somatic cell score	(Alim et al., 2014; Fontanesi et al., 2014)
GH1	Growth hormone 1	19	Milk yield, milk fat and protein yield, milk composition, somatic cell count, survival, body condition score, body size	(Fontanesi et al., 2014; Machlin, 1973)
SREBF1	Sterol regulatory element binding transcription factor 1		fatty acid composition, fat percentage, protein percentage	(Cochran, Cole, Null, & Hansen, 2013; Nafikov et al., 2013)
PRLR	Prolactin receptor	23	Milk, Protein and fat yield, protein percentage	(Fontanesi et al., 2014; Strucken et al., 2015; Viitala et al., 2006)
GHR	Growth hormone receptor	20	Milk yield and composition, protein and fat percentage	(Blott et al., 2003; Viitala et al., 2006)
LTF	Lactoferrin			
SCD1	Stearoyl-CoA desaturase	26	Milk production traits, fatty acid C10, C12, C14, C16	(Bouwman, Visker, van Arendonk, & Bovenhuis, 2012; Buitenhuis et al., 2014; Moioli et al., 2007)

Chr = chromosome; *formerly β-lactoglobulin

1.6. Causative variant discovery

Causative variant discovery is the identification and characterisation of the specific genetic variant responsible for a given QTL. While linkage analysis and GWAS can identify genetic variants associated with complex traits, in isolation they are unlikely to identify the functional variant (and gene) that influences the phenotype and explains the observed association. This reflects the 'burden of proof' required to claim causality of an individual genetic variant, which often involves genetic, informatics and experimental functional data (Glazier, 2002; MacArthur et al., 2014). As opposed to GWAS which benefit from linkage and LD, the challenge of causative variant discovery is distinguishing the functional genetic variant among the many statistically associated variants.

At many associated loci, LD extends across large genomic intervals, resulting in large confidence intervals encompassing many genes and thousands of variants. For example, of the 238 milk production QTL reviewed by Lemay et al., (2009), 63 had a median interval size of approximately 17 Mbp and were estimated to contain between 105 and 127 genes (Lemay et al., 2009). Within these QTL there will be many highly associated variants, which exhibit similar association statistics due to their correlation and LD. Given the large numbers of associated variants at QTL, the prioritisation and filtering of these variants is important to discern association from causation. Conventionally, prioritisation of candidate causative variants is addressed with statistical fine-mapping techniques. Fine-mapping involves genotyping or imputing all the polymorphisms within a QTL in a sample size sufficient to provide enough power to detect recombination between some of these associated variants and the causal variant (Spain & Barrett, 2015). By providing an exhaustive catalogue of the genetic variation at the locus and breaking up the associated haplotype block, the statistical signal at the QTL is refined and smaller numbers of variants can be prioritised as candidate causative variants (McCarthy et al., 2008). For many years, this has been the rate limiting step in causative variant discovery as it requires targeted sequencing of the locus which is technically demanding, time consuming and can be very costly (Edwards, Beesley, French, & Dunning, 2013). However, the advent of high-throughput sequencing (refer to section 1.7) has made it much easier to provide a full catalogue of genetic variants at associated loci, such that now the main issue is how to characterise the functional impact of these implicated variants.

Characterising the functional impact of the statistically-indistinguishable variants is the predominant challenge in causative variant discovery. Evaluating the functional consequences of variants requires an array of approaches, including the use of computational predictive models, comparative genomics, and database searches to investigate the impact of a variant on gene function or at the cellular and/or organism level (MacArthur et al., 2014). While in some cases, variants can be considered strong candidates for the QTL effect if they disrupt the coding sequence of a gene with physiological relevance to milk composition and production, often there is no strong non-synonymous candidate polymorphisms, and there may be scant information about the possible involvement of the gene in mammary and lactation biology. Instead, deciphering the function of the individual candidate causative variants may require experimental testing to elucidate the mechanism by which each candidate variant might be influencing the trait.

1.6.1. Non-coding genetic variants and regulatory elements

Non-coding DNA does not encode protein-coding genes, and based on our limited understanding of the function of these sequences, it has previously been described as 'junk DNA'. The non-coding genome represents about 98% of the bovine genome sequence, and this large amount of non-coding DNA sequence and the fact that it is not constrained by the genetic code like coding DNA, has limited our understanding of the non-coding portion of the genome. As non-coding DNA can accumulate many more neutral polymorphisms, and has many diverse functions and mechanisms of regulatory control, the methods currently used to prescribe function to coding variants are not applicable.

The Encyclopaedia of DNA Elements (ENCODE) project is a large-scale inter-disciplinary project that was founded to investigate the functionality of the non-coding parts of the human genome. The primary discoveries of this project were that the majority of the genome shows biochemical activity and there is a large amount of regulatory functionality of the non-coding regions of the genome (Encode Consortium, 2012). Regulatory elements are defined as discrete genome segments that encode a defined product (e.g. non-coding RNA) or display a reproducible biochemical signature (e.g. transcription factor binding). The latter includes promoters, enhancers, silencers, and other functionally active sequences that interact with the cell's transcriptional machinery (Encode Consortium, 2012). While these regulatory sequences are predominately *cis*-acting, they generally function independently of orientation and at various distances from their target, and with a range of protein, co-factors and DNA sequences (Gaffney et al., 2012; Sanyal, Lajoie, Jain, & Dekker, 2012). The dynamic nature of these elements enables them to determine when, where and at what level each gene is expressed, making it imperative to study these elements in the correct physiological context.

Genetic variations coinciding with these elements can perturb the binding sites of transcription factors, local chromatin structure or co-factor recruitment, and ultimately lead to changes in the transcription of nearby gene(s) (Paul, Soranzo, & Beck, 2014). A genetic variant that overlaps a regulatory element can be considered as a strong candidate causal variant for a given QTL, providing a potential starting point as to the mechanism by which a variant may impact the phenotype.

1.6.2. Expression quantitative trait loci

Expression quantitative trait loci (eQTL) are regions of the genome containing DNA sequence variants that influence the expression of one or more genes. Much like bovine milk production or other physiological traits, the expression level of a gene can be viewed as a phenotype, and eQTL can be identified by studying a population of genotyped individuals (Albert & Kruglyak, 2015; Smith et al., 2013). Genetic variations that are associated with both the level of expression of a given gene, and phenotypic variation in a physiological trait, provide excellent evidence to separate associated variants from causative variants, and may reveal the identity of the causative gene, pathway and mechanism of phenotypic modulation (Lappalainen, 2015; Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008; Z. Wang, Gerstein, & Snyder, 2009).

To this end, eQTL mapping has become an important approach for understanding mechanisms underlying variation in complex traits. Many studies have identified genetic variants associated with complex human diseases that also modulate gene expression in both a *cis*- and *trans*- regulatory manner, demonstrated by their enrichment in regions annotated as active promoters, strong enhancers and in transcription factor binding sites (Lappalainen et al., 2013). Commonly, eQTLs have been found near transcription start sites (TSS) of target genes (McCarthy et al., 2008; Stranger et al., 2007).

1.7. Molecular methods and technologies to identify causative variants

To uncover the causative variant, and its mechanism of action, responsible for a given QTL in GWAS, extensive experimental follow-up is required. Traditionally this has been hindered by the difficulty of testing the function of individual candidate variants, particularly for non-coding variants (Edwards et al., 2013). More recently, two paradigms are emerging for experimentally assessing the impact of candidate causal variants. Firstly, ENCODE and subsequent projects are generating functional annotations at a genome scale allowing predictive models to be applied in non-protein-coding contexts, and the prioritisation of variants based on their overlap with molecular features or interactions. Secondly, the functional characterisation of regulatory variants using targeted genome editing is facilitating the high-throughput functional testing of candidate variants, providing evidence of the impact of genetically isolated candidate variants.

Both of these experimental approaches leverage high-throughput sequencing, which is massively parallel short-read sequencing that enables rapid and cost effective determination of DNA sequence. Compared with automated Sanger sequencing, which is considered as a first generation technology, the 'sequencing by synthesis' approach defined by the Illumina platform is much more cost effective and produces vastly more data than any other technology, and has revolutionised data acquisition in molecular genetics and genomics (Shendure & Ji, 2008). Third generation sequencing technologies are also emerging, including the development of 'long-read' sequencing by PacBio and Oxford Nanopore (Branton et al., 2008; Rhoads & Au, 2015). PacBio and Nanopore sequencing technology can routinely produce single sequencing molecules upwards of tens and hundreds of kb, respectively, resulting in the transformation of reference genome sequences and solving structural sequence issues.

The following sections will discuss these concepts, firstly focusing on two high-throughput sequencing applications, whole genome and RNA sequencing, given they play an integral part in this thesis.

1.7.1. Whole genome sequencing

Whole genome sequencing (WGS) is the process of working out the order of the dinucleotides (adenine (A), guanine (G), cytosine (C) and thymine (T)) for an entire genome, which in the case of *Bos taurus*, is approximately 2.6 billion bases of DNA in total. While the sequencing of the first human genome took over 10 years, and required an international collaboration at a cost of around USD$3 billion, whole genomes can now be sequenced for as little as $1-2,000 (Hayden, 2014). The increasing availability of WGS data presents opportunities to conduct association studies using rare and population-specific variants, and provide a full catalogue of genetic variants at associated loci (Brunham & Hayden, 2013). While the cost of WGS remains prohibitive to the sequencing of entire populations, it is becoming routine to conduct targeted sequencing of individuals that have contributed significant proportions of the alleles present in a population, and then impute the DNA sequence into the remaining individuals of a population for which low density genotyping information exists (Daetwyler et al., 2014).

1.7.2. RNA sequencing

High-throughput RNA sequencing (RNAseq), provides a quantitative measure of gene expression through the assessment of transcript abundance on a genome-wide scale (M. Li et al., 2013). As such, RNAseq can provide information on all the RNA transcripts that comprise a given transcriptome, including messenger RNA (mRNA), non-coding RNA and small interfering RNA and small nucleolar RNA, and enables the quantification of each transcript under different conductions e.g. developmental stages, disease states or in relation to genotype (refer to section 1.6.2) (Simon Anders & Huber, 2010; Z. Wang et al., 2009). As the expression of each transcript is measured by the relative number of individual sequence reads mapping to that gene, RNAseq also provides a count of RNA from each known gene, exon or isoform (Simon Anders et al., 2013; Simon Anders & Huber, 2010).

Compared with microarrays, RNAseq requires less RNA input, is less noisy and has a much larger dynamic range (Mortazavi et al., 2008). Most importantly, RNAseq can identify new transcript isoforms and other RNA features permitted through reading the base composition directly (for e.g. quantification of allele-specific expression), whereas the detection capability of a microarray is limited by the pre-defined probe content of the array

(Z. Wang et al., 2009). As such, RNAseq experiments provide a unique and unparalleled resource for annotating gene structures, such as determining transcription start sites (TSS), 5′ and 3′ ends, splicing patterns and post-transcriptional modifications (Simon Anders & Huber, 2010; Smith et al., 2013).

1.7.3. Genome-wide annotation of functional regulatory elements

Given the majority of GWAS signals reside in non-coding genomic DNA sequence, considerable effort has been expended to understand the functions of non-coding DNA. The ENCODE project was founded to uncover the function of the non-coding genome in humans, (Encode Consortium, 2012). The ENCODE and subsequent projects have used a number of high-throughput sequencing approaches including: WGS, RNAseq, ChIP-sequencing (ChIPseq; Arvey, Agius, Noble, & Leslie, 2012), DNase-sequencing (DNaseseq; Thurman et al., 2012), whole-genome bisulfite sequencing (WGBS; Cokus et al., 2008), assay for transposase-accessible chromatin (ATAC; Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013) to generate genome-wide annotations of functional regulatory elements. These analyses enabled the systematic mapping of transcription factor binding sites, DNA-protein interactions, DNA methylation, chromatin structure and histone modifications, all on a genome-wide scale in hundreds of cell types to generate a catalogue of regulatory regions in the human genome.

Importantly, variants associated with these functional annotations have been shown to be enriched in GWAS loci for human diseases and complex traits (Maurano et al., 2012; Schaub, Boyle, Kundaje, Batzoglou, & Snyder, 2012). Predominately, this has been demonstrated to be the result of *cis*-regulation of the nearby genes, typically through influencing the activity of transcriptional enhancers and silencers, which often reside several kb from the gene that they influence (Mercer et al., 2013; Ziemann, Kaspi, Lazarus, & El-Osta, 2013). These functional annotations provide insight into the mechanism by which non-coding variants are potentially influencing molecular functions, linking these polymorphisms to the phenotype.

Furthermore, the use of high-throughput sequencing in this context can provide a highly sensitive quantitative measurement on a genome scale, and enables these genomic and functional annotations to be viewed as a quantitative trait (much like gene expression) and QTLs for these active regulatory elements can be identified by studying a population of genetically different individuals (Song & Crawford, 2010).

1.7.4. Genome editing to test function of candidate causative variants

Genome editing broadly describes genetic engineering approaches whereby DNA is inserted, deleted or replaced in the genome of an organism, typically using nuclease-based methods. These nucleases act as molecular scissors to create site-specific double-strand breaks (DSB) in the genome, which are repaired through the cell's own DNA-repair machinery comprising non-homologous end-joining (NHEJ) or homologous recombination (HDR) methods. The NHEJ repair pathway has the highest activity and involves the ligation of the two ends of the DSB without the use of a homologous template, whereas HDR requires a homologous sequence to guide the repair of the DSB (Fa Ann Ran et al., 2013). As the NHEJ pathway does not have a sequence to guide repair of the DSB, it is error prone and tends to result in indel mutations that may cause frame interruption and functional knockout of target genes (Rouet, Smih, & Jasin, 1994). In contrast, the HDR pathway results in accurate DNA sequence repair at the DSB, however is also much less efficient than NHEJ which can occur throughout the cell cycle (Bétermier, Bertrand, & Lopez, 2014).

Several different nuclease systems exist including zinc finger nucleases (ZFN; Urnov, Rebar, Holmes, Zhang, & Gregory, 2010), transcription activator-like effector nucleases (TALEN; Joung & Sander, 2013) and the clustered regularly-interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) protein systems (Jinek et al., 2012). These technologies have revolutionised targeted genetic editing in mammalian cells and have been used to knock-out or knock-in genes, to make allelic mutants, to change gene-regulatory control and to add reporters or epitope tags, all in the endogenous genomic context (Cyranoski, 2016; Niu et al., 2014; T. Wang, Wei, Sabatini, & Lander, 2014; Wu et al., 2014).

In particular, the CRISPR-Cas9 system is rapidly revolutionising the targeted editing of mammalian genomes with unparalleled efficiency, ease of use and scalability. This thesis evaluates the CRISPR-Cas9 mediated genome editing system, and aspects of the approach are discussed below (for a full review of targeted genome editing using TALENs and ZFNs see Joung & Sander, 2013; Petersen & Niemann, 2015).

1.7.4.1. CRISPR-Cas9 genome editing

The CRISPR-Cas9 system is part of the adaptive immune system of archaea and most bacteria, used as a defence against foreign nucleic acids, such as viruses and plasmids (Bhaya, Davison, & Barrangou, 2011; Terns & Terns, 2011). In the Type II CRISPR-Cas system, CRISPR loci typically consist of a clustered set of Cas genes and the CRISPR array, which is a series of repeat sequences that contain variable sequences called protospacers (Patrick, Eric, & Zhang, 2014). These protospacers are short segments of 20 – 30 base pair (bp) 'spacer DNA' which has accumulated from previous exposures to foreign DNA. These protospacers are transcribed into CRISPR RNAs (crRNAs) and hybridise with a second RNA, known as the transactivating crRNA (tracrRNA) and these two RNAs complex with the Cas9 nuclease (Jinek et al., 2012).

The CRISPR-Cas9 complex recognizes its DNA target through Watson–Crick base-pairing interactions between the protospacer sequence and target DNA, and through Cas9's interactions with the protospacers adjacent motif (PAM) site adjacent to the gRNA targeting site (H. Wang, Russa, & Qi, 2016). The wild-type *S.pyogenes* Cas9 is known to make a blunt cut between the 17th and 18th bases in the target sequence, which is 3 bp 5′ of the PAM site (NGG). As a result, the CRISPR-Cas9 system can easily be retargeted to cleave virtually any DNA sequence by redesigning the crRNA, assuming a PAM motif is also present.

The CRISPR-Cas9 technology provides the opportunity to systematically analyse gene function in mammalian cells and experimental systems (Doudna & Charpentier, 2014). In particular, the addition of a donor DNA template with homologous sequence either side of the DSB can be used by the HDR pathway to insert specific DNA sequences at the target location. This approach could be used to introduce specific alleles that associate with, and therefore might have a causal role in, milk composition regulation. It is clearly advantageous to obtain specific, targeted edits, versus the "take-what-you-get" modifications resulting from NHEJ, but as described above, this comes at the cost of reduced efficiency. The disadvantage of this reduction in efficiency is that HDR-mediated editing often necessitates single-cell cloning to isolate the small percentage of cells with the desired modification. This is in contrast to high efficiency NHEJ-based methods where the analyses may be conducted on heterogeneous pools of cells, or if pure colonies are desired, the screening of edited clones can be done at a lower throughput.

1.8. Scope of this thesis

This thesis describes the detailed investigation of four genomic loci that have large effects on milk production and composition traits in NZ dairy cattle. The research presented in the following chapters is focused on mechanistically linking a causative variant to the gene expression and milk composition effects at each of these loci. To this end, much of my work has utilised a high-depth mammary RNA sequence dataset to look for milk production-associated variants that also impact the expression of mammary genes, helping to identify the variants (and genes) responsible. For candidate variants identified in these analyses, further statistical, bioinformatic and lab-based experiments were conducted to provide additional lines of evidence for the genes and mutations implicated. These experiments have involved association analysis of variants with a range of novel phenotypes, cell-culture based modelling of variants, and analysis of gene expression by quantitative PCR. Each results chapter concludes with a discussion and conclusion, which are then drawn together in the last chapter, the general discussion. Taken together, this work has resulted in an improved understanding of the mechanisms linking these genetic variants to bovine milk composition, and raises the profile of these and other variants as potential tools for selection of NZ dairy animals with desirable production characteristics.

Chapter 2: General Materials and Methods

This chapter describes the general materials and methods used in the experimental chapters in this thesis. The 'Methods' section within individual chapters contains detailed descriptions of methods pertaining to specific experiments.

2.1 General materials

2.1.1 Chemicals and reagents used in this research

Chemical/Reagent	Supplier
Agarose LE multipurpose agarose	Axygen Scientific Inc.
AxyPrep Plasmid Miniprep Kit	Axygen Scientific Inc.
AxyPrep PCR Clean-up Kit	Axygen Scientific Inc.
AxyPrep DNA Gel Extraction Kit	Axygen Scientific Inc.
Real time PCR primers and probes	Integrated DNA technologies
PCR primers as specified	Integrated DNA technologies
Nuclease-Free Duplex buffer	Integrated DNA technologies
RedSafe™ Nucleic Acid Staining Solution	iNtron Biotechnology
DMEM (Dulbecco's Modified Eagle Medium)	Invitrogen/Life Technologies
1Kb Plus DNA Ladder	Invitrogen/Life Technologies
Gibco® 0.25% Trypsin-EDTA	Invitrogen/Life Technologies
Gibco® Foetal bovine serum, qualified, NZ origin	Invitrogen/Life Technologies
Gibco® Penicillin-Streptomycin	Invitrogen/Life Technologies
Gibco® 1x PBS	Invitrogen/Life Technologies
Gibco® Trypan-Blue	Invitrogen/Life Technologies
One Shot® TOP10 Chemically Competent *E.Coli*	Invitrogen/Life Technologies
SOC medium	Invitrogen/Life Technologies
Ultra-Pure water/DNase and RNase free	Invitrogen/Life Technologies
Ambion® DNA-free™ Kit	Invitrogen/Life Technologies
SuperScript® III First-Strand Synthesis SuperMix Kit	Invitrogen/Life Technologies
Recovery™ cell culture freezing media	Invitrogen/Life Technologies

DNA sample loading dye	Kapa Biosystems
KAPA Universal ladder	Kapa Biosystems
KAPA Fast Probe qPCR Mastermix	Kapa Biosystems
KAPA2G™ Robust PCR Kit	Kapa Biosystems
Universal Probe Library	Roche
Insulin solution human	Sigma-Aldrich
Progesterone	Sigma-Aldrich
Prolactin from sheep pituitary	Sigma-Aldrich
Chloroform	Sigma-Aldrich
Phenol	Sigma-Aldrich
Phenol-Chloroform	Sigma-Aldrich
dNTPs	Roche
TRIzol Reagent	Ambion
Kanomycin	Sigma-Aldrich
Dithiothreitol (DDT)	Sigma-Aldrich
AMPure Beads	Agencourt
T7EI Endonuclease	New England BioLabs
pMAXGFP Plasmid	Lonza
Lipofectamine® LTX	ThermoFisher
Lipofectamine® PLUS Reagent	ThermoFisher
Lipofectamine® RNAiMAX	ThermoFisher
BD FACS™ Pre-Sort buffer	BD Biosciences
ExoSAP-IT PCR Product Cleanup	Affymetrix

2.1.2 Buffers and media

TNE buffer: 1 M TrisHCl pH 7.5, 0.5 M EDTA pH 8.0, 5 M NaCl

Digestion buffer: 10 mM Tris pH 7.5, 10 mM EDTA pH 8.0, 10 mM NaCl, 2% SDS, 39 mM DTT

Cell lysis buffer: 10 mM Tris, 10 nM EDTA, 2% SDS, 300 mM NaCl

Cell lysis buffer for direct PCR: 10 mM Tris, 10% Triton-X 100, pH 8.0

Tris buffered EDTA (TBE) buffer: 10 mM Tris, 1 mM EDTA, pH 8.0

Luria-Bertani (LB) medium: 1 g Yeast extract, 2 g Tryptone, 2 g NaCl, H_2O to 200 mL, pH 7.5

DMEM cell culture proliferation medium: 10% v/v foetal bovine serum, 1% Penicillin-Streptomycin, 5 μg/mL insulin, 1 μg/mL progesterone in DMEM

DMEM cell culture differentiation medium: 10% v/v foetal bovine serum, 1% Penicillin-Streptomycin, 5 μg/mL insulin, 10 μg/mL dexamethasone and 5 μg/mL prolactin in DMEM

2.1.3 Animal cohorts used in this thesis

Four animal cohorts were used throughout the experiments described in this thesis. These cohorts are categorised into the Friesian x Jersey cross-bred (FJXB) animals, the mammary biopsy RNA sequencing (RNAseq) cohort, and the mixed ancestry (MA) dairy cows, and Livestock Improvement Corporation (LIC) high-merit bulls. A brief description of these cohorts is provided below, with particular focus on the phenotypic and genotypic data that were leveraged in this thesis.

All animal experiments were conducted prior to this thesis, representing pre-existing data. Data were collected in strict accordance with the rules and guidelines outlined in the New Zealand (NZ) Animal Welfare Act 1999. The majority of the data used in this thesis were generated as part of routine commercial activities, and did not require ethical approval. For the needle biopsy of mammary tissue for RNAseq (see 2.1.3.2), protocols were approved by the Ruakura Animal Ethics Committee, Hamilton, NZ (approval AEC 12845).

2.1.3.1 Friesian-Jersey trial animals

The Friesian Jersey cross-bred (FJXB) trial animals represented a pedigree of Friesian (F) and Jersey (J) animals that were extensively phenotyped with the aim of identifying QTL for traits of economic interest for the NZ dairy industry. The FJXB design rationale is extensively detailed elsewhere (Spelman, Miller, Hooper, Thielen, & Garrick, 2001). Briefly, an F_2 experimental design with a half-sib structure was undertaken with NZ F and J cattle to identify the genetic differences between the two breeds. Six F_1 sires were mated to F_1 commercial cows to produce over 800 F_2 calves. The F_2 progeny were phenotyped extensively over their lifetime for traits including milk composition and production, health and disease, and reproductive performance (see Spelman, Hooper, Stanley, Kayis, & Harcourt, 2004; Spelman et al., 2001).

2.1.3.2 RNAseq biopsy animals

The RNAseq animal cohort is comprised of 406 mostly Holstein-Friesian NZ dairy cows, representing three subgroups biopsied at different time points. Tissue samples were taken by needle biopsy for all animals (as described in Littlejohn et al. 2014) from their mammary gland during lactation and total RNA was extracted by NZ Genomics Limited

(NZGL; Auckland, New Zealand). The first cohort comprised 27 HFxJ animals from the FJXB pedigree described in section 2.1.3.1. The second cohort was made up of 193 mostly HF animals and were sampled in January 2013. The final 186 animals were sampled in December 2013, and represent mostly HF cows in their third or fourth lactation. All procedures for sample collection were undertaken with the approval of the Ruakura Animal Ethics Committee, Hamilton, NZ (approval AEC 12845).

2.1.3.3 Mixed ancestry dairy cows

The mixed ancestry dairy cow population is a large population consisting of approximately 65,000 dairy cows located on commercial dairy farms throughout NZ, forming a large phenotypic and genotypic database of animals used for evaluation of sire performance. All animals were born between 1998 and 2013, with the majority of records for animals born after 2004.

This population consists of a mixture of Holstein-Friesians, Jerseys, and Holstein-Friesian x Jersey crossbreeds; the specific numbers of each breed are detailed for each analysis described in the relevant chapters. Unless otherwise specified, 'purebred' animals were defined as having a breed proportion of at least 13/16ths. There were a small number of animals used in the analysis that were more than 4/16ths other breeds. Other breeds that were present in the date set included small numbers of Ayrshire, Brown Swiss, Guernsey, Hereford, Milking Shorthorn, and Swedish Red.

Differences in the number of animals quoted in each analysis in this thesis are a reflection of the impact of genotype and phenotype quality filtering as well as whether or not the animal had lactation records at the time the analysis was conducted.

2.1.3.4 LIC sires

Semen samples were collected from sires as part of LICs semen collection protocol, and used for genomic DNA extraction (section 2.2.3.1). Straws were manually checked to confirm correct animal key and name prior to DNA extraction.

2.1.3.5 Milk production and gene expression phenotypes

Milk composition and liveweight phenotyping: For the mixed ancestry population, the concentrations of major milk components were measured as part of standard herd testing procedures using Fourier transform infrared spectroscopy. Most milk samples were processed by LIC TestLink (Newstead, Hamilton, NZ) using the MilkoScan FT6000 instrument (FOSS, Hillerød, Denmark). Liveweight records were restricted to 2 year olds, representing weight measurements where the animal walked over a scale or weights were estimated from visual scores from certified assessors. These records were adjusted for age of calving, stage of lactation and weighted to account for unequal variances.

For the FJXB animals, milk composition was measured using the herd test results from a three-month period during the animals' second lactation. This lactation data was also used for fatty-acid analysis of milk fat, with milk samples taken at peak (September/October), mid (November) and late lactation (February). Fatty acids were extracted by a modification of the Röse Gottlieb technique, and quantified by gas-liquid chromatography on a Shimadzu GC17A instrument (Shimadzu Corporation) at Fonterra Research Centre in Palmerston North, NZ. The relative proportions of individual fatty acids were calculated as grams (g) per 100 g of total fatty acid.

RNA sequencing: RNA sequencing of the 27 samples collected in 2004 and 2012 was conducted by NZGL (Dunedin, NZ) using the Illumina HiSeq 2000 instrument. For these samples, libraries were prepared using the TruSeq RNA Sample Prep Kit v2 (Illumina). RNA sequencing of the 193 samples collected in 2013, and 186 samples collected in 2013, was carried out by the Australian Genome Research Facility (AGRF; Melbourne, Australia) using the Illumina HiSeq 2000 instrument. For these samples, libraries were prepared using the TruSeq Stranded Total RNA Sample Prep Kit (Illumina) with ribosomal depletion using Human/Mouse/Rat Ribo-Zero kit (Epicentre/Illumina). All samples were sequenced using a 100 base pair (bp) paired-end protocol, with two samples multiplexed per lane.

RNA sequence informatics: Library preparation, read-mapping and data quality filtering of the RNAseq data was conducted prior to the experiments presented in this thesis (detailed in Littlejohn et al., 2014, 2016). Briefly, RNA sequence data representing the 406 animals was mapped to the UMD3.1 genome using Tophat2 (version 2.0.12) (D. Kim et al., 2013), locating an average of 88.9 million read-pairs per sample. Cufflinks software (version 2.1.1) (Trapnell et al., 2010) was used to quantify expressed transcripts, and yielded fragments per kilobase of exon model per million mapped (FPKM) expression values. The read counts from Cufflinks were also processed using the variance-stabilising transformation (VST) normalisation method in DESeq (version 1.18) (Simon Anders & Huber, 2010) to derive gene expression phenotypes suitable for linear model analysis, and subsequent expression quantitative trait locus (eQTL) mapping.

2.1.3.6 DNA extraction and high throughput genotyping

Genomic DNA extraction for all animal populations has been previously described (S. Berry et al., 2013; M. D. Littlejohn et al., 2016; M. D. Littlejohn, Tiplady, et al., 2014; Spelman et al., 2001). Briefly, genomic DNA was extracted from ear-punch tissue or blood by GeneSeek (Lincoln, NE, USA) and processed using Qiagen BioSprints kits (Qiagen). Genomic DNA was also extracted by GeneMark (Hamilton, NZ) and processed using MagMAX system (Life Technologies). All high throughput genotyping was performed by GeneSeek, with animals typed across a range of platforms including GeneSeek Genomic Profiler BeadChip (SuperGGP; GeneSeek/Illumina), the Illumina BovineSNP50 BeadChip (Illumina), or the Illumina BovineHD BeadChip (Illumina).

Those animals genotyped using the SuperGGP or Illumina Bovine SNP50 BeadChip, were imputed to the Illumina Bovine HD BeadChip markers using BEAGLE software (Browning & Browning, 2009). Subsequently, sequence-based genotypes were imputed into these animals, using a reference population of 556 animals using Beagle v4 (Browning & Browning, 2009), as described in Littlejohn et al., (2016). Briefly, imputation was done in sequential steps, where those animals genotyped on the Super GGP and SNP50 genotypes first imputed to the BovineHD variant set. After which, all animals with BovineHD genotypes were imputed to the sequence-based variant set.

Whole genome sequencing: Whole genome sequencing was conducted on a collection of sires which represent the wider commercial dairy population of NZ. Outbred individuals were selected for sequencing based on specific phenotypes of interest, or their representativeness of the population structure found in NZ. Whole genome sequencing has been described previously (M. D. Littlejohn et al., 2016; M. D. Littlejohn, Henty, et al., 2014).

Briefly, 100 bp paired end sequencing was performed by Illumina FastTrack using the Illumina HiSeq 2000 instrument.

Table 1.1 Animal populations investigated in this thesis

Population	Animal N	Chapter	Analysis
FJXB	826 F_2 cows	6	Genotyped *AGPAT6* variable number tandem repeat (VNTR) via GeneScan
	6 F_1 sires	6	Genotyped *AGPAT6* VNTR via Sanger sequencing
	617 F_2 cows	4	Association analysis with fatty acid profiles
	6 F_1 sires and dams	4	Local refinement of reference assembly
RNAseq	375 cows	4, 5, 7	Lactating mammary gland transcriptome profiling i.e. eQTL mapping, *trans*-eQTL analysis, gene expression profiling, splicing efficiency phenotyping
Mixed Ancestry (MA)	39442 cows	7	Association analysis of chromosome 14 liveweight locus with milk production traits
	37236 cows	6	Association analysis of chromosome 27 milk fat percentage locus with milk production traits
LIC bulls	13 sires	4	Detection of genomic location of breakpoints of copy number variant (CNV) and genotyping of this candidate variant for association analysis

2.2 General methods

2.2.1 Databases used

Sequence databases: Bovine genomic DNA and mRNA sequences used in this research were obtained from GenBank (https://www.ncbi.nlm.nih.gov/genbank/). The genome reference build UMD3.1/Btau6.1 was used for all work presented in this thesis. Accession numbers are quoted where relevant.

Primer design: Primers for PCR were designed using Primer 3.0 software (version 0.4.0, website). Primers for qPCR were designed using the Roche Universal Probe Library Assay Design Centre (https://lifescience.roche.com/en_nz/brands/universal-probe-library.html) and Primer 3.0. The Basic Alignment Search Tool (BLAST) (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was used to check specificity of primers. All PCR primers were manufactured by Integrated DNA Technologies (IDT; Singapore).

2.2.2 Software used

Alignment visualisation software: Geneious software (version 6.0.3) was used to store and visualise sequence information when designing genotyping assays, and examine Sanger sequence traces. Integrated Genomic Viewer (IGV) was utilised to visualise RNA and DNA alignments to the reference genome (UMD3.1).

LightCycler software: Roche LightCycler® 480 software (Release 1.5.0 SP4, version 1.5.0.39) was used for the analysis of real-time qPCR data. The Abs Quant/2nd Derivative Max function was used to construct standard curves for assays and calculate the PCR efficiency. The Advanced Relative Quantification function was used to quantify the expression of target genes between samples and relative to that of selected reference genes.

BD FACS Aria II: FlowJo v10.1 (www.flojo.com) was used for visualisation and analysis of cell sorting and fluorescence data.

2.2.3 Cellular and molecular biology

2.2.3.1 Genomic DNA extraction from semen and from cell culture

Genomic DNA was extracted from two semen straws for each sire. Briefly, 500 µL of TNE buffer was added to the semen straw solution in a 1.5 mL Eppendorf tube (Eppendorf). The solution was vortexed, and then centrifuged for 5 minutes at 10,000 g. The supernatant was removed before another 500 µL of TNE buffer was added to the pellet and 'muddled' with a pipette tip. The solution was vortexed, and then centrifuged for another 5 minutes at 10,000 g. After which, the supernatant was removed and 400 µL digestion buffer (as described above) was added to the pellet. The pellet was 'muddled' again before vortexing. Then, 10 µL proteinase K solution (10 mg/mL) was added to the solution and incubated at 37°C overnight.

The following day, the solution was subjected to a standard phenol-chloroform extraction and ethanol precipitation protocol. First, 300 µL phenol was added to the solution which was inverted 20 times before being centrifuged for 10 minutes at 10,000 g. The upper phase was recovered into a new 1.5 mL tube, and the organic phase discarded. Then, 400 µL phenol-chloroform was added to the tube, which was inverted 20 times and centrifuged for 10 minutes at 10,000 g. Then, the upper phase was recovered into a new 1.5 mL tube and 400 µL chloroform was added, and once again the tube inverted 20 times and centrifuged for 10 minutes at 10,000 g. Finally, the upper phase was collected into a new 1.5 mL tube, and 40 µL 3M NaAc (pH 5.2) and 880 µL 100% cold ethanol added. The tube was inverted 30 times before being centrifuged at top speed for 20 minutes at 4°C. The supernatant was then carefully removed and the pellet washed with 300 µL of 80% ethanol and centrifuged at top speed for 15 minutes at 4°C. Again, taking care not to disturb the pellet, the supernatant was removed and the pellet was left to air dry for 15 minutes before re-suspending in 100 µL TE. The re-suspended pellet was incubated at 37°C overnight before quantification.

Genomic DNA was extracted from cell culture (see section 2.2.4) using a similar protocol, with some slight modifications. Briefly, cells were removed from the wells and pelleted before media was replaced with 200 µL cell lysis buffer. The cells were incubated in the cell lysis buffer for a minimum of 5 minutes at 37°C before 200 µL phenol was added to the solution and inverted 20 times. Then, the above semen extraction protocol was carried out, adjusting the volumes where relevant i.e. adding an equal volume of reagent. The resulting DNA pellet was resuspended in 10 µL TE before quantification.

2.2.3.2 DNA/RNA quantification

DNA and RNA were quantified using the NanoDrop spectrophotometer (Nanodrop Technologies; Wilmington, USA), the Qubit Fluorometer (Invitrogen) or 2100 Bioanalyser (Agilent), as specified.

2.2.3.3 Polymerase chain reaction

Polymerase chain reaction (PCR) was performed using KAPA Robust DNA polymerase (Kapa Biosystems) according to manufacturer's specification. Unless otherwise specified, a 10 µL reaction contained 0.5 µM of each forward and reverse primer, 200 µM of each dNTP, 0.2 U of DNA polymerase and 20 ng of genomic template DNA. Cycling conditions were: initial denaturation at 95^{o}C for 3 minutes, 35 cycles of denaturation at 95^{o}C for 30 seconds, annealing at 56-60^{oC} for 30 seconds, extension times were specific to PCR product size, but were generally 1 minute/kilobase (kb).

2.2.3.4 Gel electrophoresis of DNA

PCR products and genomic DNA were separated by gel electrophoresis using Agarose LE multipurpose agarose. Agarose gels were 1-3% w/v made in 1x TBE. DNA sample loading dye (Kapa Biosystems) was added to samples to a 1x final concentration and DNA size standards used were either KAPA Universal ladder or Express ladder (Kapa Biosystems) as specified. Samples were separated at 100 V or 70 V in Biorad gel tanks. Immediately following electrophoresis, gels were visualised using a GelDoc imager (Biorad).

2.2.3.5 Sanger sequencing of DNA

Sanger sequencing was performed by Kristine Boxen at the University of Auckland Centre for Genomics, Proteomics and Metabolomics (CGPM) on an Applied Biosystems 3130XL Genetic Analyser, using BigDye version 3.1 terminator chemistry on Applied Biosytems 9700 Gold Block thermal cyclers. PCR products were prepared for sequencing at 5 ng/100 bp/10 µL reaction. Sequencing primers were prepared at 5 pmol per reaction.

2.2.3.6 Transformation of chemically competent *E.coli* cells

One 50 µL vial of One shot® cells (Invitrogen) was used for each plasmid transformation. Cells were thawed on ice completely. One microliter of plasmid DNA (~10 – 20 ng) was added to each vial of cells. The tube was gently tapped several times to disperse the DNA and then incubated on ice for 30 minutes. The cells were heat shocked by incubating in a 42^{o}C water bath for 30 seconds before returning to ice. 200 µL of pre-warmed S.O.C medium was added to each vial before the vials were shaken horizontally at 37^{O}C for 1 hour at 225 rpm.

Twenty µL from each transformation vial was spread on a separate, labelled LB agar plate containing the selecting antibiotic (100 µg/mL ampicillin (Amp) or 50 µg/mL kanamycin (Kan) at final concentration) and incubated overnight at 37^{o}C. Alongside the experimental plasmids, 10 pg of pUC19 control plasmid DNA was transformed as a positive control.

2.2.3.7 Plasmid DNA purification

Single colonies were selected from Amp or Kan selective LB agar plates with a sterile pipette tip and used to inoculate 5 mL LB Broth (containing selecting antibiotic as above). The broth cultures were shaken at 180 rpm overnight at 37^{o}C. The following day, glycerol stocks were prepared from 600 µL culture and 600 µL of 100% glycerol, for storage at -80^{o}C. The remaining culture was centrifuged at 1500 g for 3 minutes and the resulting supernatant discarded before plasmid purification using the AxyPrep plasmid Miniprep Kit (Axygen) as per the manufacturers' standard spin protocol. Purified plasmid DNA was suspended in 50 µL eluent, quantified by Nanodrop, and stored at -20^{o}C.

2.2.4 Cell culture

The bovine mammary alveolar cell line (MAC-T; Huynh, Robitaille, & Turner, 1991) was used for all cell culture experiments. Cells were maintained in 75 mL flasks at 37^{O}C in 5% CO$_2$ in 15 mL DMEM (Life Technologies) based media. The proliferation media included 10% v/v foetal bovine serum, Penicillin-Streptomycin (each at 100 Units per mL), insulin and progesterone in DMEM. Cells were passaged every 3-4 days once 70-90% confluent, using 0.25% trypsin 0.03% EDTA (Invitrogen) to detach cells from the plastic surface. These cells were maintained in differentiation media during transfections with plasmids (see below), where the media contained 10% v/v foetal bovine serum, Penicillin-Streptomycin (each at 100 U per mL), insulin, dexamethasone, and prolactin in DMEM (concentrations as per Section 2.1.2).

2.2.4.1 Fluorescence assisted cell sorting

Fluorescence assisted cell sorting (FACS) was performed by Alicia Didsbury at the University of Auckland using a FACSAria™ II flow cytometer (BD Biosciences). Cell debris was excluded from analysis using bivariate, forward/side scatter (FSC/SSC) parameters and dead cells were gated from analysis using DAPI (provided). Cells were binned into negative and positive GFP fluorescent cell populations and sorted into 15 mL falcon tubes or 96-well plates.

2.2.4.2 Transfection of MACT-T cells

Lipofectamine® LTX & PLUS Reagent (Invitrogen) was used to transfect plasmid DNA into the MAC-T cells. For transfection with Lipofectamine® LTX PLUS Reagent, plasmids were transfected using 0.5 µL Lipofectamine® LTX, 0.5 µL PLUS Reagent and 50 µL Opti-MEM media. Reaction mixtures were incubated at room temperature for 5 minutes before transfection to allow for complex formation. Cells were transfected as a monolayer when they were ~70% confluent and incubated at 37°C.

2.2.4.3 Optimisation of Lipofectamine® LTX Reagent as a method for cell transfection

Lipofectamine® LTX was used for all transfections involving *DGAT1* plasmids. It was first necessary to determine what concentration would achieve maximum delivery of plasmid DNA into the MAC-T cell line. Four different concentrations of reagent are recommended for use with cell lines; 1U, 2U, 3U and 4U. To test the efficiency of these concentrations, transfections were conducted in duplicate at Passage 10, 18 and 19, both with and without Lipofectamine® LTX PLUS Reagent. Based on the ~40% transfection efficiency we observed across all of these concentrations, we then conducted transfections with lower concentrations of reagent in order to limit the cell death evident with increasing amounts of transfection reagent. Transfections were conducted in duplicate at passage 11, with 0.25, 0.5, 1 and 2 U of Lipofectamine® reagent.

2.2.4.4 RNA extraction from transfected MAC-T cells

After 24-48 hours (as specified), MAC-T cells transfected with nucleic acid (plasmids or gene editing reagents) were subjected to RNA extraction using TRIzol Reagent (Ambion). After removal of cell culture media, 1 mL of PBS was added directly to each well of the 24-well plate and gently washed. The PBS was removed, and an equal volume of TRIzol Reagent was added directly to each well of a 24-well plate. To ensure the detachment of the cells from the bottom of the well, the solution was passed through the pipette several times until there was a change in the consistency of the solution. The solution was then transferred to a 1.5 mL eppendorf tube and passed through a 25-gauge needle 20 times. Lysates were rested at room temperature for 5mins, 200 µL chloroform added and tubes shaken vigorously for 15 seconds before being left at room temperature for another 3 minutes. Samples were then centrifuged at 12,000 g for 15 minutes at 4 ^{0}C. The upper aqueous phase was transferred to a new microcentrifuge tube and the remaining interphase and organic phase discarded. 500 µL of isopropanol was added and inverted 20 times before leaving at 4^{o}C for 20 minutes. Then, the samples were centrifuged at 12,000 g for 15 minutes at 4^{o}C.

The RNA was suspended in 20 µL Ultra-Pure water (Invitrogen) was used to elute the RNA and transferred to a new microcentrifuge tube, and stored at -80^{o}C or immediately subjected to DNase treatment and cDNA synthesis.

2.2.4.5 DNase treatment of RNA

To remove traces of genomic DNA, RNA samples were DNase treated using the Ambion ® DNA-free™ Kit (Life Technologies). RNA samples (20 µL) were incubated with 2 U DNase 1 and 0.1 volume 10x DNase I buffer at 37^{o}C for 20-30 minutes. Then, 0.1 volume DNase Inactivation Reagent was added and incubated for 2 minutes at room temperature with occasional mixing before centrifugation at 10,000 g for 1.5 minutes. The RNA solution was then transferred to a clean eppendorf tube and subjected to a second DNase treatment, following the same steps as above.

RNA samples were quantified by Nanodrop prior to storage at -80°C. Samples with 260/280 absorbance ratio of 1.85 or above were used in subsequent experiments.

2.2.4.6 First strand cDNA synthesis

Complementary DNA (cDNA) was synthesised from DNase-treated RNA by reverse transcription polymerase chain reaction (RT-PCR) using SuperScript® III First-Strand Synthesis SuperMix Kit (Invitrogen) as per the manufacturer's instructions. Initial reaction mixtures contained RNA (amount specified in relevant chapters) from each sample, along with 1 µL of both random hexamer primers and annealing buffer and Ultra-Pure water in a total volume of 8 µL. Mixtures were incubated for 5 minutes at 65°C before being placed on ice for at least 1 minute, prior to the addition of 2x First-Strand reaction mix and 2 µL SuperScript® III/RNaseOUT™ enzyme mix to a final volume of 20 µL. Reactions were vortexed and centrifuged briefly before incubation for 10 minutes at 25°C, followed by 50 minutes at 50°C and termination of reactions at 85°C for 5 minutes. The reactions were then transferred to ice and cDNA aliquots were diluted 1:10 (unless otherwise specified) the resultant cDNA samples were either used immediately as templates for qPCR or stored at -20°C.

Reverse transcriptase (RT) negative controls were generated for each sample by conducting the same protocol, with the exception of excluding the enzyme.

2.2.4.7 Real-time PCR

Real-time PCR assays for target genes and intron-exon boundaries were designed to publicly available bovine gene sequences (NCBI; http://www.ncbi.nlm.nih.gov/gene) using the Roche Universal Probe Library Assay Design Centre and Primer3 (refer to section 2.2.1). The specificity of PCR primers was tested using BLAST (http://blast.ncbi.nlm.nih.gov/) and with the exception of the assays designed for the quantification of pre-mRNA assays, were designed to span an intron-exon boundary to prevent amplification of genomic DNA. For each primer pair, the Roche Assay Design Centre also identified an appropriate 5'FAM-labelled short (8-9 nucleotide) hydrolysis probe, which were part of the Roche Universal Probe Library (Roche; Mannheim, Germany). All primer and probe sequences are detailed in the relevant results chapters.

Quantification of transcripts was based on Roche probe chemistry and performed on the Roche LightCycler® 480 (Roche Diagnostics, Mannheim, Germany). The reaction volume was 10 µL consisting of 3 µL cDNA and 7 µL of master mix (5 µL of KAPA Fast Probe Mastermix; 0.4 µL of each 5 µM primer; 0.02 µL of probe and 1.58 µL of water). Reactions were set up using an ep*Motion* 5075 robot (Eppendorf). Standard cycling conditions were used; 95^{0}C for 10 minutes, 45 cycles of 95^{0}C for 10 seconds, 60^{0}C for 30 seconds, followed by a hold at 40^{0}C for 40 seconds.

Real-time PCR experiments were carried out in 384-well format with a single reaction per well. Each assay included a negative control (i.e. a no-template control with water added instead of cDNA) and a reverse transcriptase (RT) negative control for each experimental sample. Triplicate measurements were performed for all samples and standard curves with standard deviations less than 0.2 cycles were used for quantification. Following amplification, amplicons were subjected to gel electrophoresis to ensure each assay provided a single product of the expected size.

2.2.4.7.1 cDNA standards and relative quantification analysis

Serial 5x standard cDNA dilution series of 1:1, 1:5, 1:25, 1:125, 1:625 were created in Ultra-Pure water for the generation of standard curves by real-time PCR for each reference and target gene assay. Serial dilutions were created from pooled diluted cDNA from each experimental sample. Standard curves for all dilution series were generated using the Abs Quant/2nd Derivative Max LightCycler480® software function. The standard curves generated using this function were used to normalise the expression of samples for each real-time PCR assay. This function also calculates the efficiency of the PCR reaction across the known concentration range for each assay, with 2 being 100% efficient. The efficiencies of assays included in this thesis were between 1.79 and 1.97.

To quantify gene expression levels, relative quantification analysis was used which compared the expression levels of the target genes to the geometric mean of selected of endogenous control genes (*RPS15A*, *EIF3K* and *GFP*). This method was based on the E-Method of relative quantification (Tellmann, 2006), which is able to compensate for differences in the amplification efficiency of target and reference genes by normalising to the standard curve for each gene.

2.2.5 Genetic association analyses

Unless otherwise stated, all association analyses presented in this thesis were conducted using ASReml-R (A R Gilmour, Gogel, Cullis, & Thompson, 2009; Arthur R. Gilmour, Thompson, & Cullis, 1995). This statistical software package was selected as it efficiently fits linear mixed models using restricted maximum likelihood to large and complex datasets, such as those described in 2.1.3.

Association analyses between SNPs and milk production traits and gene expression phenotypes were quantified using restricted likelihood (REML) using pedigree-based mixed models in ASReml-R. Each SNP was fitted in a separate sire-maternal grandsire single trait model, with SNP treated as a quantitative variable based on the number of copies of the alternative allele and variance components estimated in a restricted maximum-likelihood (REML) framework. Covariates for sequencing cohort, the proportions of NZ Holstein-Friesian ancestry, US Holstein-Friesian ancestry, Jersey ancestry and heterosis effects were also included in the models. The additive genetic variance for each SNP was calculated using $\sigma^2_{SNP} = 2p(1-p)a^2$, where p is the frequency of the highest frequency allele and a is the estimated allele substitution effect. Polygenic genetic variances were evaluated as $\sigma^2_{anim} = 4\sigma^2_{sire}$ where $\sigma^2_{sire}$is the estimate of sire variance from the model. Total genetic variance was evaluated as $\sigma^2_g = \sigma^2_{SNP} + \sigma^2_{anim}$ and phenotypic variance was evaluated as $\sigma^2_p = \sigma^2_{SNP} + \sigma^2_{anim} + \sigma^2_e$ where σ^2_e is the residual variance. The proportion of phenotypic variance explained by each SNP for each phenotype was calculated as $\sigma^2_{SNP}/\sigma^2_p$ and the proportion of genetic variance explained by each SNP was calculated as $\sigma^2_{SNP}/\sigma^2_g$.

Chapter 3: Optimisation of CRISPR-Cas9 genomic editing of mammary cells

3.1. Overview

In recent years, the widespread application of genome-wide association studies (GWAS) has identified sequence variants associated with bovine milk composition and production. The ultimate goals of these studies are to provide information for genomic selection (GS), and define the genetic underpinnings of these traits to provide insight into mammary gland physiology. However, accomplishing the latter goal requires defining the causative variant(s) that is responsible for the genetic signal at a given locus; its mechanism of action, and implicated gene(s). In GWAS, it is often difficult to determine which genetic variant is responsible for the phenotype using statistical methods alone, as the tight linkage disequilibrium (LD) between markers provides multiple candidates to choose from. As such, dissecting statistically indistinguishable variants from each other requires functional experiments to provide information about the effect of specific associated variants. While the prediction of functional consequences for coding variants may be more straightforward, it is difficult to predict the effects of non-coding variants, given the diverse functions of non-coding DNA, the incomplete annotation of regulatory elements, and the many mechanisms of regulatory control.

Recent advancements in genomic and genetic technologies are providing new approaches to understand the function of non-coding genetic variants, gene function, and how this contributes to complex traits. Namely, the re-engineering of mammalian genomes using genome editing technologies; zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), engineered meganucleases, and most recently the clustered regularly-interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) protein system, enables the targeted investigation of specific variants in isogenic backgrounds. In particular, the CRISPR-Cas9 system is revolutionising the targeted editing of mammalian genomes with unparalleled efficiency, ease of use and scalability. This is reflected in the rapid evolution of the technology and its widespread application, including functional knockout of individual genes (Wagner, Platt, Goldfless, Zhang, & Niles, 2014), large scale knock-out and knock-in screens (T. Wang et al., 2014), *in vivo* mouse models (Aida et al., 2015), and human clinical trials (Cyranoski, 2016).

The first experimental chapter of this thesis describes the optimisation of CRISPR-Cas9 genome editing of a bovine mammary cell line. Experiments presented include the optimisation of the delivery method and CRISPR-Cas9 expression system, and the evaluation of the NHEJ and HDR editing efficiency of 28 gRNAs. Experiments towards multiplexing targets and the enrichment of transfected cells are also presented, and ultimately continued research using the tools generated here will allow for the establishment of efficient editing of mammary cell lines to study the effects of candidate causative variants for bovine milk composition and production traits.

3.2. General aim

Investigate the use of the CRISPR-Cas9 system as a method for the genome editing of a bovine mammary cell line.

3.3. Specific aims

1. Identify the optimal Cas9 and gRNA delivery method and expression system that achieves efficient genome editing at the *AGPAT6* chr27:36198117T>TGGC target locus.
2. Use the optimal CRISPR-Cas9 protocol identified above to compare the efficiency of the NHEJ and HDR method of gene editing at 28 target loci.
3. Identify the optimal concentration of ssODN HDR repair template to achieve maximal HDR at the *MGST1* chr5:93946027T>A target locus.
4. Identify the optimal CRISPR-Cas9 system that achieves efficient gene editing at multiple loci.
5. Identify a cell selection protocol that enriches for putative CRISPR-Cas9 genome editing events.

3.4. Methods

3.4.1. Research strategy

The research strategy presented in this chapter is adapted from the IDT CRISPR-Cas9 genome editing protocol (http://sg.idtdna.com/; see Chapter 1 for review of CRISPR-Cas9 genome editing). This protocol has been optimised for the lipofection of Alt-R CRISPR-Cas9 RNPs in HEK293 cells but has shown to be translatable to other adherent, immortalised eukaryotic cell lines. Optimisation experiments based on this protocol were conducted to find the conditions that demonstrate maximal editing efficiency and minimal cell toxicity in the bovine mammary epithelial cell line, MAC-T (Huynh et al., 1991). Following this, the efficiency of NHEJ and HDR editing at 28 loci, encompassing candidate causative variants at four QTL influencing bovine milk production, were investigated. The 28 target variants represent 13 candidate causative variants within the *MGST1* locus, 13 candidate causative variants in the *AGPAT6* locus, and one variant in both the *DGAT1* and *LGB* loci identified previously (Grisart et al., 2004; Kuss et al., 2003; M. D. Littlejohn et al., 2016; M. D. Littlejohn, Tiplady, et al., 2014).

To conduct CRISPR-Cas9 mediated genome editing of the MAC-T mammary cell line, it was first necessary to determine the genetic background of these cells in order to design appropriate gRNAs and ssODN HDR templates. Further, the endogenous expression of *AGPAT6*, *DGAT1*, *LGB* and *MGST1* in this cell line needs to be detectable by real time quantitative PCR (RT-qPCR) to ensure that the influence of specific genetic variants introduced through CRISPR-Cas9 on gene expression could be investigated.

3.4.1.1. 'Wild-type' genotype determination for target variants of the MAC-T cell line

The genotypes for the target genetic variants in Table 3.1 were established from whole genome sequencing (WGS) of genomic DNA extracted from an early passage of MAC-T cells. At passage (~90% confluent), the MAC-T cells were stripped from a T75 flask using 3 mL trypsin-EDTA (Invitrogen). Following trypsin treatment, cells were resuspended in 1 mL PBS (Invitrogen) and passed through a 25-gauge needle 20 times. Then, genomic DNA was isolated using an AxyPrep MiniPrep Kit (Axygen) as per the manufacturer's protocol. The isolated DNA was then purified using AMPure Bead purification (Agencourt), by adding a 1.2x volume of AMPure beads to the sample and following the manufacturer's protocol. The DNA was quantified and the quality of DNA was checked by electrophoresis on a 2% agarose gel. After quantification, an aliquot of the DNA was sent to the Australian Genome Research Facility (AGRF; Brisbane, Australia), who performed WGS using v3 sequencing chemistry and 2 x 125 bp read lengths on the Illumina HiSeq 2000 instrument.

Alignments were performed using Borrow-Wheeler Aligner mem (BWA mem; version 0.7.12; Li, 2013) using the UMD3.1 reference genome and default parameters. The mapped sequence data was interrogated at each of the target loci to determine the MAC-T genotypes for the variants in Table 3.1, and was performed manually to ensure high confidence of genotype assignment.

3.4.1.2. Gene expression phenotypes

RT-qPCR was used to quantify the endogenous expression of *AGPAT6*, *DGAT1*, *LGB* and *MGST1* in the MAC-T cell line. Assays for these genes were designed as described in General Methods, with the primer and probe sequences presented in Appendix I.

RNA was extracted from untransfected MAC-T cells (as described in General Methods), and was DNase-treated and quantified before use as input for cDNA synthesis by RT-qPCR. These reactions used the SuperScript® III First-Strand Synthesis SuperMix Kit (Invitrogen), with 2.5 µg DNase-treated RNA in each 20 µL reaction.

Serial 5x cDNA standard curves were generated from the cDNA and each concentration was run in triplicate for each of the assays. Assays for *EIF3K* and *RPS15A* were also conducted to serve endogenous control calculations, performed in the same way as that described for other assays.

All assays readily detected the different transcripts at PCR efficiencies between 1.82 and 2.04. The average relative expression of *AGPAT6*, *DGAT1*, *LGB* and *MGST1* is shown in Figure 3.1.

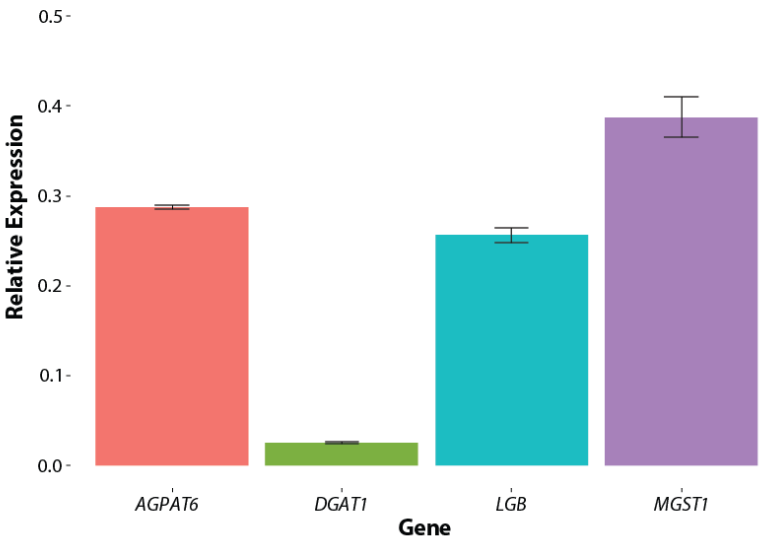


Figure 3.1 Gene expression of CRISPR-Cas9 target genes in the MAC-T mammary cell line
Gene expression values are relative to the geometric mean of *EIF3K* and *RPS15A* expression in the same samples, and normalised.

Table 3.1 Genotypes for target variants for CRISPR-Cas9 genome editing of mammary cell line. The favourable allele is associated with increased milk fat percentage.

Gene	Target bp	Target Variant	Genotype	Favourable allele
AGPAT6	27:36198117	TGGC/T	TGGC/T	TGGC
	27:36200888	T/C	T/C	T
	27:36200968	T/C	T/C	T
	27:36202188	T/A	T/A	T
	27:36202636	GT/G	G/TC	GT
	27:36203904	G/C	G/C	G
	27:36204066	T/C	T/C	T
	27:36204680	CAG/ATC	CAG/ATC	GAG
	27:36206783	C/A	C/A	C
	27:36209319	T/G	T/G	T
	27:36211257	GA/T	GA/T	GA
	27:36211708	T/C	T/C	T
	27:36212352	G/A	G/A	G
MGST1	5:93944937	T/C	T/T	C
	5:93945655	T/G	T/T	G
	5:93945738	T/C	T/T	C
	5:93946027	T/A	T/T	A
	5:93946548	G/C	G/G	C
	5:93947761	C/T	C/C	T
	5:93947989	T/A	T/T	A
	5:93948357	C/T	C/C	T
	5:93948646	C/G	C/C	G
	5:93948718	G/C	G/G	C
	5:93948804	T/C	T/T	C
	5:93949810	G/A	G/G	A
	5:93954748	T/C	T/T	C
DGAT1	14:1802265	GC/AA	GC/GC	AA
LGB	11:103301781	G/A	G/G	G

3.4.2. CRISPR-Cas9 design and components

Following sequence characterisation of the MAC-T cell line, design of the CRISPR-Cas9 components was undertaken using the reference sequence for chromosome 5 (chr5:93,925,899-93,972,185), chromosome 11 (chr11:103,301,488-103,306,381), chromosome 14 (chr14:1,795425-1,804,838) and chromosome 27 (chr27:36,198,042-36,229,006), in conjunction with the genotypes obtained as described in 3.4.1.1.

CRISPR RNAs (crRNAs) and ssODNs were designed for the target variants in Table 3.1 using Geneious software (version R9). For each target variant, one crRNA was selected based on 3 criteria: the proximity of the protospacer adjacent motif (PAM; NGG) site to the target variant, and requirement for high on-target and low off-target activity scores. In the case of the *MGST1* locus, there were four candidate causative variants that were in close proximity to the variants in Table 3.1. Given their proximity, these variants were targeted with the same gRNA, such that both alternative alleles were included in a single haplotype template (Table 3.2, see Appendix I). The ssODN repair templates targeting each of the genetic variants in Table 3.1 are presented in Appendix I. These templates were designed using the criteria described in (Richardson, Ray, DeWitt, Curie, & Corn, 2016), generating ~130 bp ssODN with asymmetric homology arms.

Table 3.2 ***MGST1*** **target variant 'haplotypes', where two target variants are included in the HDR template**

Target bp	Genotype	Second target bp	Genotype	Distance (bp)
chr5:93946027	TT	93945991	GG	36
chr5:93946548	GG	93946570	GG	22
chr5:93948718	GG	93948725	GG	7
chr5:93954748	TT	93954751	GG	3

Two CRISPR-Cas9 delivery systems were investigated in the current study; a Cas9 DNA expression vector used in conjunction with transfected gRNAs, and Cas9-gRNA RNP complexes. The components of these two systems are detailed below:

tracrRNA: Alt-R™ CRISPR tracrRNA is a conserved 67 nucleotide RNA provided by IDT.

crRNA: Alt-R™ CRISPR crRNAs containing a target-specific 19 or 20 nucleotide protospacer domain and a 16 nucleotide sequence complementary to the above tracrRNA. These RNAs were synthesised by IDT.

Cas9 plasmids: pU6-(BbsI)_CBh-Cas9-T2A-mcherry-P2A-Ad4E4orf6 and pU6-(BbsI)_CBh-Cas9-T2A-BFP-P2A-Ad4E1B were a gift from Ralf Kuehn (Addgene plasmid # 64222 and # 64218, respectively). Hereafter, these plasmids are referred to as mCherry and BFP, respectively. pSpCas9(BB)-2A-Puro (PX459) was a gift from Feng Zhang (Addgene plasmid # 62988)

Cas9 protein: Alt-R™ S.p. Cas9 Nuclease 3NLS is an endonuclease derived from S. *pyogenes*, and contains 1 N-terminal nuclear localisation sequence (NLS), 2 C-terminal NLSs, and a C-terminal 6-His tag. This Cas9 protein was synthesised by IDT.

gRNA complex formation: An equimolar concentration of tracrRNA and crRNA were added together to create a final duplex concentration of 1 µM in Nuclease-Free Duplex buffer (IDT). To form duplexes, the solution was heated at 95^{o}C for 5 minutes then allowed to cool to room temperature. Once at room temperature, the gRNA complex was either transfected into cells, used to form the RNP complexes, or stored at -20^{o}C until required.

RNP formation: The RNP complex was prepared immediately prior to each experiment. Cas9 protein was diluted to 1 µM using Opti-MEM reduced serum media (Invitrogen). Then equimolar concentrations of Cas9 and gRNAs were combined with Opti-MEM and incubated at room temperature for 5 minutes to assemble the RNP complexes.

HDR templates: The ssODNs described above were synthesised by IDT. All ssODNs were diluted to 1 µM/µL using Ultra-Pure water (Invitrogen).

Primers: PCR products encompassing each of the target sites were designed with primers flanking the locus. Primers were designed using Primer 3 software, and were synthesised with the addition of a Nextera adapter sequence to facilitate next generation sequencing (NGS) library preparation. Primer sequences for the 28 target variants are presented in Appendix I. Genomic DNA encompassing the target site was PCR amplified and analysed with the mismatch assay using T7E1 and used as input into NGS as described below.

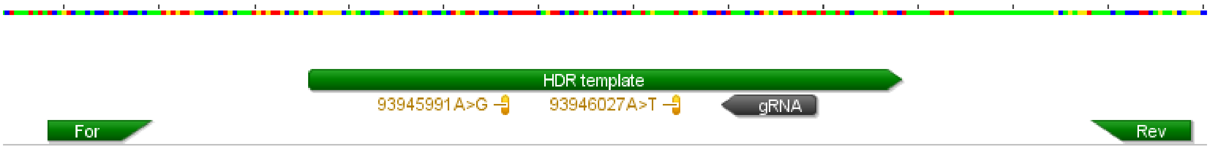


Figure 3.2 Location of CRISPR-Cas9 components for the chr5:93946027A>T target locus.
DNA binding sites for the gRNA is shown (grey) as well as the forward (For) and reverse (Rev) primers used for PCR of the target locus, and the HDR template for HDR (all dark green). The cleaving position of the gRNA is chr5:93946030, just 12 bp from the target variant.

3.4.3. DNA extraction and genome editing detection

For all experiments described in this chapter that test the efficiency of the above CRISPR-Cas9 components, genomic DNA was isolated using a standard phenol-chloroform extraction and ethanol precipitation following cell lysis of transfected cells. Briefly, 48 hours after transfection, the media was removed from the cells, and each well was washed using 100 µL pre-warmed PBS (Invitrogen). Following washing, 40 µL trypsin-EDTA (Invitrogen) was added to the cells and incubated at 37^{o}C for 10 minutes, or until the cells had detached. An aliquot of 120 µL of full proliferation media was then added to deactivate the trypsin, with the solution gently mixed before 80 µL of the cell suspension was transferred to a sterile 1.5 mL Eppendorf tube. The remaining 80 µL was transferred to a labelled 15 mL falcon tube and centrifuged at 1200 rpm for 5 minutes. The media was then removed and replaced with 100 µL freezing media (Invitrogen). The cells were quickly resuspended and the cell suspension transferred to a labelled 1 mL cyrovial tube (NUNC) and frozen at -1^{o}C/sec and stored at -80^{o}C.

The 1.5 mL tubes containing the cell suspensions were centrifuged at 12,000 g for 1 minute to pellet the cells. The media was gently removed and replaced with 200 µL cell lysis buffer (described in General Methods). The cells were incubated in the cell lysis buffer for a minimum of 5 minutes at 37^{o}C before being subjected to a standard phenol-chloroform and ethanol precipitation DNA extraction protocol (described in General Methods). DNA was quantified using the Nanodrop, where possible, approximately 20 ng of genomic DNA was used for input for the PCR reaction. For those samples that were under 20 ng/µL, 2 µL (<10 ng/µL) or 3 µL (<5 ng/µL) of DNA was used as the template for PCR.

Genomic DNA flanking the target site was amplified (see Appendix I for primer sequences) by PCR using the KAPA Robust PCR system (Kapa Biosystems). Amplification was performed in 25 µL reactions containing 0.5 U KAPA 2G Robust enzyme, and up to 20 ng genomic DNA (see General Methods for PCR reaction mix). Amplification conditions included initial denaturation at 95^{o}C for 3 minutes, followed by 35 cycles of 95^{o}C for 15 seconds, between 56^{o}C and 62^{o}C for 15 seconds (with annealing temperature dependent on the target locus), and 72^{o}C for 15 seconds. A final extension at 72^{o}C for 5 minutes was performed before cooling to 15^{o}C. The annealing temperature for the individual loci is included with the primer sequences in Appendix I.

3.4.3.1. T7E1 endonuclease assay

The editing efficiency of CRISPR-Cas9 can be estimated from the T7E1 mismatch endonuclease assay. In this assay, the PCR products from edited loci are denatured and re-annealed to allow heteroduplex formation between wild-type (WT) and edited DNA which are cleaved based on any mismatches as follows.

Aliquots of 1.5 µL of 10X NEBuffer 2 (New England BioLabs) and Ultra-Pure water were added to 10 µL of PCR product. In a thermal cycler, these products were incubated at 95^{o}C for 10 minutes before being cooled from 95-85^{o}C at a ramp rate of -2^{oC}/sec, and from 85-25^{o}C at a ramp rate of -0.3^{o}C/sec. Then, 6.5 µL of the PCR product is transferred to a new 0.2 µL PCR tube containing 1 µL T7 Endonuclease I (1U/µL; New England BioLabs), and incubated at 37^{o}C for 60 minutes. Both digested and undigested PCR products were visualised following separation by gel electrophoresis (3% w/v agarose) for 60 minutes.

3.4.3.2. Deep sequencing

Given the low sensitivity of the T7E1 mismatch detection assay and its inability to detect HDR events, sequencing of the PCR products encompassing each target was also conducted. The sequencing strategy for this work was 2 x 150 bp paired-end sequencing using the Illumina MiSeq platform. PCR products were barcoded and purified DNA samples were quantified using a Qubit 2.0 Fluorometer, that were subsequently pooled in an equimolar ratio. Sequencing libraries were then sequenced with the Illumina MiSeq Sequencer using a MiSeq 300 cycle Nano kit (Life Technologies), by New Zealand Genomics Limited (NZGL; Auckland, NZ).

3.4.3.3. Quantifying gene editing efficiency

The quantitative analysis of cleavage efficiency was conducted using the sequence data from PCR products encompassing the target loci. Alignments were performed using BWA mem (version 0.7.12) for each barcode using the UMD3.1/bosTau6 reference genome and default parameters. The mapped sequence data were sorted using samtools (version 1.3.1) before igvtools (version 2.3.82) was used to count the base depth and content at single nucleotide resolution across the amplicons. The aligned bam files were also visualised using IGV to manually interrogate the cleavage sites at each locus in the different experimental conditions.

To quantify NHEJ gene editing, the per base deletions or insertions were summed and reported as a proportion of the total read depth at that nucleotide, giving a quantitative measure of editing efficiency for each base in the amplicon. To quantify HDR gene editing, the presence of the introduced allele was summed and reported as a proportion of the max read depth at that nucleotide.

3.4.4. Optimisation of Cas9 plasmid-mediated CRISPR-Cas9 genome editing

To optimise the plasmid-based CRISPR-Cas9 expression system, experiments were conducted to identify the Lipofectamine® transfection reagent for maximal delivery of the Cas9 plasmid into the mammary cell line. Then, transfections of the Cas9 plasmid were performed using the optimised conditions followed by forward and reverse transfections of the gRNA complex 24 hours later. Prior to the transfection of the gRNAs, an aliquot of cells were also subjected to a FAC sort based on GFP fluorescence (Figure 3.3; detailed in Appendix I).

3.4.5. Optimisation of Cas9 protein-mediated CRISPR gene editing

To optimise the protein-based CRISPR-Cas9 expression system, experiments were conducted to identify the optimal Cas9 RNP and HDR concentration that results in maximal editing in the mammary cell line. Then, transfections of RNPs targeting 28 variants were performed using the optimised conditions. Additional experiments, targeting more than one locus in a single transfection and co-transfection of GFP plasmid were also conducted as detailed in Appendix I and Figure 3.4.

All experiments using CRISPR-Cas9 RNP complexes (with the exception of Experiment 5 described in 3.4.5.5) were conducted in 96-well format according to the IDT protocol (version 3.1). Cells were seeded at 15,000 cells/well approximately 18 hours prior to forward transfection of RNP complexes. All transfections were incubated for 48 hours, with a media change to full proliferation media 24 hours post-transfection. Following this incubation, DNA extraction and genome editing detection was carried out as described in 3.4.3 (Figure 3.4).

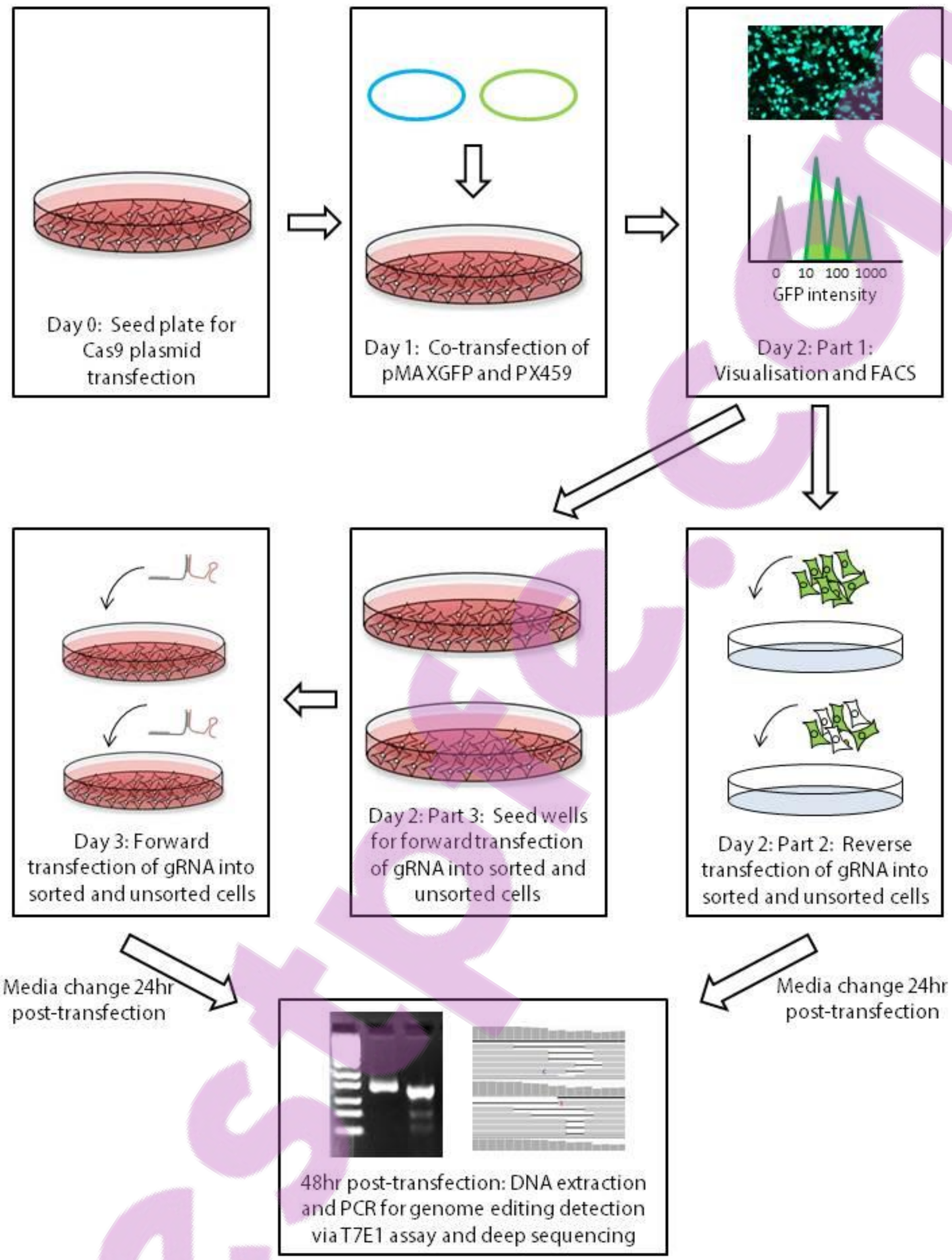

Day 0: Seed plate for Cas9 plasmid transfection
Day 1: Co-transfection of pMAXGFP and PX459
0 10 100 1000
GFP intensity
Day 2: Part 1: Visualisation and FACS
Day 2: Part 2: Reverse transfection of gRNA into sorted and unsorted cells
Day 2: Part 3: Seed wells for forward transfection of gRNA into sorted and unsorted cells
Day 3: Forward transfection of gRNA into sorted and unsorted cells
Media change 24hr post-transfection
Media change 24hr post-transfection
48hr post-transfection: DNA extraction and PCR for genome editing detection via T7E1 assay and deep sequencing

Figure 3.3 Schematic of the plasmid-based CRISPR-Cas9 genome editing protocol
Experiments were conducted to investigate the best delivery method for the CRISPR-Cas9 expression plasmid components, using the co-transfection of a GFP expression plasmid to enrich for transfected cells, and a forward and reverse transfection of gRNAs. The efficiency of these delivery methods was determined by deep sequencing of targeted loci.

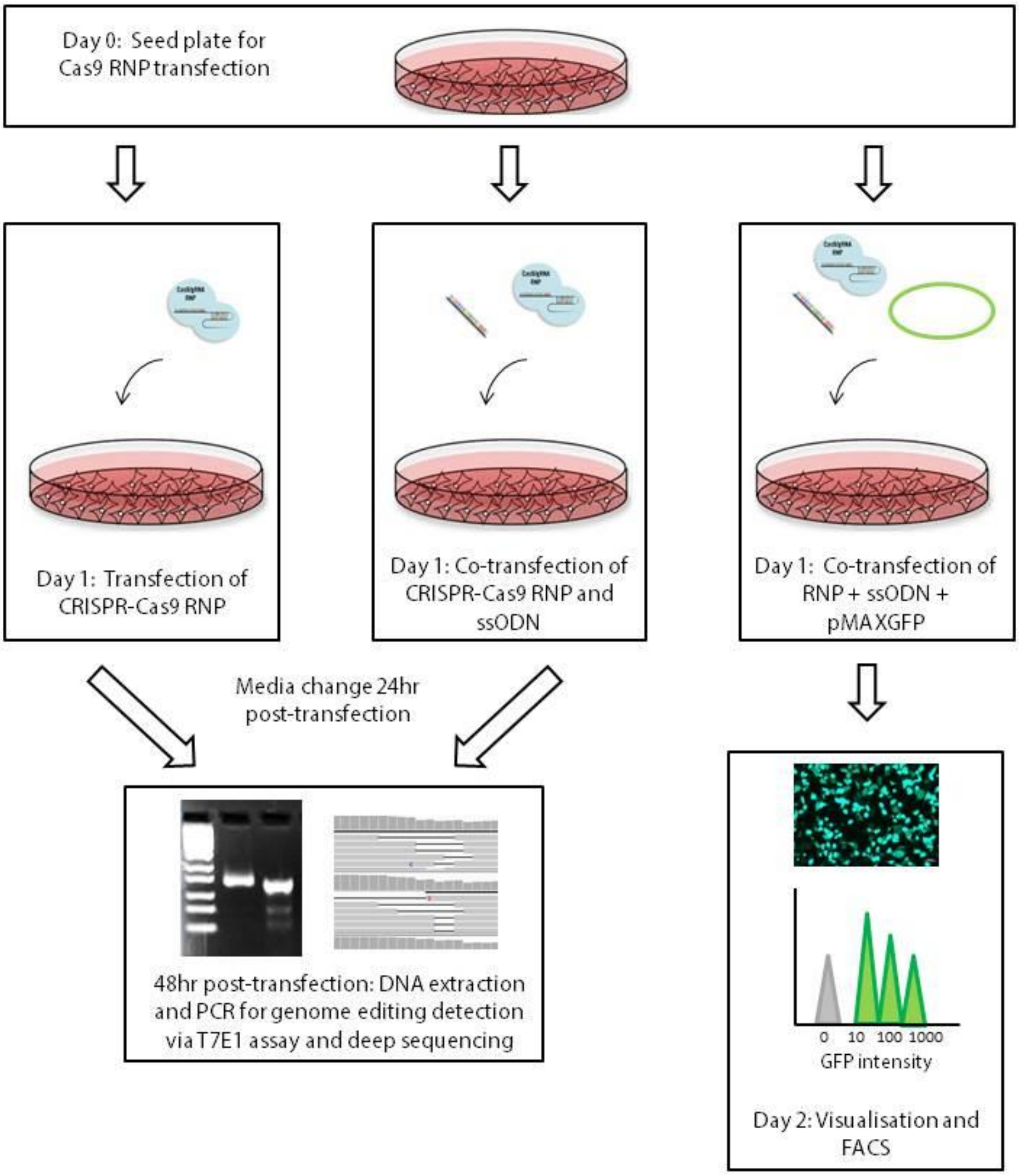

Day 0: Seed plate for Cas9 RNP transfection
Day 1: Transfection of CRISPR-Cas9 RNP
Day 1: Co-transfection of CRISPR-Cas9 RNP and ssODN
Day 1: Co-transfection of RNP + ssODN + pMAXGFP
Media change 24hr post-transfection
48hr post-transfection: DNA extraction and PCR for genome editing detection via T7E1 assay and deep sequencing
0
10
100
1000
GFP intensity
Day 2: Visualisation and FACS

Figure 3.4 Schematic of the RNP-based CRISPR-Cas9 genome editing protocol
Experiments were conducted to investigate the best delivery method for the CRISPR-Cas9 RNP components using the co-transfection of a fluorescent marker (GFP) to enrich for transfected cells and HDR template to test the efficiency of both NHEJ and HDR repair pathways. The efficiency of these delivery methods was determined by deep sequencing of targeted loci.

3.5. Results

3.5.1. Optimisation of transfection reagent conditions for delivery of Cas9 plasmid

The use of Lipofectamine® LTX and Lipofectamine® RNAiMAX as a transfection agent for the delivery of Cas9 plasmids into MAC-T cells was investigated using the recommended protocols for each reagent. For both the mCherry Cas9 plasmid and the co-transfection of pMAXGFP and PX459 plasmids, the transfection efficiency of Lipofectamine® LTX was greater than Lipofectamine® RNAiMAX (Figure 3.5). The transfection efficiency of the co-transfected pMAXGFP and PX459 plasmids was approximately 70% and 60% for Lipofectamine® LTX and Lipofectamine® RNAiMAX, respectively (Figure 3.5A and Figure 3.5B). The transfection efficiency of the mCherry Cas9 plasmid was approximately 30% and 10% with Lipofectamine® LTX and Lipofectamine® RNAiMAX, respectively (Figure 3.5C and Figure 3.5D).

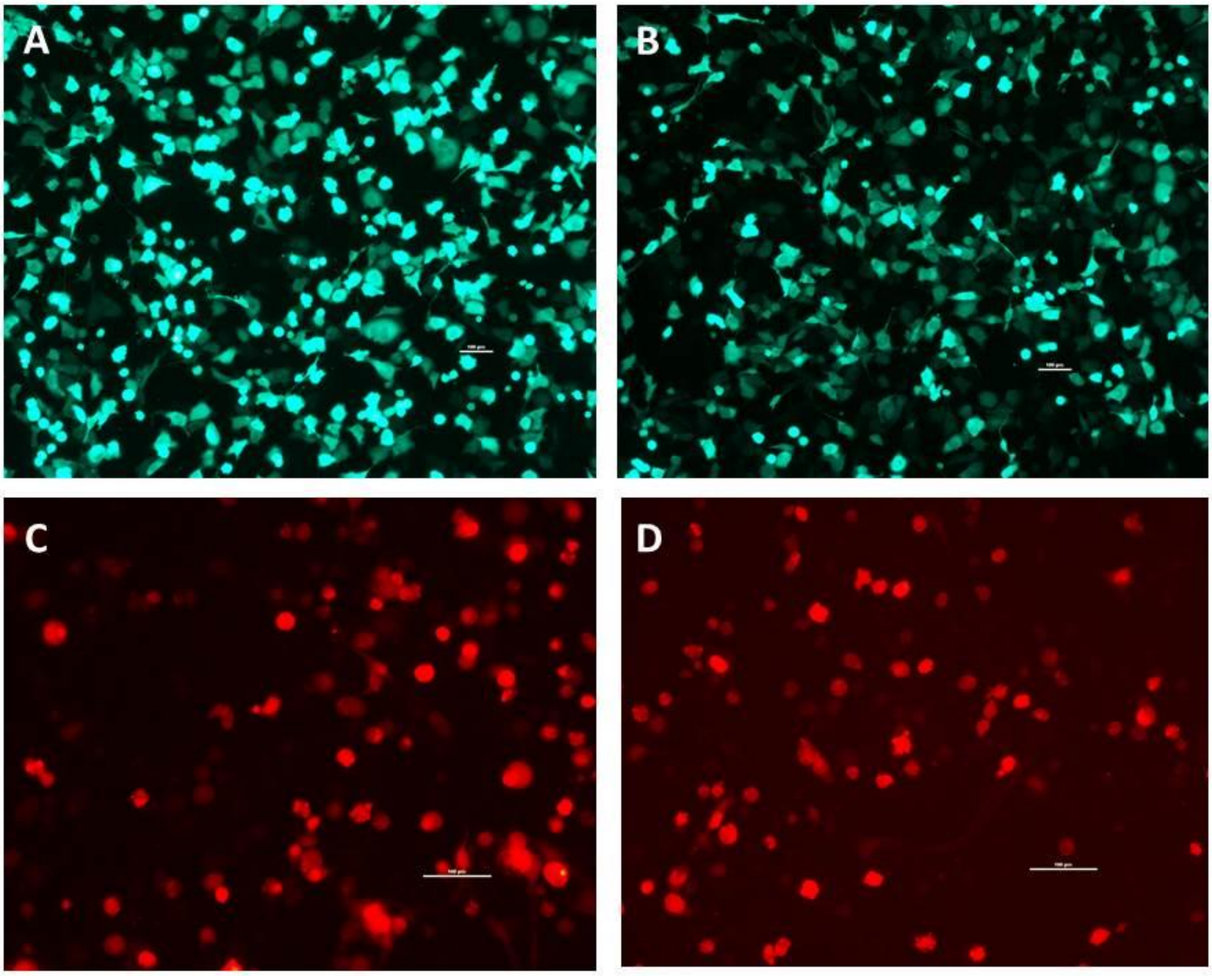


Figure 3.5 Visualisation of transfection efficiency for the co-transfection of PX459 and pMAXGFP and the mCherry Cas9 plasmid

A) Represents the co-transfection of pMAXGFP and PX459 plamids using Lipofectamine® LTX, while B) represents the co-transfection of pMAXGFP and PX459 plasmids (1:1) using Lipofectamine® RNAiMAX. C) Represents the transfection of mCherry Cas9 plasmid using Lipofectamine® LTX, while D) represents the transfection of mCherry Cas9 plasmid using Lipofectamine® RNAiMAX.

3.5.2. Determination of the optimal conditions for Cas9 plasmid-mediated gene editing

To test the efficiency of Cas9 plasmid-based editing efficiency, experiments were conducted using a variety of transfection protocols targeting the chr27:36198117T>TGGC variant in the mammary cell line. The amount of Cas9 expression plasmid was titrated by co-transfecting in a 1:1 ratio and 2:1 ratio with pMAXGFP prior to sorting via FACS. The chr27:36198117T>TGGC gRNA was transfected intro these cells by either forward or reverse transfection. Unsorted cells (transfected with Cas9:pMAXGFP in 1:1 and 2:1 ratio) were also transfected by either a forward or reverse transfection. Unfortunately, not all components of these experiments generated valid results, since pipetting and FACS sorting errors rendered replicates from the '2:1 forward sorted', '2:1 reverse sorted', and '1:1 reverse unsorted' cells unusable.

The sequencing of PCR products amplified from genomic DNA encompassing the chr27:36198117T>TGGC locus revealed the Cas9 plasmid-mediated editing was highly active across a number of these protocols (Figure 3.6). The indel frequency at the seven nucleotides surrounding the gRNA cleavage site was used to determine the optimal Cas9 plasmid-mediated protocol for editing at this locus. Of the transfection protocols tested, enriching for GFP positive cells via FACs sorting had the most profound effect on editing efficiency for both the forward and reverse transfection of the gRNA (Table 3.3; Figure 3.7). The protocol involving the forward or reverse transfection of gRNA into FACS sorted cells resulted in an average indel frequency of 34.35% and 22.25%, respectively (Table 3.3; Figure 3.7). Interestingly, the average indel frequency across the samples involving the forward transfection of the gRNA into GFP positive cells was highly consistent, with all replicates demonstrating an average indel frequency between 31.31% and 36.21% (Table 3.3). In contrast, there was poor replication been the reverse sorted samples, with average indel frequencies of 33.19%, 11.31%, and 8.68% (Table 3.3; Figure 3.7).

Notably, there was no influence on indel frequency at the target locus as the result of increasing the amount of Cas9 plasmid transfected. The cells transfected in a 2:1 and 1:1 ratio of PX459:pMAXGFP had a similar indel efficiency in the sorted and unsorted cells (Table 3.3; Figure 3.7). Surprisingly, the '2:1 reverse unsorted' protocol resulted in an average indel frequency of 16.75% and 18.53%, which was higher than the reverse transfection of the gRNA into sorted cells (Table 3.3; Figure 3.7). Taken together, the results from this experiment support the forward transfection of gRNAs into cells that have been enriched for Cas9 expression through a selectable marker such as GFP fluorescence.

It is important to note that the accuracy of estimating the editing efficiency at the chr27:36198117T>TGGC locus is constrained by the wild-type genotype for this position. As such, the maximum editing efficiency reported for this position (which is also the gRNA cleavage site) is likely being influenced by the wild-type indel and possible allele bias during PCR amplification. As the range of allele frequencies reported in control samples varied from ~25% to ~45% (where theoretically the indel frequency should be 50%), the average of the indel frequency reported for the surrounding nucleotides is used to establish an average indel frequency for this target site.

In addition, the indel frequency at this locus was also measured using the T7E1 assay. The cleavage products and full-length amplicons were visualised by gel electrophoresis (Figure 3.8). However, the inherent insensitivity of this assay in combination with the heterozygous wild-type genotype for chr27:36198117T>TGGC made it difficult to interpret the results and the use of this assay to estimate editing efficiency was not pursued further.

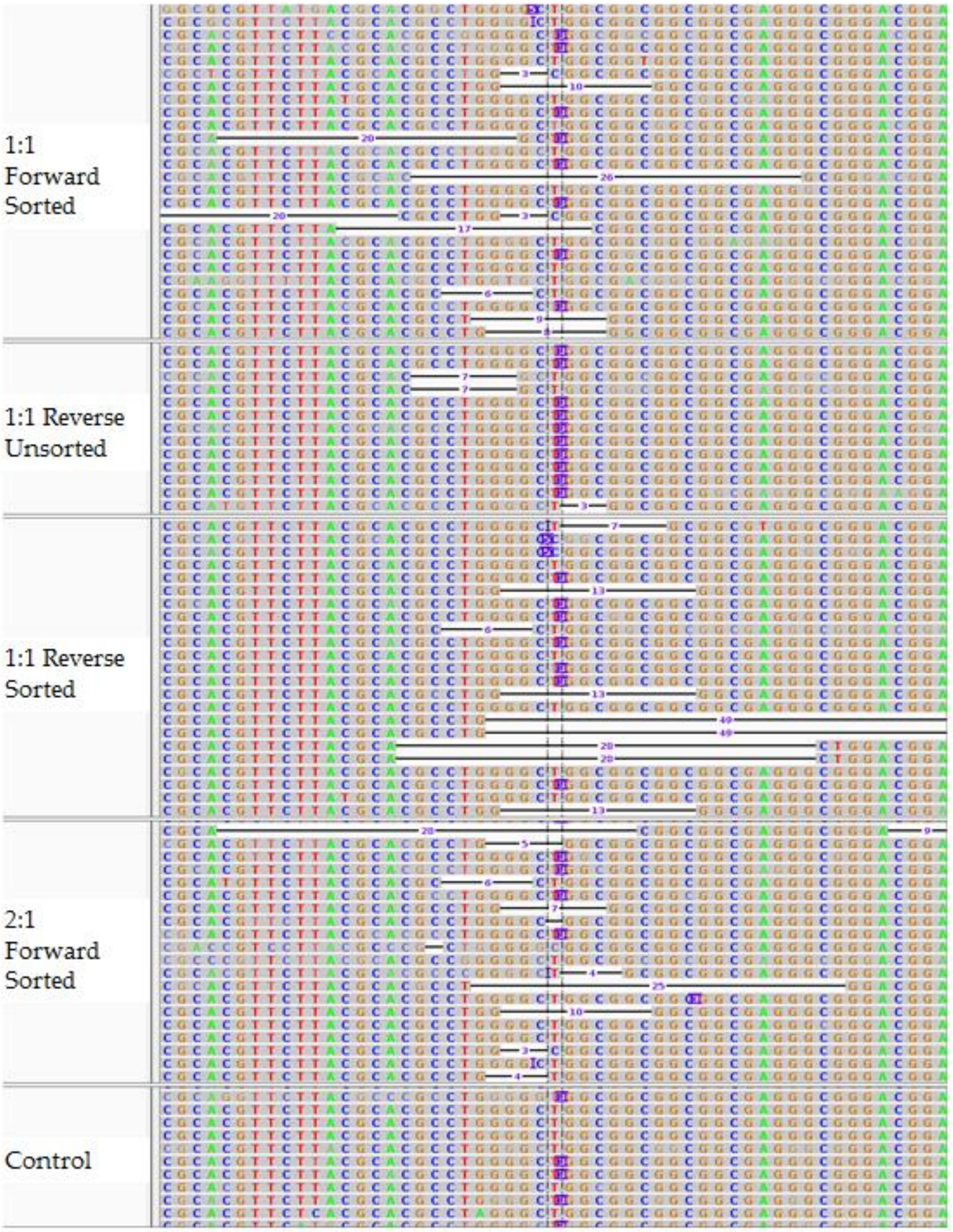


Figure 3.6 Alignments of sequence data demonstrating Cas9 plasmid mediated editing of chr27:36198117 T>TGGC target variant. CRISPR-Cas9 editing can be seen in the four treatments represented in this image in the form of deletions (black lines) which are absent from the control (below). The T>TTGC variant (for which the cell line is heterozygous) is marked in purple in the centre of the image.

Table 3.3 Indel frequency at the 7 bp around the gRNA cleavage site for Cas9 plasmid mediated editing of chr27:36198177 T>TGGC target variant. Cleavage site (gRNA offset = zero) of the Cas9 is chr27:36198177 for which the cell line is heterozygous for the indel as shown in control.

	Max indel frequency gRNA offset (%)							
Transfection Protocol	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**Average**
1:1 Forward Sorted	31.72	33.30	30.41	61.53	20.70	18.51	16.99	35.53
1:1 Forward Sorted	28.48	29.16	25.47	55.97	17.25	16.33	15.23	31.31
1:1 Reverse Sorted	0.03	0.02	0.95	64.36	0.88	0.78	0.83	11.31
1:1 Reverse Sorted	28.51	30.12	26.72	56.94	19.21	19.45	18.19	33.19
1:1 Reverse Unsorted	4.38	5.06	5.01	44.47	3.88	3.79	3.72	11.72
2:1 Forward Sorted	31.01	33.12	29.85	58.19	21.27	21.12	22.72	36.21
2:1 Reverse Sorted	4.06	4.01	0.66	39.34	1.08	0.98	1.92	8.68
2:1 Reverse Unsorted	9.86	10.76	9.66	48.28	7.40	7.26	7.26	16.75
2:1 Reverse Unsorted	11.69	14.97	12.08	60.49	4.05	3.90	4.00	18.53
Control	0.02	0.01	0.63	43.16	1.14	1.05	1.13	7.86

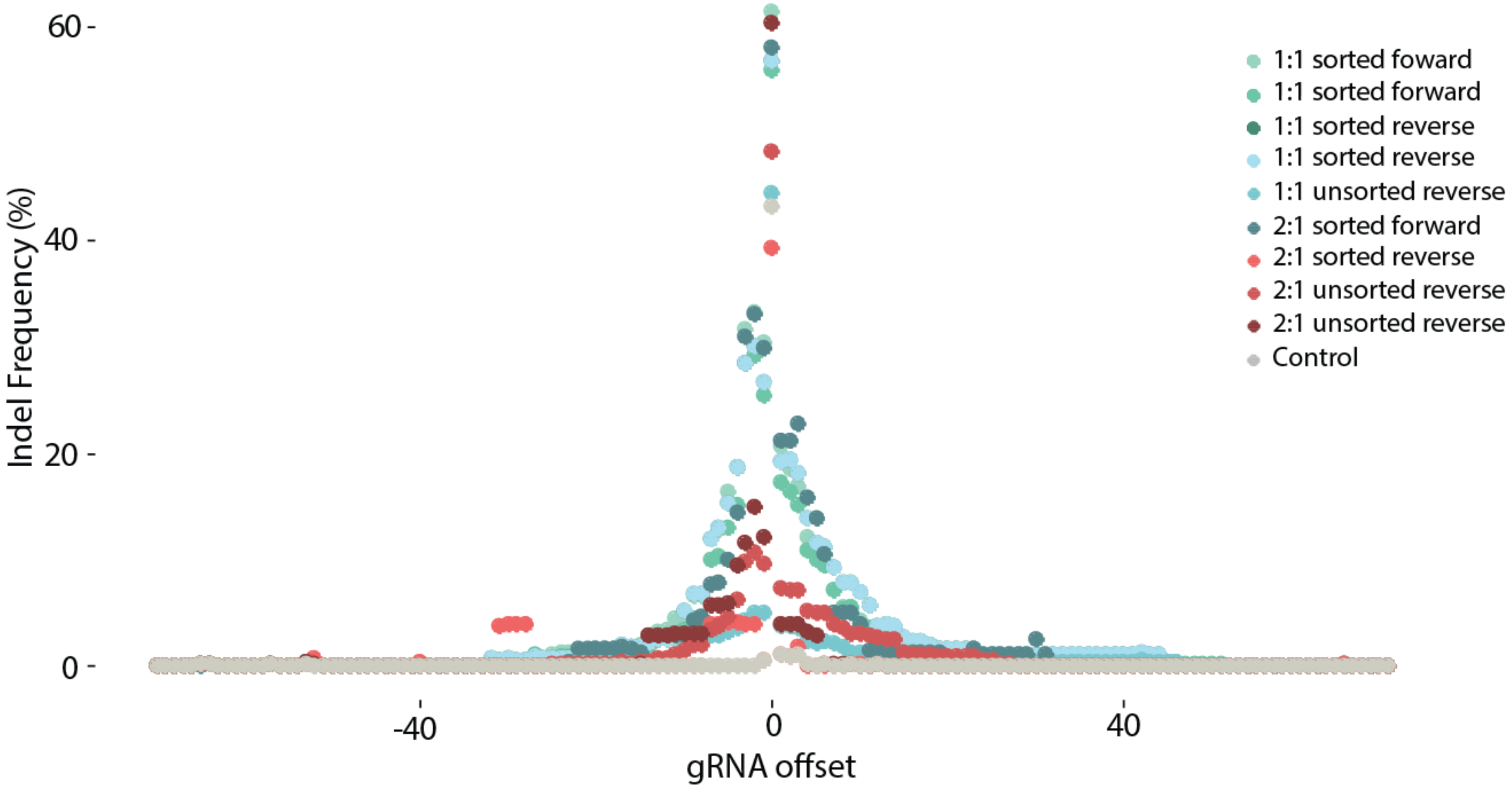


Figure 3.7 Indel frequency at the gRNA cleavage site for Cas9 plasmid-mediated editing of the chr27:36198117 T>TGGC target variant using different transfection protocols for PX459.
Cleavage site of the Cas9 is chr27:36198177 for which the cell line is heterozygous for the indel as shown in control (grey).

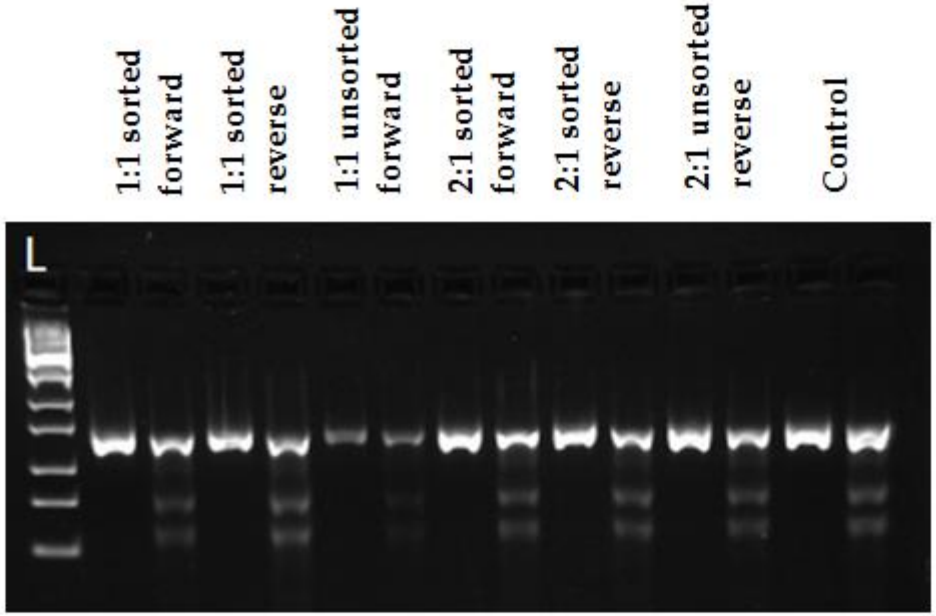


Figure 3.8 Gel electrophoresis of T7E1 digested and undigested PCR products from CRISPR-Cas9 editing of the chr27:36198117T>TGGC target locus

3.5.3. Determining the optimal conditions for Cas9 protein-mediated gene editing

3.5.3.1. Optimisation of transfection conditions for CRISPR-Cas9 RNP complexes

To establish the optimal conditions for the transfection of CRISPR-Cas9 RNP complexes, a number of transfection conditions were tested, using the protocol recommendations of IDT as a guide for method development. In the first optimisation experiment, 0.5 µL, 1 µL and 2 µL Lipofectamine® RNAiMAX was tested in conjunction with 3, 10 and 20 nM RNP complex for the chr27:36198117T>TGCC locus. The RNP complexes were incubated with the different concentrations of Lipofectamine® RNAiMAX, transfected, and cells cultured for 48 hours.

The sequencing of PCR products amplified from genomic DNA encompassing the target loci revealed the RNP complexes were highly active, generating indel frequencies similar to the maximum achieved using the most efficient parameters used for plasmid-mediated editing (see 3.5.2). Specifically, using 2 µL Lipofectamine® RNAiMAX and 20 nM RNP complex resulted in the highest average indel frequency (30.00%) at the seven nucleotides surrounding the gRNA cleavage site. At this concentration of transfection reagent, the indel frequency also appeared to diminish in a dose-response manner across the 10 nM and 3 nM RNP concentrations (15.36% and 9.60%, respectively; Table 3.4; Figure 3.9). Similarly, indel generation was lowest for the samples treated with 0.5 µL Lipofectamine® RNAiMAX. At this concentration of transfection reagent, there appeared to be no relationship between indel frequency and RNP concentration, with the average indel frequency of 9.49%, 6.72% and 8.35% in the 3 nM, 10 nM and 20 nM RNP samples, respectively (Table 3.4; Figure 3.9). The cells treated with 1 µL of Lipofectamine® RNAiMAX also demonstrated fairly stable (and intermediate) editing efficiencies across the RNP concentrations, with an average indel frequency of 13.97%, 13.81% and 20.24% in the 3 nM, 10 nM and 20 nM samples, respectively (Table3.4; Figure 3.9).

Given the observation of increased editing efficiency with increased RNP and transfection reagent concentration, further transfections were conducted using 2 µL

Lipofectamine® RNAiMAX in conjunction with 10 nM, 20 nM, 30 nM, 40 nM and 80 nM RNP complex. Sequencing performed on PCR products amplified from genomic DNA encompassing the target loci revealed the highest average indel frequency at the chr27:36198117 T>TGGC locus was derived from the protocol using 20 nM RNP complex (19.99%; Table 3.5; Figure 3.10). Interestingly, the higher RNP concentrations, 30 nM, 40 nM and 80 nM did not result in increased editing at this locus (14.71%, 12.10% and 12.60%, respectively; Table 3.5; Figure 3.9). The lowest average indel frequency was seen in the 10 nM RNP sample (6.17%, Table 3.5; Figure 3.10)

As a consequence of the high indel frequency in cells transfected with 2 µL Lipofectamine® RNAiMAX and 20 nM RNP complexes (made up of equimolar gRNA and Cas9 protein), these transfection conditions were used to test the efficiency of NHEJ and HDR editing for the 27 other gRNAs presented in Table 3.1 (see 3.6.3.2).

Table 3.4 Indel frequency at the 7 bp around the gRNA cleavage site for Cas9 protein mediated editing of chr27:36198177 T>TGGC target variant using 0.5 – 2 µL Lipofectamine® RNAiMAX and 3 – 20 nM RNP complex

Cleavage site (gRNA offset = zero) is 36198177 for which the cell line is heterozygous for the indel as shown in control.

	Max indel frequency gRNA offset (%)							
Transfection Protocol	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**Average**
0.5 µL Lipofectamine + 3 nM RNP	2.96	2.73	3.15	48.77	2.92	2.82	3.08	9.49
0.5 µL Lipofectamine + 10 nM RNP	0.02	0.01	0.66	42.53	1.25	1.15	1.39	6.72
0.5 µL Lipofectamine + 20 nM RNP	0.99	0.99	1.70	48.29	2.18	2.07	2.19	8.35
1 µL Lipofectamine + 3 nM RNP	7.94	7.94	8.60	54.75	6.22	6.10	6.24	13.97
1 µL Lipofectamine + 10 nM RNP	9.49	9.78	7.63	51.86	6.23	5.76	5.88	13.81
1 µL Lipofectamine + 20 nM RNP	21.77	21.70	17.18	51.45	10.53	9.90	9.15	20.24
2 µL Lipofectamine + 3 nM RNP	3.15	3.25	3.59	47.49	3.38	3.17	3.16	9.60
2 µL Lipofectamine + 10 nM RNP	11.47	12.07	10.86	51.13	7.70	7.35	6.94	15.36
2 µL Lipofectamine + 20 nM RNP	33.38	35.38	30.04	62.80	17.17	16.17	15.07	30.00
Control	0.04	0.02	0.73	41.03	1.31	1.17	1.27	6.34

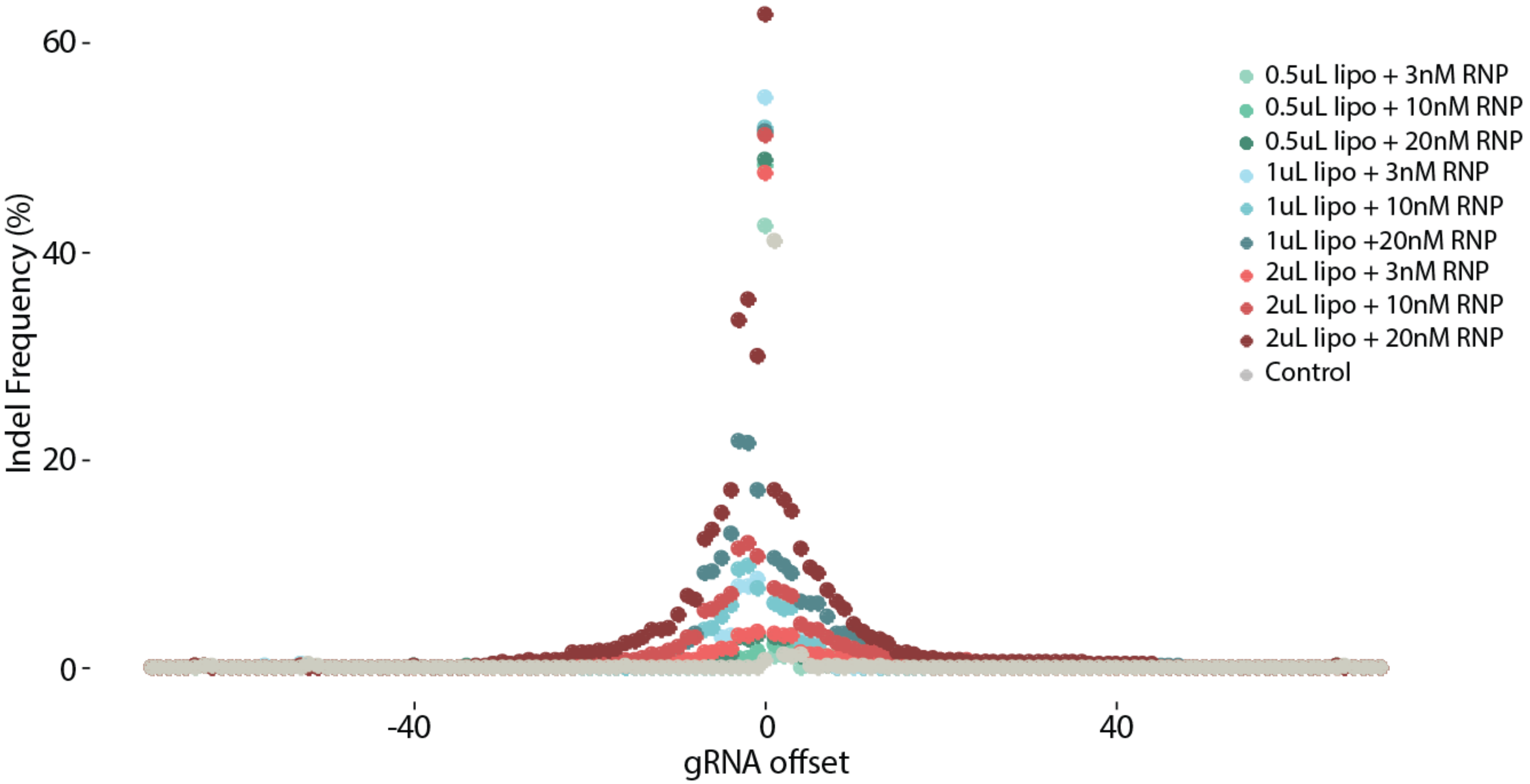


Figure 3.9 Indel frequency for Cas9 protein-mediated editing of the chr27:36198177 T>TGGC target locus using 0.5 – 2 µL transfection reagent and 3 – 20 nM RNP complex

Cleavage site (gRNA offset = zero) is 36198177 for which the cell line is heterozygous for the indel as shown in control (grey). Lipo = Lipofectamine® RNAiMAX

Table 3.5 Indel frequency at the 7 bp around the gRNA cleavage site for Cas9 protein-mediated editing of chr27:36198117 T>TGGC using 10 – 80 nM RNP complex
Cleavage site (gRNA offset = zero) is 36198177 for which the cell line is heterozygous for the indel as shown in control.

	Max indel frequency gRNA offset (%)							
Transfection Protocol	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**Average**
10 nM RNP	0.67	0.67	1.11	34.21	1.53	1.53	3.50	6.17
20 nM RNP	22.52	26.58	20.58	39.97	10.32	10.00	9.93	19.99
30 nM RNP	18.81	19.97	10.35	37.41	4.78	4.91	6.73	14.71
40 nM RNP	11.64	12.54	11.73	31.40	5.80	5.49	6.11	12.10
80 nM RNP	10.96	11.81	9.91	38.44	5.79	5.60	5.70	12.60
Control	0.00	0.00	1.50	24.64	0.85	0.21	2.90	4.30

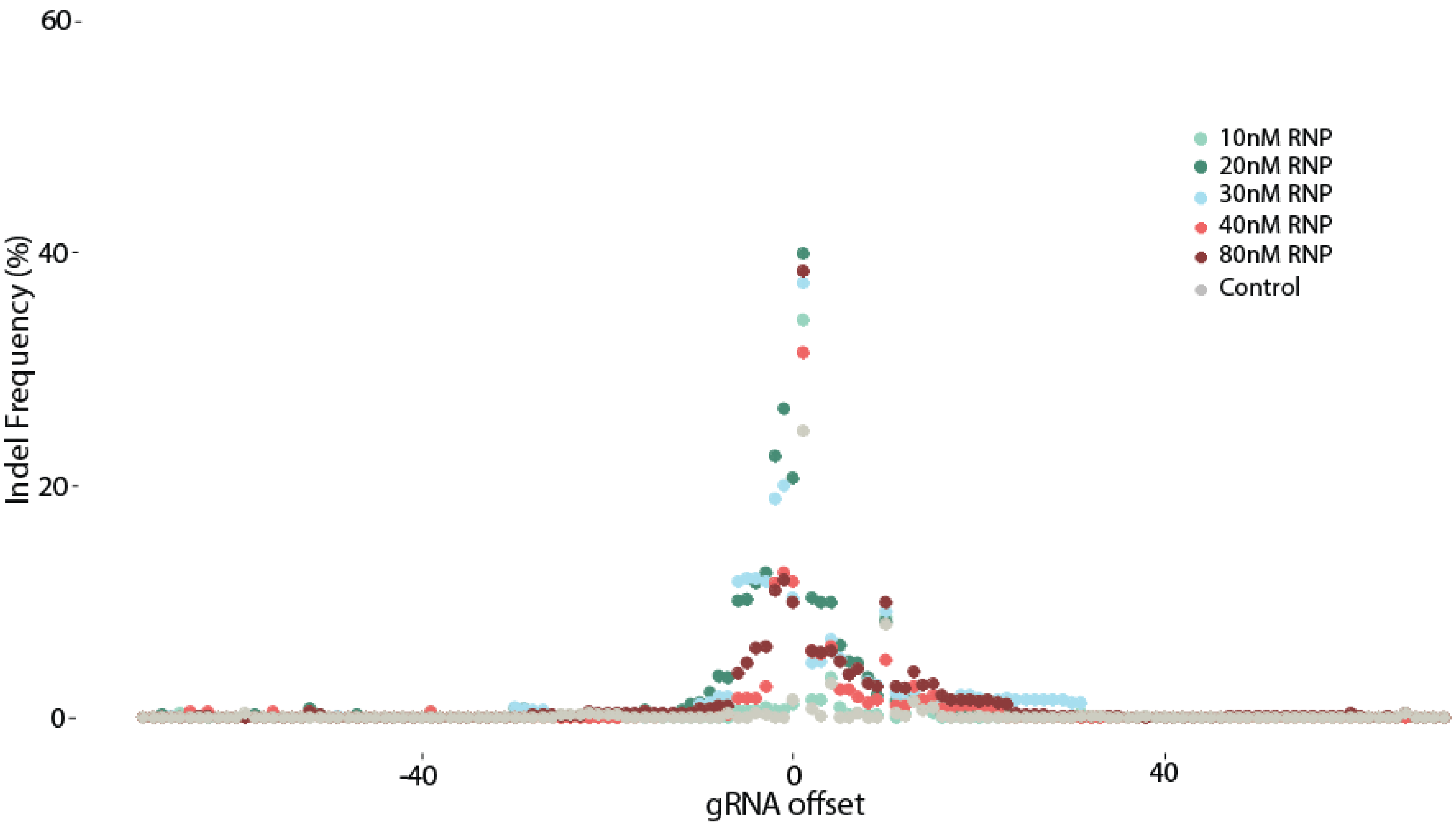


Figure 3.10 Indel frequency for Cas9 protein-mediated editing of the chr27:36198177 T>TGGC target locus using 10 – 80 nM RNP complex. Cleavage site (gRNA offset = zero) is 36198177 for which the cell line is heterozygous for the indel as shown in control (grey).

3.5.3.2. Testing the efficiency of NHEJ and HDR editing for gRNAs targeting 28 loci

To test the efficiency of NHEJ and HDR for the gRNAs in Table 3.1, transfections were conducted using 20 nM RNP complex, 2 μL Lipofectamine® RNAiMAX and 10 nM ssODN for each gRNA. Genomic DNA spanning each of the targets was amplified by PCR and deeply sequenced to quantitatively assess the efficiency of NHEJ and HDR editing at each locus. Of the 28 loci targeted for CRISPR-Cas9 editing, a subset of 20 locations yielded sequence data of a quality sufficient for downstream analysis (Table 3.6).

Of the 20 loci that yielded high quality data, the indel frequency ranged from 1.05% to 57.17%, with a median frequency of 29.25% across all targets (Table 3.6; Figures 3.11-3.13). The activities of RNP complexes targeting chr27:36202188T>A, chr27:36204066T>C, and chr5:93948357C>T were very high, inducing indel frequencies of 57.14%, 51.17 and 48.48%, respectively. However, the mean sequence read depth for the chr27:36204066 T>C locus was less than 20X, so the interpretation of data for this locus needs to be treated with caution.

Activities of the RNP complexes designed to target chr5:93945655T>G, chr5:93946027T>A, chr5:93946548G>C, chr5:93947761C>T, chr27:36200968T>C, chr27:36202636GT>G, chr27:36203904G>C, chr27:36209319T>G, chr27:36211708T>C and chr11:103301781G>A were also substantial, inducing mutations at frequencies between 28.57% and 45.45% (Table 3.6). Conversely, the RNP complexes designed to target chr5:93948646C>G, chr5:93948718G>C, chr5:93954748T>C, chr27:36206783C>A, chr27:36211257GA>T and chr27:36212352G>A only induced indel frequencies between 1.05% and 20% (Table 3.6).

Interestingly, the addition of the HDR template to the transfection mix completely ablated NHEJ at the *MGST1* chromosome 5 target loci (with the exception of chr5:93946027T>A target locus; Figures 3.11-3.13). Conversely, at the *AGPAT6* chromosome 27 location (encompassing 9 target loci), there was NHEJ editing at all sites with the addition of the HDR template to the transfection mix (Figures 3.11-3.13). However, the indel frequency at these loci was much lower with the addition of the HDR template.

Unsurprisingly, the percentage of HDR was much lower than the indel frequency, ranging from 8.33% to 0%, with only two of the 8 loci with available sequence data demonstrating any HDR repair (Table 3.6). These loci were chr5:93945655T>G, chr5:93948646C>G which had 3.88% and 8.33% and HDR, respectively. Notably, the sequence read depth for the chr5:93948646C>G locus was less than 20X so the interpretation of these data should be treated with caution.

Due to the heterozygous background of the cell line for the genetic variants at the *AGPAT6* locus, the percentage of HDR could not be determined for any of these targets. As previously mentioned, the possibility of amplification bias during PCR made it impossible to discern if the allele frequencies reported in the sequence data was due to HDR, since the anticipated magnitude of HDR events and subtleties in the representation of alleles would be likely similar. Additionally, CRISPR-Cas9 editing efficiency could not be established for the following eight targets as these samples failed sequencing: chr5:93944937T>C, chr5:93945738T>C, chr5:93947989T>A, chr5:93948804T>C, chr5:93949810G>A, chr27:36200888T>C, chr27:36204680CAG>ATC, chr14:1802265GC>AA.

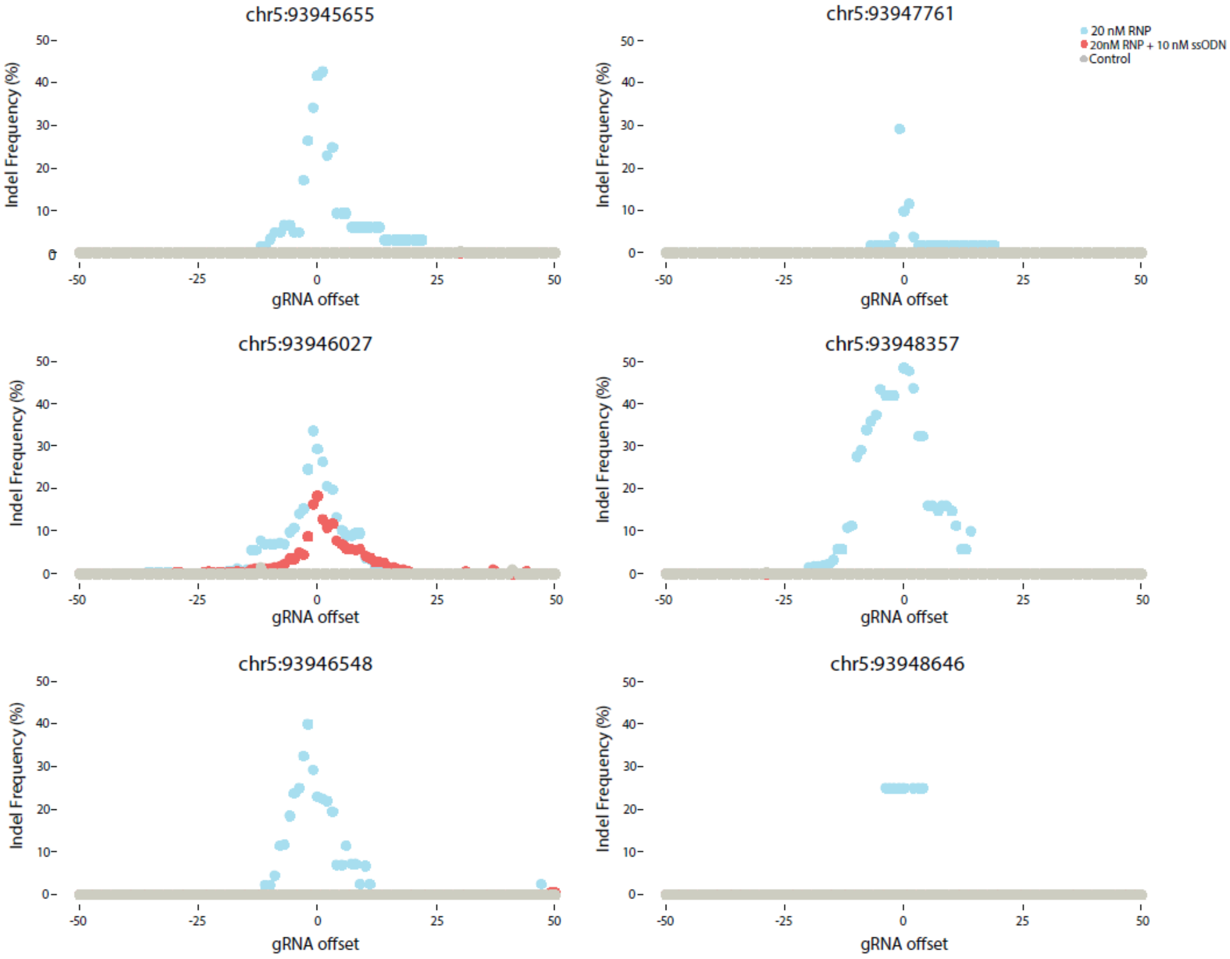


Figure 3.11 Indel frequency at six chromosome 5 target loci; chr5:93945655, chr5:93947761, chr5:939460257, chr5:93948357, chr5:93946548 and chr5:93948646

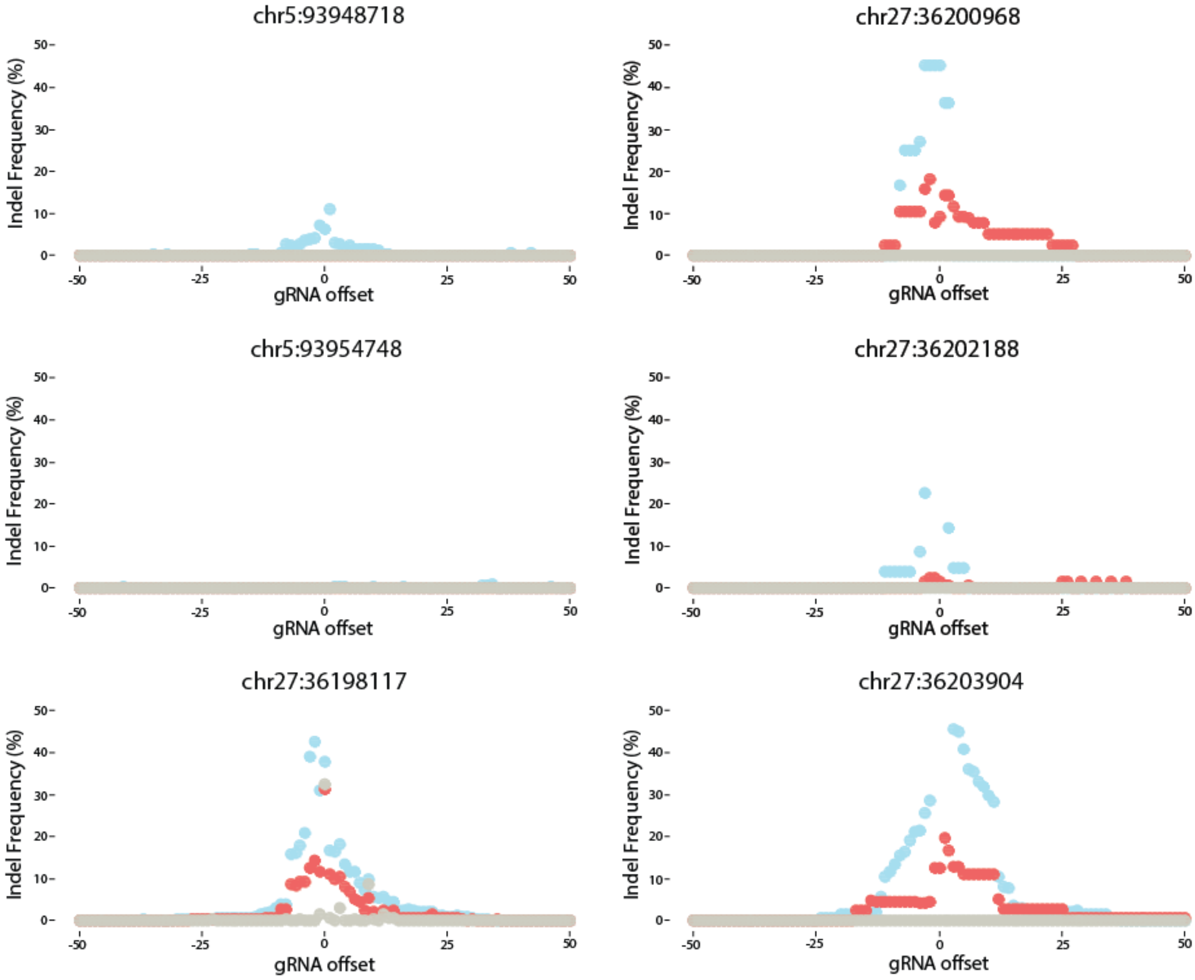


Figure 3.12 Indel frequency at two chromosome 5 target loci; chr5:93948718 and chr5:93954748 and four chromosome 27 target loci: chr27:36200968, chr27:36202188, chr27:36198177 and chr27:36203904

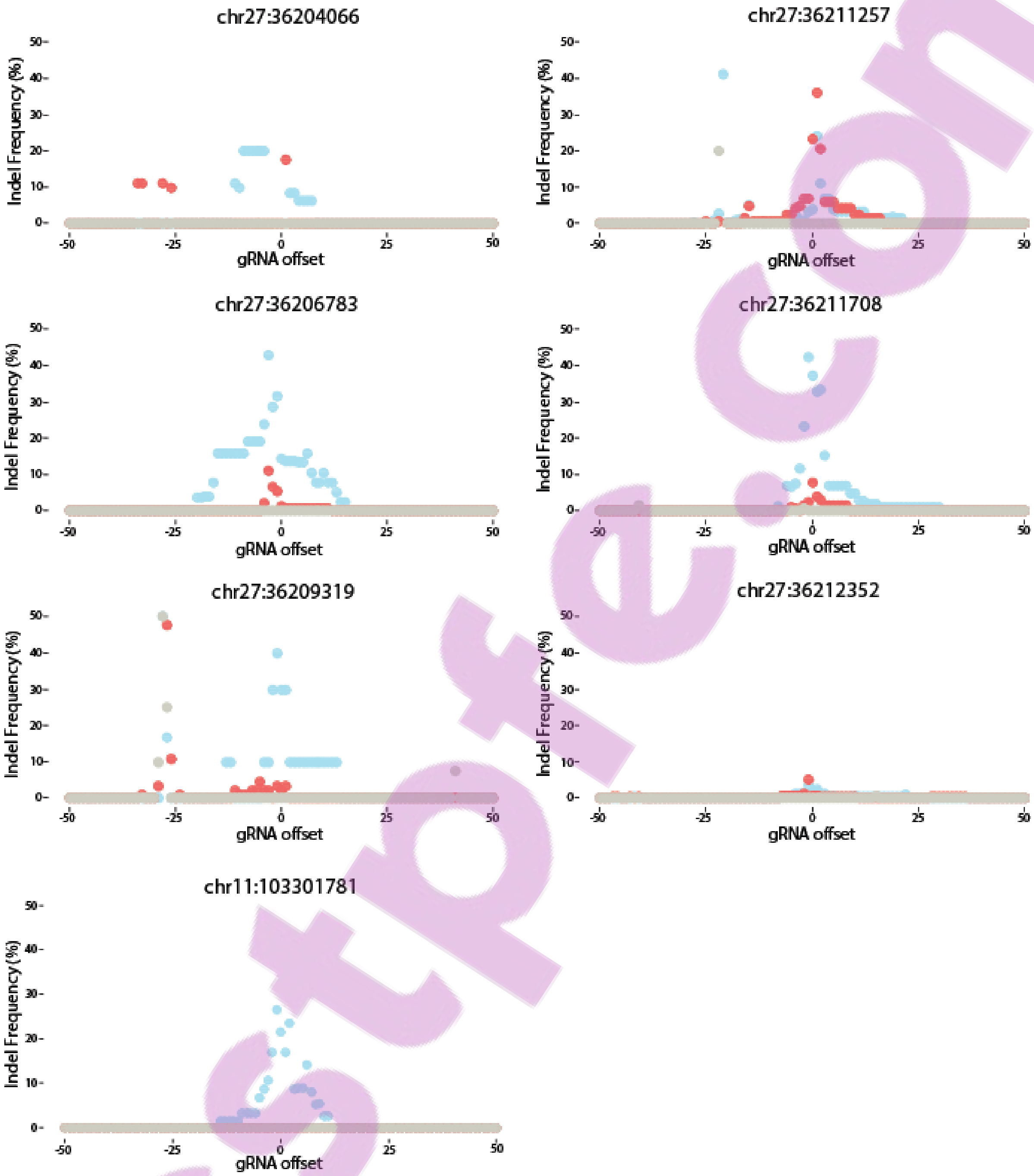


Figure 3.13 Indel frequency at six chromosome 27 target loci; chr27:36204066, chr27:36211257, chr27:36206783, chr27:3621178, chr27:36209319 and chr27:36212352, and one chromosome 11 target locus: chr11:103301781

Table 3.6 NHEJ and HDR editing efficiency of gRNAs targeting 28 loci in a mammary cell line

	Chr	bp	gRNA	Target variant	Indel frequency	HDR efficiency
MGST1	5	93944937	CTTGGGTTCTTCTCCCAGTG	T/C	N/A	N/A
	5	93945655	AAGATTCTCATAGAATCAGA	T/G	42.86%	3.88%
	5	93945738	ATGAGAAGATACAATAAATC	T/C	N/A	N/A
	5	93946027	TTATCTTGCACTGAGAAATG	T/A	33.53%	0%
	5	93946548	GTGCACTGTGAAGTCGGAGA	G/C	29.27%	0%
	5	93947761	TTTATTAACCTCATGTTGCA	C/T	29.23%	N/A
	5	93947989	GTAAGTGCTAGGTAAGTATT	T/A	N/A	N/A
	5	93948357	GGTGGGGGTGGGATTCTAGG	C/T	48.48%	0%
	5	93948646	AAAGAGAAAAGACAGTTCAG	C/G	20% *	8.33%*
	5	93948718	CTTCAGGGCCCAGGTGTTCG	G/C	11.03%	0%
	5	93948804	TTTTTCTGAGGGTTTGAGAG	T/C	N/A	N/A
	5	93949810	TTGGCTTGAGAATTCAAAGT	G/A	N/A	N/A
	5	93954748	TAATCTTACAAAGATTATTG	T/C	1.05%	0%
AGPAT6	27	36198117	TTACGCACGCCTGGGGCTGG	TGGC/T	42.64%	ND
	27	36200888	TGTGCTGGAGAATATGGGCC	T/C	N/A	ND
	27	36200968	TTACGTCTTCCTGTATCATT	T/C	45%*	ND
	27	36202188	TGAGCTGTAAAAACAGACAC	T/A	57.14%	ND
	27	36202636	TGTGCCGTCAGGGAAGTTTG	GT/G	37.10%	ND
	27	36203904	TGTAAGAAACTTGCTTGAGT	G/C	45.45%*	ND

	27	36204066	CCTGGGCTCTATTTTGCTCT	T/C	57.17%*	ND
	27	36204680	TAACAGACTGGGCTTCGCAG	CAG/ATC	N/A	ND
	27	36206783	AGACCACCTTCCCTCCCGAA	C/A	15.78%	ND
	27	36209319	AAAGTGGCCAGAAAGGCTGG	T/G	28.57% *	ND
	27	36211257	GCACACTCCAAGGAGAAGAT	GA/T	11.11%	ND
	27	36211708	AAACCTGGATGAAACGCCTG	T/C	42.42%	ND
	27	36212352	GCTCTTGGGCAGGAGATACA	G/A	3.52%	ND
DGAT1	14	1802265	CGCTTGCTCGTAGCTTTGGC	GC/AA	N/A	N/A
LGB	11	103301781	ATTGTCACCCAGACCATGAA	G/A	28.30%	0%

* Low sequence depth (≥20X), N/A no sequence data available, ND not done.

3.5.3.3. Optimising HDR template concentration

To investigate the influence of ssODN concentration on the frequency of NHEJ and HDR, transfections were conducted with 0 nM, 3.33 nM, 10 nM and 20 nM ssODN, in conjunction with 20 nM RNP complex for the chr5:93946027T>A locus. The HDR template was mixed with the RNP complex and cell cultures were incubated for 48 hours to permit editing. Genomic DNA was then extracted and PCR amplicons spanning the target were generated and subjected to deep sequencing.

The sequencing of the target locus revealed 29.34%, 31.68%, 18.30% and 21.49% indel frequency at the gRNA cleavage site in the 0 nM, 3.33 nM, 10 nM and 20 nM ssODN samples, respectively (Figure 3.14; Table 3.7). The sample treated with 20 nM RNP and no ssODN had the highest editing efficiency as determined by indel frequency at the seven nucleotides surrounding the gRNA cleavage site (24.22%; Table 3.7; Figure 3.14 and 3.15). Interestingly, there was a decrease in the indel frequency at this locus associated with increased ssODN concentration, with the biggest drop in NHEJ seen between 3.33 nM and 10 nM ssODN samples (22.59% and 11.88% indel frequency, respectively). The sample treated with 20 nM RNP and 10 nM ssODN had the lowest indel frequency of 9.10% (Table 3.17; Figures 3.14 and 3.15).

Notably, there was 0.39%, 2.62%, 5.66 and 5.67% HDR in 0 nM, 3.33 nM, 10 nM and 20 nM ssODN samples, respectively, as determined by the frequency of the A nucleotide at position chr5:93946027 (Figure 3.15; Table 3.7). There was no HDR detected in the control sample. The ssODN for the chr5:93946027 locus also contained a second variant (chr5:93945991G>A), located 36 bp from the chr5:939456027 target variant. The majority of sequence reads that had incorporated an A nucleotide at position chr5:939456027 also had an A nucleotide at position chr5:93945991, providing confirmation that the frequency of HDR reported in the current study represents genuine incorporation of synthetic repair template at this locus (Figure 3.16). As not all sequencing reads contained both alleles, the discrepancies in the frequency of their relative incorporation may be due to the partial recombination of the HDR template, and/or PCR and sequencing error.

The indel frequency observed at the chr5:93946027 position (which is 12 bp from the gRNA cleavage site) was 7.69%, 5.28%, 1.40%, and 1.91% in the 0 nM, 3.33 nM, 10 nM and 20 nM ssODN samples, respectively (Figure 3.15; Table 3.7). This is compared to the control sample which had 0% indel frequency at the nucleotides immediately surrounding the cleavage site and 1.64% indel frequency at chr5:93946027. The reported HDR in the control and RNP-only samples is likely due to PCR error given that a non-proof reading enzyme was used to maximise robustness of amplification and the A>T variant sits in the middle of a run of seven T's (TTTTATT). Additionally, the reported indels at this position in the sample control is likely a reflection of the close proximity of this position to the end of the sequencing read (18 bp from the end of the read; Figure 3.16).

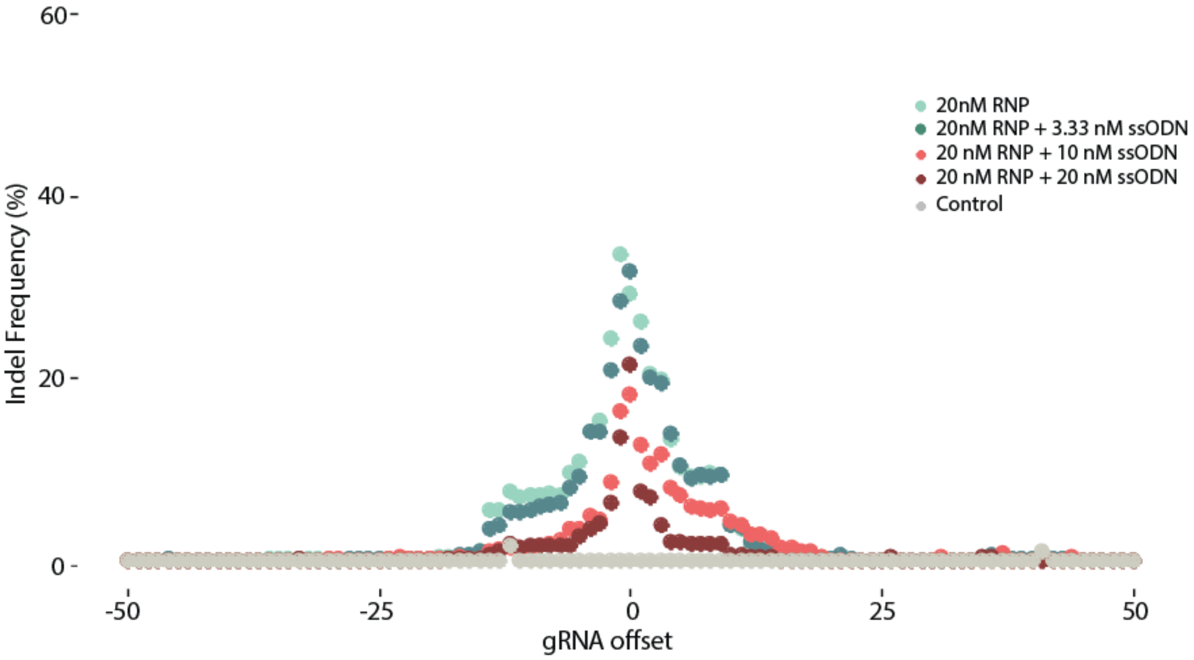


Figure 3.14 Indel frequency for Cas9 protein-mediated editing of the chr5:93946027T>A locus using 0 – 20 nM ssODN with 20 nM RNP complex.

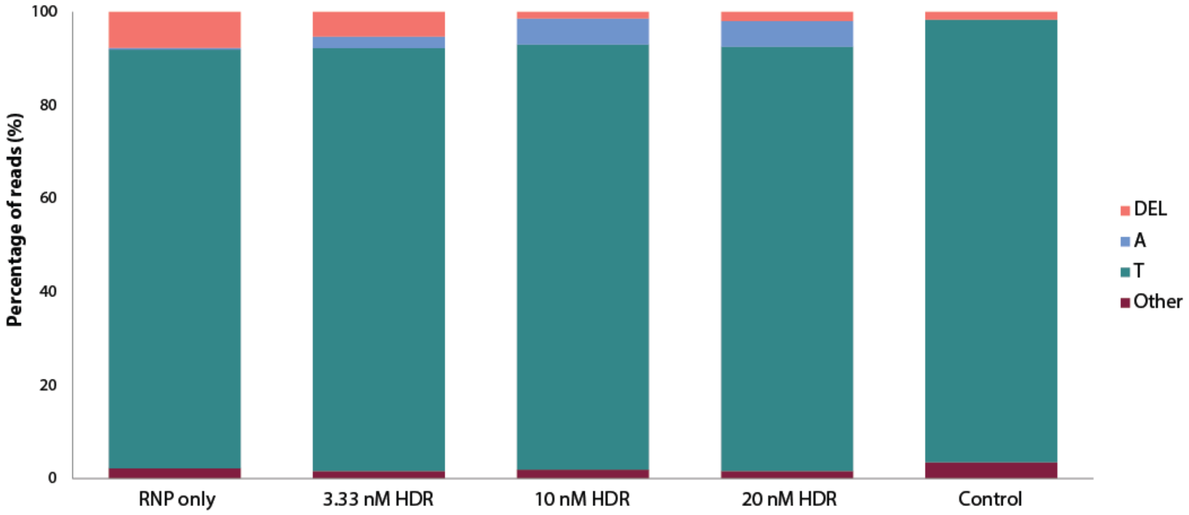


Figure 3.15 The influence of increasing ssODN HDR template concentration on the frequency of CRISPR-Cas9 mediated NHEJ and HDR editing at the chr5:93946027T>A locus.

Table 3.7 Indel frequency at the chr5:93946027T>A target locus for the 7 bp surrounding the Cas9 cleavage site.
The frequency of NHEJ and HDR at the target variant site (which is 12 bp from the cleavage site) is also presented.

	Max indel frequency gRNA offset (%)								**chr5:93946027**	
Transfection Protocol	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**Average**	**Indel (%)**	**HDR (%)**
20 nM RNP	15.35	24.49	33.53	29.34	26.34	20.57	19.88	24.22	7.69	0.36
20 nM RNP + 3.33 nM ssODN	14.09	20.88	28.50	31.68	23.48	20.08	19.42	22.59	5.28	2.48
20 nM RNP + 10 nM ssODN	4.49	8.73	16.33	18.30	12.82	10.73	11.76	11.88	1.40	5.58
20 nM RNP + 20 nM ssODN	4.06	6.28	13.49	21.49	7.53	6.95	3.85	9.10	1.91	5.67
Control	0	0	0	0	0	0	0	0	1.64	0.17

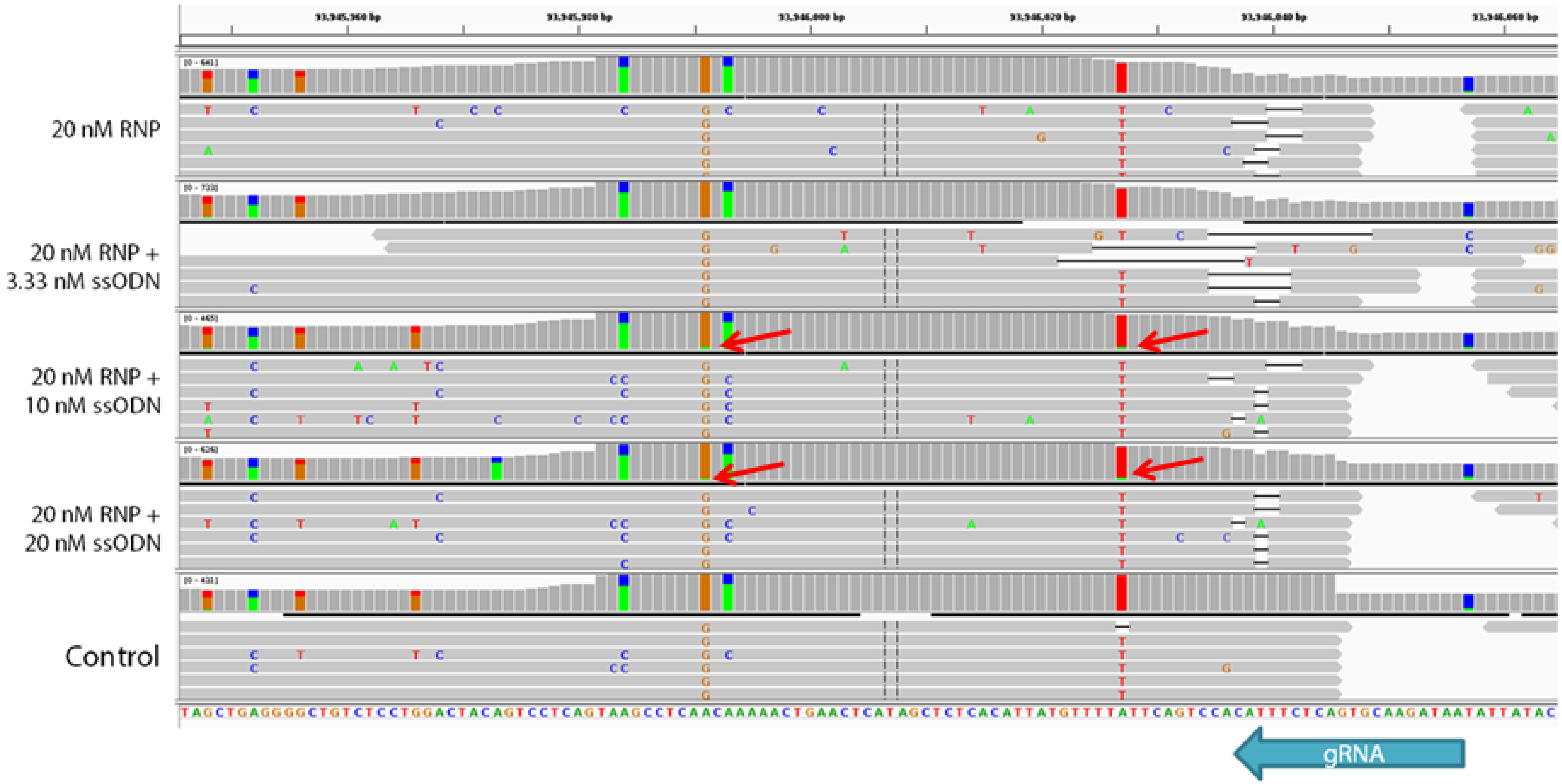

20 nM RNP
20 nM RNP + 3.33 nM ssODN
20 nM RNP + 10 nM ssODN
20 nM RNP + 20 nM ssODN
Control
gRNA
TAGCTGAGGGGCTGTCTCCTGGACTACAGTCCTCAGTAAGCCTCAACAAAAACTGAACTCATAGCTCTCACATTATGTTTTATTCAGTCCACATTTCTCAGTGCAAGATAATATTATAC

Figure 3.16 Aligned bam files illustrating the NHEJ and HDR-mediated introduction of two variants at the chr5:93946027T>A locus
Indel mutations are present in the four treatment samples centred on the gRNA cleavage site (3 bp 5′ of PAM site) and largely absent in the control (bottom). The HDR mediated introduction of an A nucleotide at both position chr5:93946027 and chr5:93945991 in the 20 nM RNP and 10 nM ssODN and 20 nM of both RNP and ssODN samples (middle) can be seen by the small amount of green as indicated by the four red arrows.

3.5.3.4. Targeting more than one locus in a single transfection

To test if more than one locus could be targeted in a single transfection, four gRNAs were multiplexed in the same transfection mix. Transfections were conducted with a final concentration of 20 nM, 40 nM and 80 nM RNP complex, which represented 5 nM, 10 nM and 20 nM of each of the four gRNAs. Two combinations of gRNAs were used targeting different variants for the *AGPAT6* and *MGST1* locus, and the two targets at the *DGAT1* and *LGB* loci. These 4-plex multiplexes were denoted Multiplex A (chr27:36198117, chr11:103301781, chr5:93946027 and chr14:1802265), and Multiplex B (chr27:36211257, chr11:103301781, chr5:93946548 and chr14:1802265).

The sequencing of the target intervals revealed almost no editing at the loci in the 20 nM and 40 nM multiplex samples for the two combinations of targets Table 3.8 and 3.9; Figure 3.17 and 3.18). Comparatively, the gRNAs targeting the chr27:36198117, chr11:103301781, chr5:93946027 loci in Multiplex A and chr11:103301781 and chr5:93946548 in Multiplex B were all active in the 80 nM multiplex sample (17.83% and 18.8%, respectively Table 3.8 and 3.9; Figure 3.17 and 3.18). Interestingly, this corresponds to 20 nM of each RNP complex, which is the amount of RNP which was previously shown to be the most active (3.5.3.1).

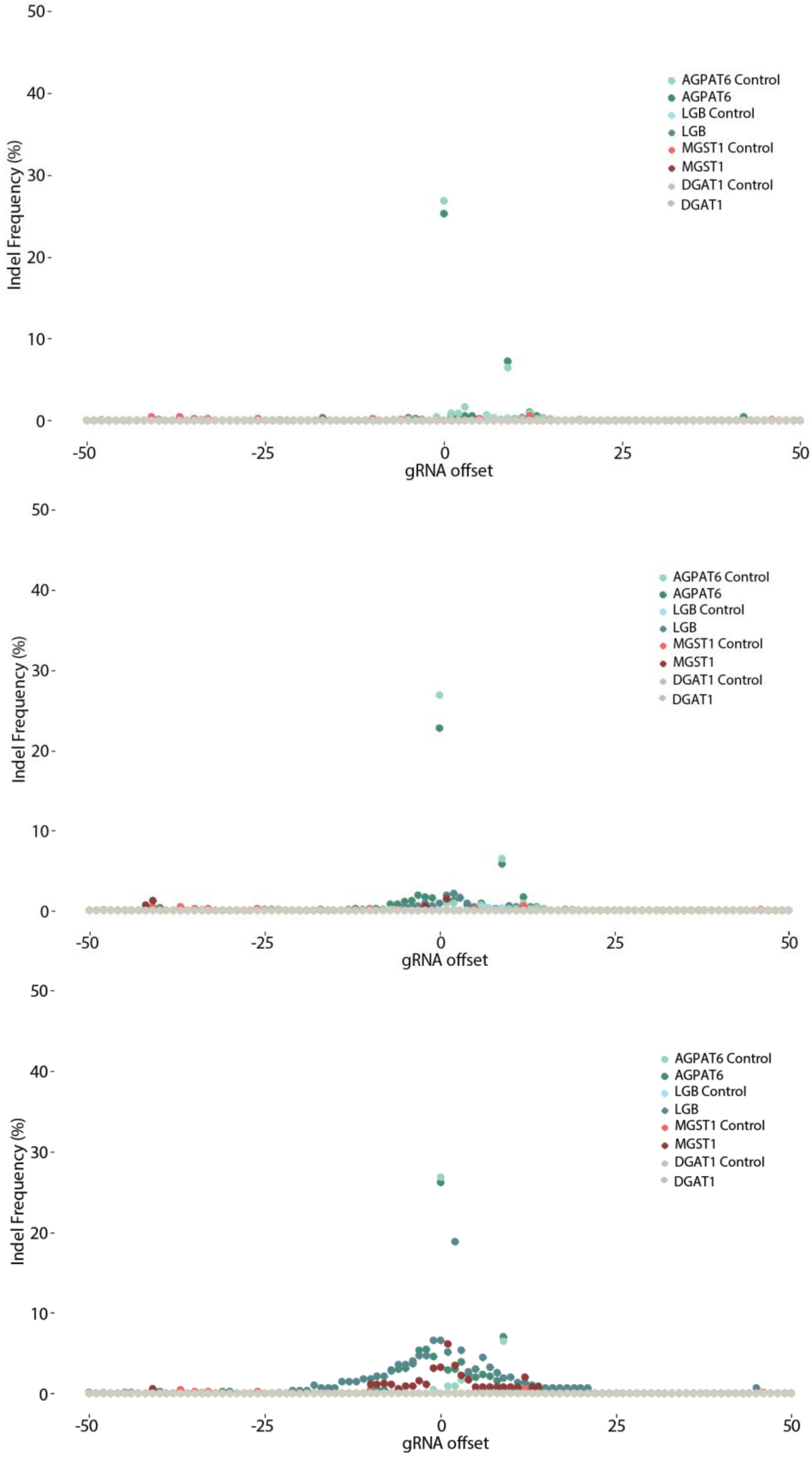


Figure 3.17 Indel frequency for 'Multiplex A'
The four variants targeted were chr27:36198117, chr11:103301781, chr5:93946027 and chr14:1802265 at the *AGPAT6*, *LGB*, *MGST1* and *DGAT1* loci, respectively.

Table 3.8 Indel frequency within 7 bp of the cleavage site for 'Multiplex A'

The four variants targeted were chr27:36198117, chr11:103301781, chr5:93946027 and chr14:1802265 at the *AGPAT6*, *LGB*, *MGST1* and *DGAT1* loci, respectively.

		Max indel frequency gRNA offset (%)								
Multiplex	**Target**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**Average**	**Total**
20nM Multiplex	chr27:36198117	0.62	0.45	0.51	25.26	0.23	0.00	0.12	3.88	
	chr11:103301781	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	chr5:93946027	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	chr14:1802265	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.88
40nM Multiplex	chr27:36198117	1.61	1.21	1.39	22.71	1.60	1.74	1.92	4.60	
	chr11:103301781	1.57	2.18	1.88	0.93	0.62	0.93	0.62	1.25	
	chr5:93946027	0.00	0.00	1.52	0.00	0.00	0.61	0.00	0.30	
	chr14:1802265	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.15
80nM Multiplex	chr27:36198117	3.97	3.06	2.91	26.14	4.58	5.54	5.40	7.37	
	chr11:103301781	5.36	18.89	5.17	6.64	6.62	4.74	4.74	7.45	
	chr5:93946027	2.31	3.51	6.15	3.22	3.17	1.17	1.55	3.01	
	chr14:1802265	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	17.83
Control	chr27:36198117	1.76	0.89	0.94	26.87	0.52	0.00	0.06	4.43	
	chr11:103301781	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.02	
	chr5:93946027	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	chr14:1802265	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.45

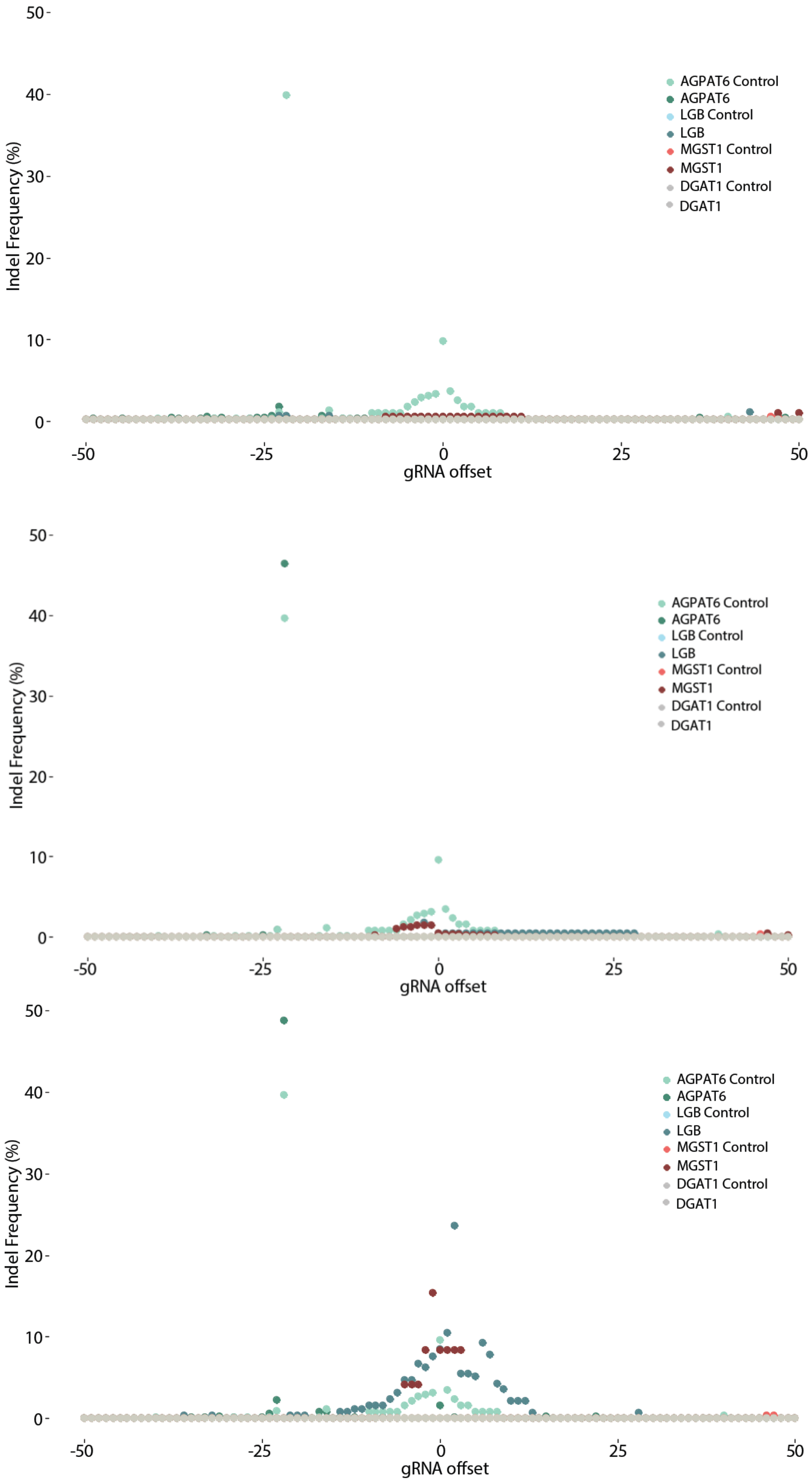


Figure 3.18 Indel frequency for 'Multiplex B'

The four variants targeted were chr27:36211257, chr11:103301781, chr5:93946548 and chr14:1802265 at the *AGPAT6*, *LGB*, *MGST1* and *DGAT1* loci, respectively.

Table 3.9 Indel frequency within 7 bp of the cleavage site for 'Multiplex B'

The four variants targeted were chr27:36211257, chr11:103301781, chr5:93946548 and chr14:1802265 at the *AGPAT6*, *LGB*, *MGST1* and *DGAT1* loci, respectively.

		Max indel frequency gRNA offset (%)								
Multiplex	**Target**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**Average**	**Total**
20nM Multiplex	chr27:36211257	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.02	
	chr11:103301781	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	chr5:93946548	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39	
	chr14:1802265	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.41
40nM Multiplex	chr27:36211257	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	chr11:103301781	0.00	1.82	0.00	0.46	0.46	0.46	0.46	0.52	
	chr5:93946548	1.50	1.50	1.49	0.43	0.42	0.42	0.42	0.88	
	chr14:1802265	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.4
80nM Multiplex	chr27:36211257	0.00	0.00	0.00	1.61	0.00	0.12	0.00	0.25	
	chr11:103301781	6.72	6.30	7.63	8.47	10.44	23.61	5.46	9.80	
	chr5:93946548	4.17	8.33	15.38	8.33	8.33	8.33	8.33	8.75	
	chr14:1802265	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.8
Control	chr27:36211257	2.73	2.94	3.13	9.64	3.52	2.33	1.56	3.69	
	chr11:103301781	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	chr5:93946548	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
	chr14:1802265	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.69

3.5.3.5. Co-transfecting with GFP plasmid to enrich for RNP transfected cells

Since the overall aim of CRISPR-Cas9 optimisation experiments was to obtain the highest efficiencies possible and permit the generation of clonal CRISPR-Cas9 edited cell lines, experiments were also conducted to try and increase the efficiency of RNP-mediated edits. To this end, having observed the significant enrichment of Cas9 plasmid-based editing efficiency when selecting for GFP positive cells via FACS cell sorting, I wondered if RNP-based HDR editing efficiency could similarly be enhanced. The first step was to establish if CRISPR-Cas9 RNP complexes could be co-transfected with both an HDR template and pMAXGFP plasmid. To test if this was possible, six transfection mixes were transfected, containing: [chr27:36198117 RNP, long form of ssODN, and pMAXGFP], [chr27:36198117 RNP, short form of ssODN, and pMAXGFP], [chr27:36211257 RNP, long form of ssODN, and pMAXGFP], [chr27:36211257 RNP, short form of ssODN, and pMAXGFP], [chr27:36212352 RNP, ssODN and pMAXGFP], and pMAXGFP only.

After 24 hours, the cells were visualised to assess transfection based on GFP fluorescence (Figure 3.19). While the transfection efficiency did not appear to be as high as previous experiments (see 3.5.1 and Figure 3.6), the addition of the RNP complex and ssODN had minimal influence on the transfection efficiency (Figure 3.19F). Importantly, the yield of cells from this transfection should be sufficient for input for FACS sorting to derive clonal CRISPR-Cas9 edited cell lines. Further directions for this project will be discussed in section 3.6.5.

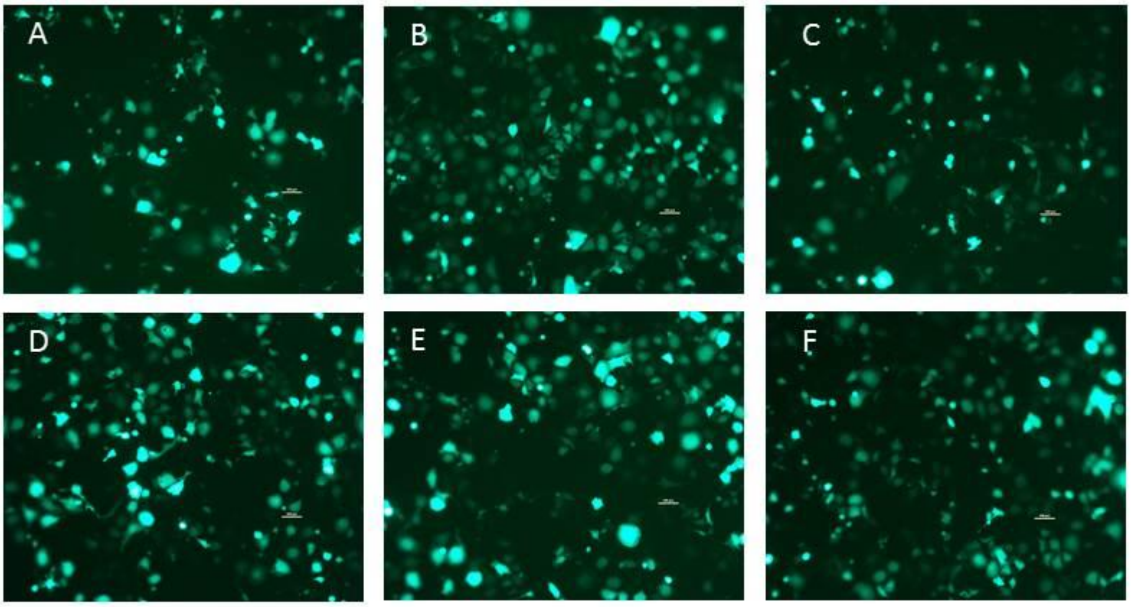


Figure 3.19 Visualisation of transfection efficiency when co-transfecting RNP complex, ssODN and pMAXGFP plasmid

A) Represents the co-transfection of chr27:36198117 RNP, long form of ssODN and pMAXGFP plasmid. B) Represents the co-transfection of chr27:36198117 RNP, short form of ssODN and pMAXGFP plasmid. C) the co-transfection of chr27:36211257 RNP, long form of ssODN and pMAXGFP plasmid. D) the co-transfection of chr27:36211257 RNP, short form of ssODN and pMAXGFP plasmid. E) the co-transfection of chr27:36212352 RNP, ssODN and pMAXGFP plasmid. F) Represents the transfection of only pMAXGFP plasmid.

3.6. Discussion

The motivation for the research described in this chapter was to generate robust gene editing protocols to enable eventual characterisation of candidate causative variants identified through GWAS of bovine milk composition. Such methodology has the potential to identify causative variants within clusters of identically associated candidates in perfect LD, enabling the effect of individual alleles to be assessed outside of their native haplotypic context. This approach serves a pressing need in the identification of causative variants for expression-based effects, where functional prediction of variant effects, or further statistical delineation, is otherwise difficult or impossible. A further advantage of such an approach is that, due to their status as eQTL candidates, the effects of these variants could be assessed using a single, common 'phenotype', namely differential expression of the implicated gene.

To this end, the aims of the experiments presented in this chapter were to develop protocols for the genomic editing of bovine mammary cells, facilitated by CRISPR-Cas9. Overall, I demonstrated successful targeting of 20 loci and the introduction of at least two specific genetic variants via the HDR pathway. Together these results provide strong proof of principle for CRISPR-Cas9-mediated engineering of the MAC-T cell line, and comprise a major step towards the molecular characterisation of genetic variants that impact bovine milk production and composition.

3.6.1. Critical parameters for the delivery of CRISPR-Cas9 components into a bovine mammary cell line

This study appears to be the first report of CRISPR-Cas9 mediated genome editing in the MAC-T cell line, and one of the first in mammary cells of any species. As such, the protocols for the delivery of the CRISPR-Cas9 components should be useful for other investigators working with MAC-Ts, providing the following observations regarding engineering success in this line.

First, as with any cell transfection, the quantity of material to transfect, quantity of transfection reagent, and cell density is paramount. The careful titration of the quantity of material and transfection reagent revealed that this cell line was relatively robust and could tolerate reasonably high levels of both nucleic acid and transfection reagent, with editing achieved using the highest recommended concentrations of both. Indeed, transfections were conducted using the maximum recommended volume of transfection reagent for both the Cas9 plasmid and protein protocols. This was in line with previous experiments using this cell line for the transfection of other expression plasmids (as described in General Methods and Chapter 5).

Second, there are several methods for the delivery of gRNAs and Cas9 endonuclease into mammalian cells, all of which were highly successful in creating targeted genome edits. As the transfection of DNA expression constructs into this cell line has been highly successful (see Chapter 2), the use of a Cas9 expression plasmid was investigated in the current study. Alongside this approach, the delivery of purified Cas9 protein was also investigated, based on recent reports of this being an efficient alternative to plasmid-mediated delivery of CRISPR-Cas9 complexes (Richardson et al., 2016). To test the efficiency of these CRISPR-Cas9 protocols, the gRNA targeting chr27:36198117T>TGGC was used for all optimisation experiments, selected since this variant is a putative causative mutation for an AGPAT6 eQTL, and several milk composition and yield phenotypes. Editing efficiency was quantified as the number of indels as a proportion of read depth at a given base position in the amplicon. Due to the heterozygous genotype of this variant in MAC-Ts (where the variant is also itself an indel), the use of the commonly used T7E1 assay was abandoned. Instead, deep sequencing was used to quantify CRISPR-Cas9 editing efficiency, and although this method is potentially more expensive, the clonal nature of Illumina-based sequencing makes visualisation of individual edits more straightforward and vastly higher throughput when dealing with large numbers of samples.

Analysis of genomic DNA isolated from the CRISPR-Cas9 treated cell-pools by deep sequencing revealed that the gRNA designed to target the chr27:36198117T>TGGC locus was highly active across the majority of protocols in the current study. For the Cas9 plasmid-based editing there was a large increase in efficiency when enriching GFP positive cells via FACS sorting prior to the addition of the gRNA. In contrast, there was no impact on doubling the amount of Cas9 plasmid transfected into the cells, suggesting that those cells successful transfected with the pMAXGFP and PX459 plasmids may be expressing the encoded genes to saturation. Additionally, the forward transfection of the gRNA was highly efficient, resulting in an indel frequency above 30% across all experimental samples, compared to the reverse transfection of the gRNA that, while resulting in editing efficiencies above 30% in some conditions, overall had lower editing efficiencies, and resulted in some transfections with no apparent editing. Reasons for the variability seen in the CRISPR-Cas9 editing efficiency in reverse transfected samples remains unclear at this stage, and was surprising given that reverse transfections are reported to derive higher editing efficiencies than the traditional, forward procedure (https://www.idtdna.com/pages/docs/default-source/catalog-product-documentation/crispr-2-part-rna-transfection.pdf?sfvrsn=11). This finding may be the result of individual transfection variability, or may be that, as an adherent cell line, MAC-Ts are more amenable to transfections were the cells have had a chance to adhere to the surface of the well.

For the Cas9 RNP-based editing, the forward transfections of the RNP complex resulted in an average indel efficiency of 30%, similar to that seen with FACS enriched, plasmid based transfections. While these two protocols exhibited similar levels of editing at the target locus, the Cas9 plasmid-based protocol took at least six days to complete, involved two separate transfections, and the use of an expensive FACS machine (the availability of which was limited and with use requiring co-ordination with the service provider). By contrast, the Cas9 protein-based protocol was easier to employ as it involved just a single transfection, could be completed within 3 days (not including downstream analysis of editing efficiency), and could be conducted entirely in house.

Additionally, the use of a Cas9 plasmid-based protocol has a greater risk of off-target effects than the Cas9 protein based protocol (Liang et al., 2015; T. Wang et al., 2014). The introduction of exogenous DNA into cells raises the possibility of permanent recombination into the genome, potential disruption of endogenous genes, and longer-term exposure to the encoded Cas9 (Zuris et al., 2015). Despite these risks, the use of Cas9 expression plasmids provide the unique opportunity to multiplex targets and enrich for editing events by having multiple gRNAs and selectable markers (e.g. fluorescence or antibiotic resistance) encoded on the same plasmid. As demonstrated by other studies, further investigations into the use of such plasmids could result in almost all cells containing genomic edits (H. Kim et al., 2011; Patrick et al., 2014).

Given the goals of the current study did not include outcomes that would be specifically served via plasmid-based delivery, transfection of nonreplicable protein Cas9 appears to be the most desirable approach by far. Previous studies have demonstrated that the delivery of Cas9 RNP complexes yields a higher frequency of edits in mouse embryos and mammalian cells (Aida et al., 2015; Liang et al., 2015; Lin, Staahl, Alla, & Doudna, 2014). Ultimately, the forward transfection of CRISPR-Cas9 RNP complexes resulted in the highest editing efficiency at the chr27:36198117 target locus and would be the approach used for future analyses.

3.6.2. CRISPR-Cas9 RNP complexes for targeted genome editing

Having defined optimised reagents and transfection protocols enabling high efficiency editing of MAC-Ts, the next step was to use the optimised conditions to test the efficiency of NHEJ and HDR editing for 28 gRNAs. These targets included candidate causative eQTL variants for the *AGPAT6*, *MGST1*, *DGAT1* and *LGB* genes. The approach taken was to amplify genomic DNA isolated from cell pools treated with the RNP complex and ssODN templates to measure NHEJ and HDR efficiency using deep sequencing. For the assessment of NHEJ, high quality sequence data was generated for 20 of the 28 target loci, while, high quality sequence data was generated for 8 loci of the 15 target loci for the assessment of HDR (HDR could not be measured at the *AGPAT6* locus). At these 20 loci, small indels, the signature of error-prone double-strand beak (DSB) repair via NHEJ were observed at all but one target site. Interestingly, there appeared to be greater indel frequency across the *AGPAT6* target locus compared with the *MGST1* and *LGB* loci. This is the result of the gRNAs targeting this locus having a greater activity than those targeting both the *MGST1* and *LGB* loci. Previous studies have demonstrated that certain gRNAs may not work for unknown reasons (F Ann Ran et al., 2013). Additionally, it is still relatively unclear how nucleosomes and chromatin structure affect CRISPR-Cas9 activity at distinct loci. Recently, Cas9 activity was shown to be enhanced by the use of chromatin remodelers and the activity of Cas9 is variable over several orders of magnitude depending on the dynamic properties of the DNA sequence and the distance for the PAM site from the nucleosome dyad (Isaac et al., 2016). As such, it could be that the broader *AGPAT6* locus is more susceptible to editing due to differences in chromatin accessibility and structure. However, the current study only tested one gRNA for each target site and as it is recommended that the efficiency of more than one gRNA should be tested in the intended cell type, it could also be that the chosen gRNAs were less active in the mammary cells for other reasons (Tsai, Joung, Capecchi, & Evans, 2016).

Interestingly, the relative frequency of these indels was reduced across all sites with the addition of the ssODN. Reasons why the addition of the HDR template appeared to reduce the indel frequency at the target sites is unclear at this stage. It is possible that the ssODN sequestered the Cas9 protein, causing the Cas9 to preferentially cut at the PAM site in the template rather than the genomic target sequence. Alternatively, this could have been the result of poor transfection efficiency in those particular transfections. Despite the rapidly accumulating body of literature describing approaches for HDR-mediated, CRISPR-Cas9-based editing, scant information exists as to how to introduce ssODN template during transfections. I took the approach of adding the ssODN to the transfection mix during the RNP incubation in the Lipofectamine RNAiMAX, but it is possible another approach, such as adding the ssODN to the RNPs immediately after they have formed (Yu et al., 2016) would yield higher efficiency incorporation of HDR templates, and reconcile the observation of reduced NHEJ–mediated indel rates in ssODN transfected samples.

The rate of HDR occurred much less frequently than NHEJ at the majority of target loci. Unfortunately, due to the heterozygous nature of the cell line for all (perfectly linked) *AGPAT6* candidate variants investigated, the rate of HDR could not be established for these 13 sites. Of the 8 loci representing the *MGST1*, *DGAT1*, and *LGB* genes for which sequence data was available, only three demonstrated HDR, highlighting the less efficient nature of this repair pathway. Interestingly, one locus demonstrated higher levels of HDR than NHEJ, which was highly unexpected. Reasons for this remain unclear at this stage, although may represent an inflated estimate of HDR efficiency due to the shallow read depth for the HDR sample, and/or the possibility of technical issues (e.g. poor transfection efficiency) of the RNP complex in the NHEJ sample.

To test whether the HDR frequency could be further optimised, transfections were conducted using different concentrations of ssODN, along with the RNP complex targeting the chr27:93946027T>A locus. The deep sequencing of the genomic DNA isolated from cell pools treated with 0 nM – 20nM ssODN revealed that both the HDR and NHEJ frequency was relatively high for this gRNA. Importantly, the frequency of HDR increased with increasing concentration of ssODN, with a 43.7% increase in HDR from 3 nM ssODN to 20 nM ssODN. This is in line with previous studies, which have demonstrated increasing the ratio of gRNAs and HDR templates increasing HDR frequency at the target locus (Paix, Folkmann, Rasoloson, & Seydoux, 2015). Since the most concentrated ssODN yielded the highest frequency of editing, the upper limit of this does response is not known.

Similar to the previous experiment, the frequency of NHEJ decreased with the addition of the HDR template in a dose-response manner. Again, reasons for this remain unclear, but it could be that the cell only has a certain level of repair capability, and the increase in the efficiency of one pathway may be at the expense of the other. This observation might also relate to sequestration of Cas9 as hypothesised earlier, or derive from some unknown mechanism based on the many different factors known to influence repair fates. HDR is highly dependent on the cell type and state, genomic locus and sequence characteristics of the repair template, and the phase of the cell cycle (Chu et al., 2015; Lin et al., 2014).

3.6.3. Targeting more than one locus using CRISPR-Cas9 RNP complexes

Following the identification of the optimal conditions for CRISPR-Cas9 editing of a single locus, the next step was to target more than one locus in the same transfection. The approach taken was to conduct transfections using four gRNAs and the Cas9 protein in an equimolar concentration up to final concentrations of 20-80 nM total RNP. This approach demonstrated that more than one locus could be edited in a single transfection of RNPs in a dose-response manner, with those samples treated with 80 nM total RNP demonstrated the highest indel frequency. At this concentration, the four gRNAs were present at approximately 20 nM RNP complex each, corresponding to the RNP concentration demonstrated to result in the highest indel frequency when targeting a single locus (see 3.6.1). Interestingly, there didn't appear to be much cell death associated with increasing the amount RNP transfected, with the cells across all concentrations reaching confluency at the same time. However, it has previously been demonstrated that Cas9 off-target edits, are much more frequent at higher concentrations of Cas9 (Hsu, Lander, & Zhang, 2014). This may limit the level to which this technology can be multiplexed, given the poor efficiency across the target sites at the lower concentrations of RNP complex.

Ultimately, further work is required to establish if more than one locus can be edited in the same cell, since these results are from DNA extracted from cell pools. Despite two or more loci demonstrating editing in these experiments, it was not possible to establish if these events were occurring in single cells, or occurring as single edits in multiple independent cells. Although clonal isolation would be required to answer this question, it seems feasible that multiplex editing may be occurring, perhaps at frequencies higher than the combinatorial efficiencies demonstrated for the individual loci. This hypothesis is based on the fact that editing efficiency is quantified based on data from both transfected and transfected cells, so the chance that single successfully transfected cells bear multiple edits may be higher than this metric suggests. Regardless, these results provide promise towards the efficient editing of multiple distinct target loci in parallel, which would be the ultimate realisation of these tools in the generation of animals simultaneously engineered for multiple traits.

3.6.4. Future directions

The successful genome editing of a mammary cell line constitutes the first step towards generating modified cell lines to characterise and probe the functionality of genetic variant(s) highly associated with bovine milk production and composition. The next step in this project is to conduct single-cell sorting via FACS of the GFP positive cells to establish clonal cell lines, which can be screened for the presence of the variants targeted in this study. Based on the editing frequencies achieved in this current study without any enrichment or selection, such a protocol could result in a significant proportion of these isolated cell lines containing genomic edits.

Once genome-edited cell lines have been isolated, they can be utilised for downstream applications, such as the analysis of their impacts on gene expression. Such experiments have the potential to provide previously unattainable information about the role of specific genetic variants statistically associated with milk production. As the resolution of GWAS is limited by the LD structure of the study population, such analyses typically identify haplotypes (which can encompass up to several Mbp). It is often difficult to determine which of several genetic variants in tight LD are responsible for the phenotype, so decoupling variants synthetically using CRISPR-Cas9 may allow pinpointing the variants responsible.

3.7. Summary and conclusion

CRISPR-Cas9 mediated genome editing is revolutionising the generation of transgenic *in vitro* and *in vivo* models. By recapitulating genetic mutations found via GWAS in study populations, CRISPR-Cas9 mediated editing can be used to rapidly model the causal roles of specific genetic variants. In this chapter, I have developed methodology to enable these research questions to be asked in a bovine mammary context. These results show that highly efficient engineering of the MAC-T cell line is possible, and should be useful for other researchers investigating aspects of mammary biology and bovine genetics.

Chapter 4: Detailed investigation of a milk fat percentage QTL underpinned by the *MGST1* gene

4.1 Overview

Bovine milk fat percentage is a complex trait determined by numerous genetic effects. The majority of these quantitative trait loci (QTL) are small effects, while a handful comprise major effects (Ben J. Hayes, Pryce, Chamberlain, Bowman, & Goddard, 2010). Of these loci, a QTL on bovine chromosome 5 with a large effect on milk fat percentage has been identified in several independent genetic studies (Ben J. Hayes et al., 2010; Kathryn E Kemper et al., 2015; Raven, Cocks, & Hayes, 2014; X. Wang et al., 2012). In a study of a German Holstein-Friesian population, this locus represented the second largest effect on milk fat percentage, with only the chromosome 14 *DGAT1* QTL having larger impacts (X. Wang et al., 2012). Despite the large effect size at this locus, the gene (and causative variant) responsible has yet to be established.

Both *EPS8* and *MGST1* have been proposed as underpinning the chromosome 5 milk fat percentage QTL (Raven, Cocks, & Hayes, 2014; X. Wang et al., 2012). The ambiguity around the causative status of these genes is a reflection of the fact that the most significantly associated markers do not map to protein-coding sequences, and neither gene represents a strong biological candidate for milk fat modulation. This chapter describes the detailed investigation of the chromosome 5 milk fat percentage locus, which was conducted as part of a broader study led by Littlejohn et al., (2016). This paper focused on investigating *MGST1* as the candidate gene underlying the QTL, where the gene was prioritised due to the identification of a strong *cis* eQTL in mammary RNAseq data (M. D. Littlejohn et al., 2016). The work described here examines candidate variants for those effects, and mechanistic aspects of the role of *MGST1* in milk composition. Analyses include local sequence refinement conducted to identify any 'hidden' candidate causative variants in a 360 base pair (bp) reference sequence gap in the first intron of *MGST1*; investigation of a copy number variant (CNV) adjacent the transcription start site (TSS) of *MGST1*, also assessed as a possible causative variant of the expression and milk composition QTLs, and a *trans*-eQTL analysis to investigate if the expression of any other genes and/or molecular pathways might give a clue to the mechanism of phenotype modulation. Additionally, association analysis was conducted using milk fatty acid profiles, an analysis that was not reported in the Littlejohn et al., (2016) paper, and is novel to this thesis.

4.2 General aim

Investigate the causative status of the *MGST1* gene as underpinning the chromosome 5 milk fat percentage QTL.

4.2.1 Specific aims

1. Conduct refinement of the reference assembly representing the *MGST1* locus by PCR amplifying and Sanger sequencing of an approximately 364 bp reference gap in intron 1 of *MGST1*.
2. Determine the genomic sequence of the breakpoints for a CNV detected adjacent the TSS of *MGST1* via PCR and Sanger sequencing. Use these assays to genotype 13 whole-genome sequenced bulls.
3. Conduct association analysis of individual fatty acid milk profiles using *MGST1* locus genotypes in 617 Friesian-Jersey crossbred F2 animals.
4. Conduct *trans*-eQTL analysis to identify co-associated gene expression networks based on *MGST1* QTL genotype.

4.3 Methods

4.3.1 Animal cohorts, phenotypes and genotypes

The work described in this chapter was conducted using four independent animal populations; the FJXB cohort, high genetic merit Livestock Improvement Corporation (LIC) sires, the mammary RNAseq animals, and the mixed ancestry dairy cow population, all of which are detailed in General Methods (Chapter 2). The specific numbers of animals studied, and details of phenotypes and genotypes pertaining to this work are stated below.

4.3.2 PCR and Sanger sequencing of a reference sequence gap in *MGST1*

PCR and Sanger sequencing of the reference sequence gap in intron 1 of *MGST1* was conducted on genomic DNA from six F1 sires and six F1 dams from the FJXB population.

PCR was conducted with the primers in Table 4.1 designed to amplify the 364 bp 'N' sequence (chr5:93942388-93942751; UMD 3.1 genome build) using KAPA 2G Robust enzyme in conjunction with the KAPA GC-rich buffer (Kapa Biosystems; see General Methods for

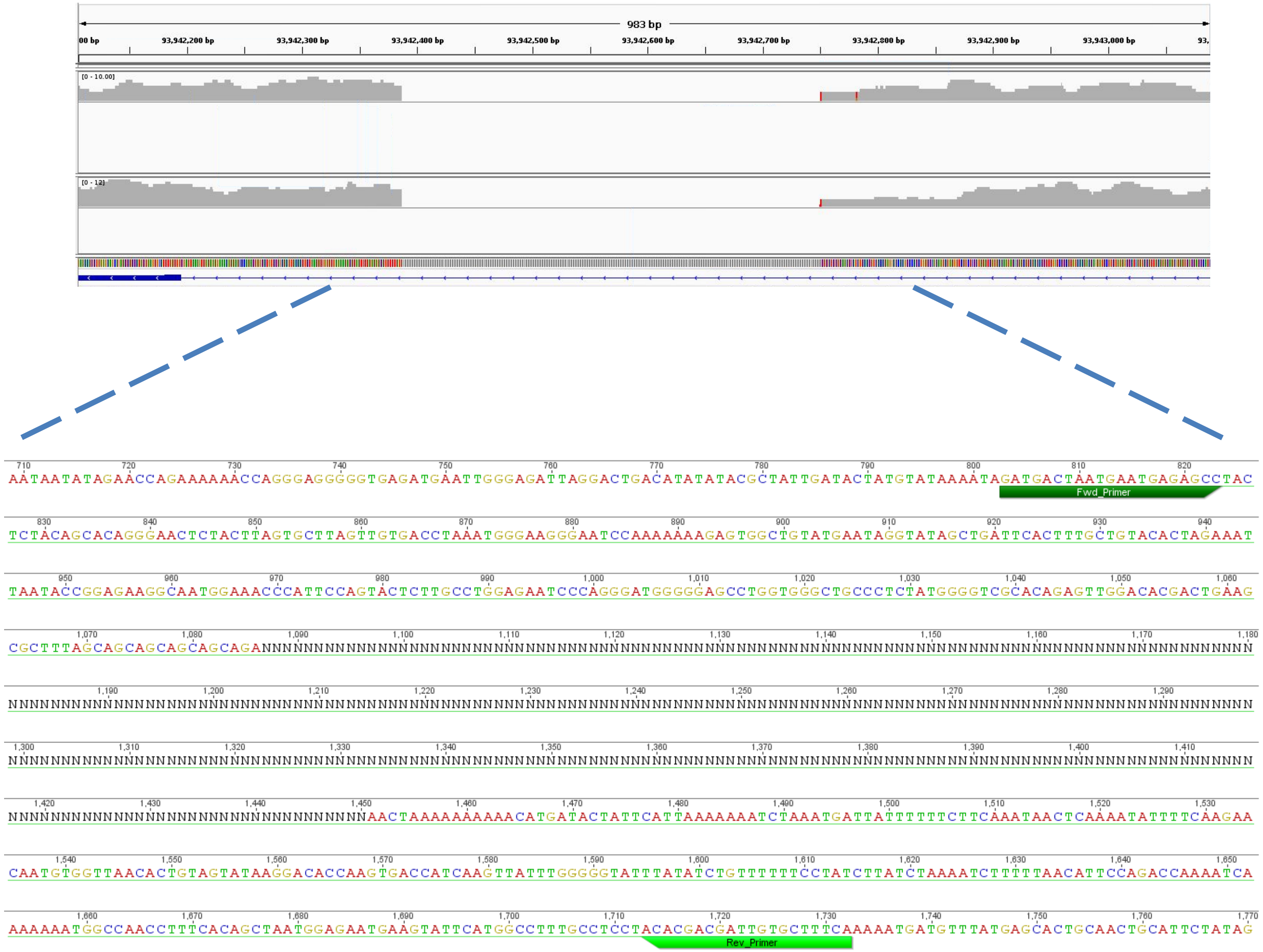
PCR reaction mix). Cycling conditions were: 95°C for 3 minutes; and 35 cycles of 95°C for 30 seconds, 59°C for 30 seconds, 72°C for 30 seconds.

For Sanger sequencing, the PCR products from two sires and two dams were purified using ExoSAP-IT (Affymetrix) as per the manufacturer's instructions. Briefly, 15 µL of PCR reaction product was mixed with 6 µL ExoSAP-IT and incubated at 37^{o}C for 15 minutes followed by incubation at 80^{o}C for 15 minutes. Then, the PCR products were prepared along with the primers in Table 4.1 and sequenced in both the forward and reverse directions by the Genomics Centre, Auckland Science Analytical Services, The University of Auckland (Auckland, NZ).

The resultant sequences were visualised and aligned to each other using Geneious software (version 6.1.7) with the default multiple alignment settings to identify a consensus sequence through this gap region. This consensus sequence was then aligned to the UMD3.1 reference sequence for *MGST1* intron 1 (Accession number AC_000162.1). Genetic variants were annotated in the consensus sequence by combining all the high quality sequence variants in these animals.

Table 4.1 PCR primers for Sanger sequencing of a gap in *MGST1* intron 1

Primer Name	Primer Sequence	Product Size
MGST1_Gap_Fwd	GATGACTAATGAATGAGAGCC	930 bp
MGST1_Gap_Rev	TGAAAGCACAATCGTCGTGT	



```
AATAATATAGAACCAGAAAAAACCAGGGAGGGGGTGAGATGAATTGGGAGATTAGGACTGACATATATACGCTATTGATACTATGTATAAAATAGATGACTAATGAATGAGAGCCTAC
TCTACAGCACAGGGAACTCTACTTAGTGCTTAGTTGTGACCTAAATGGGAAGGGAATCCAAAAAAAGAGTGGCTGTATGAATAGGTATAGCTGATTCACTTTGCTGTACACTAGAAAT
TAATACCGGAGAAGGCAATGGAAACCCATTCCAGTACTCTTGCCTGGAGAATCCCAGGGATGGGGGAGCCTGGTGGGCTGCCCTCTATGGGGTCGCACAGAGTTGGACACGACTGAAG
CGCTTTAGCAGCAGCAGCAGCAGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAACTAAAAAAAAAACATGATACTATTCATTAAAAAAATCTAAATGATTATTTTTTCTTCAAATAACTCAAAATATTTTCAAGAA
CAATGTGGTTAACACTGTAGTATAAGGACACCAAGTGACCATCAAGTTATTTGGGGGTATTTATATCTGTTTTTTCCTATCTTATCTAAAATCTTTTTAACATTCCAGACCAAAATCA
AAAAAATGGCCAACCTTTCACAGCTAATGGAGAATGAAGTATTCATGGCCTTTGCCTCCTACACGACGATTGTGCTTTCAAAAATGATGTTTATGAGCACTGCAACTGCATTCTATAG
```

Figure 4.1 Reference sequence gap in UMD3.1/Btau6 within *MGST1* intron 1.
The top image shows the coverage in two whole-genome sequenced animals for the genomic region surrounding the reference sequence gap in *MGST1* intron 1 where no sequence reads are able to map to this location. The bottom image shows the 364 bp 'N' sequence (chr5: 93942388-93942751), which is encompassed by the MGST1_Gap_For and MGST1_Gap_Rev primers for PCR amplification of these nucleotides.

4.3.3 PCR and Sanger sequencing for identification of breakpoint sequences and genotyping of a candidate causative CNV

Inspection of whole-genome sequence (WGS) alignments revealed read depth anomalies upstream of the *MGST1* TSS, and using CNVnator software, an approximately 8 kilobase (kb) polymorphic deletion was detected (detailed in Littlejohn et al., 2016). To determine the breakpoint sequences for the CNV adjacent the TSS of *MGST1*, PCR and Sanger sequencing of the breakpoints was conducted on genomic DNA isolated from the semen of 13 high genetic merit LIC sires. The DNA extraction protocol is described in General Methods.

PCR products were designed to span the breakpoints apparent from sequence alignments (Figure 4.2), using the primers in Table 4.2. PCR reactions used the KAPA 2G Robust enzyme in conjunction with the KAPA GC-rich buffer and used the following cycling parameters: 95°C for 3 minutes; and 35 cycles of 95°C for 30 seconds, 56°C for 30 seconds, 72°C for 90 seconds (see General Methods for PCR reaction mix).

PCR products from MGST1_CNV_F1 (F1) and MGST1_CNV_R1 (R1) and MGST1_CNV_F1 (F1) and MGST1_CNV_R6 (R6) were visualised following separation by gel electrophoresis (2% w/v agarose) for 60 minutes. Then, the PCR products from two bulls from each genotype class were purified using AxyPrep PCR clean-up kit (Axygen), as per the manufacturers' protocol. The PCR products were prepared along with the primers in Table 4.2 and sequenced in both the forward and reverse directions by the Genomics Centre, Auckland Science Analytical Services, The University of Auckland, (Auckland, NZ; see General Methods).

The resultant sequences were visualised and aligned to each other using Geneious software (version 6.1.7) with the default multiple alignment settings. Gap open penalty and gap extension penalty parameters were relaxed to prevent any penalisation for the presence of the CNV within the PCR products.

For animal genotyping, the presence and/or absence of amplification products for the F1/R1, and F1/F6 junction products was used to determine the genotypes for the 13 sires. Under the above cycling conditions, a PCR product for the F1/R1 primer pair would represent a copy number (CN) of at least one for the CNV, while a product for the F1/F6 would suggest a CN of zero i.e. a deletion of the CNV. As such, an animal that had no amplification from the F1/F6 primer pair but had amplification products from F1/R1 would be genotype CN2, while an animal that had no amplification from F1/R1 but had amplification products from F1/R6 would be genotype CN0. An animal that had amplification products from both primer pairs would be genotype CN1 (Figure 4.2). For each sire both PCR assays were conducted in duplicate to ensure accuracy of genotype calls. These genotypes were also compared to the CNVnator-derived genotypes called from whole genome sequence (WGS) for each sire.

Table 4.2 Primer sequences for identification of breakpoint signatures and sire genotyping.

Primer Name	Primer Sequence
MGST1_CNV_F1	TCGAAAGGCTGGCACTGACAACGA
MGST1_CNV_R1	GCAGCTGCAGTAGCACATAT
MGST_CNV_F6	TGGCTGAGTAATACTGATCTGCC
MGST1_CNV_R6	TCCCCACTTTCCCCTTTACT

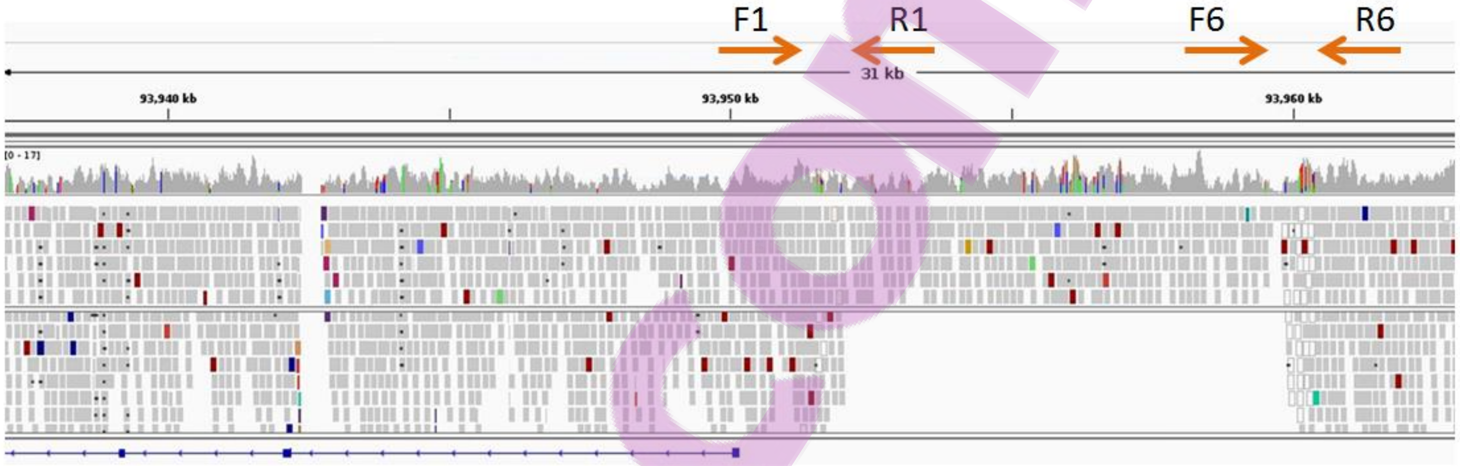


Figure 4.2 CNV upstream of the TSS of *MGST1*.
The image illustrates the coverage in two whole-genome sequenced animals encompassing the CNV. The top alignment represents a CN2 animal while the bottom alignment represents a CNV0 as evidenced by the lack of reads mapping to this location. The arrows corresponding to F1 and R1 are primer pairs designed to amplify each breakpoint and were subsequently used to genotype the 13 WGS bulls.

4.3.4 Association analyses

4.3.4.1 Fatty acid profiles in the FJXB cohort

Association analysis was conducted in 617 F2 dairy cows from the FJXB population using the individual milk fatty acid composition phenotypes in conjunction with genotypes of the chr5:93945738T>C SNP.

Fatty acids were extracted by a modification of the Röse Gottlieb technique and quantified by gas-liquid chromatography on a Shimadzu GC17A instrument (Shimadzu Corporation) at Fonterra Research Centre (Palmerston North, NZ). The phenotypes were generated outside the remit of this thesis and were presented as the proportion of individual milk fatty acids (in grams (g)) per 100 g of fatty acid).

Genotypes for the chr5:93945738T>C SNP were imputed into the FJXB population using a reference population of 556 animals as described in General Methods. These genotypes were extracted using samtools (version 0.1.19), and recoded using PLINK2 (version 1.90b2c; Chang et al., 2015) to 0, 1, or 2 to represent the number of alternative alleles for this marker (i.e. 0, 1, and 2 to represent the homozygous reference, heterozygous, and homozygous alternative genotypes, respectively).

To examine the effect of QTL genotype on milk fatty acid composition in the FJXB F2 animals, association analysis was conducted using the relative proportions of individual fatty acids in conjunction with *MGST1* genotype. Associations were quantified using pedigree-based mixed models in ASReml-R (A R Gilmour et al., 2009; Arthur R. Gilmour et al., 1995). The chr5:93945738T>C SNP was fitted in a separate sire-maternal grandsire single trait model for each milk fatty acid, and treated as a quantitative variable based on the number of copies of the alternative allele and variance components estimated in a restricted maximum-likelihood (REML) framework. Covariates included the proportions of NZ Holstein-Friesian ancestry, US Holstein-Friesian ancestry, Jersey ancestry and heterosis effects were also included in the models. The additive genetic variance, polygenic genetic variances, total genetic variance, and phenotypic variance for each milk fatty acid phenotype was calculated as described in General Methods. The proportion of phenotypic and genotypic variances explained by the chr5:93945738T>C SNP was also calculated as described in General Methods.

4.3.4.2 *Trans*-eQTL analysis in the RNAseq animals

To attempt to identify co-associated gene expression networks based on *MGST1* QTL genotype, *trans*-eQTL analysis was conducted using the high-depth mammary RNAseq dataset (described in General Methods). Gene expression for genome-wide genes was quantified using sequence read counts, normalised and transformed as described in General Methods.

The chr5:93945738T>C SNP genotypes were imputed into the RNAseq animals using a reference population of 556 animals, as per analysis of the FJXB cohort described above.

Trans-eQTL analysis was conducted using ASReml-R to fit similar models to those used for fatty acid profile association mapping, in this case testing the effect of chr5:93945738 genotype on the 9,348 nominally expressed genes (detailed below) in the RNAseq cohort. Compared to analysis of fatty-acids, an additional effect in these models included sequencing cohort fitted as a fixed effect.

Nominal Gene Expression level cut-off

To minimise false positive associations through inclusion of genes with insufficient read depth for meaningful analysis, a nominal gene expression filter was applied for *trans*-eQTL analysis. To this end, only genes with minimum of 0.5 fragments per kilobase of exon model per million mapped (FPKM) in 75% or more of the 375 RNAseq animals were considered for analysis (yielding 9,382 genes).

4.4 Results

4.4.1 Local refinement of the reference assembly reveals a 410bp sequence gap in *MGST1* intron 1

To determine the nucleotides of the 364 bp 'N' sequence in *MGST1* intron 1, PCR and Sanger sequencing was conducted on six F1 sires and six F1 dams from the FXJB cohort. A number of different primer pairs were designed to target the gap, with the MGST1_Gap_Fwd and MGST1_Gap_Rev pair in particular providing a clean, single band of approximately 1000 bp, when visualised by gel electrophoresis (Figure 4.3). Notably, the size of the amplification product was slighter larger than anticipated, where the expected size from the reference sequence was only 930 bp.

The PCR products from two sires and two dams were purified and sequenced. The resultant sequence was aligned to the UMD3.1 reference sequence for *MGST1* intron 1 revealing a 410 bp gap - 46 nucleotides longer than the 364 bp 'N' sequence indicated in the reference genome build. The aligned sequences from these animals were inspected for additional genetic variants, revealing 7 SNPs and 3 small insertions within the sequence gap (Figure 4.4). The resolved gap-sequence was also used to supplement the reference assembly for mapping, imputation, and association analysis of whole genome sequence data, described in detail in Littlejohn et al. (2016).

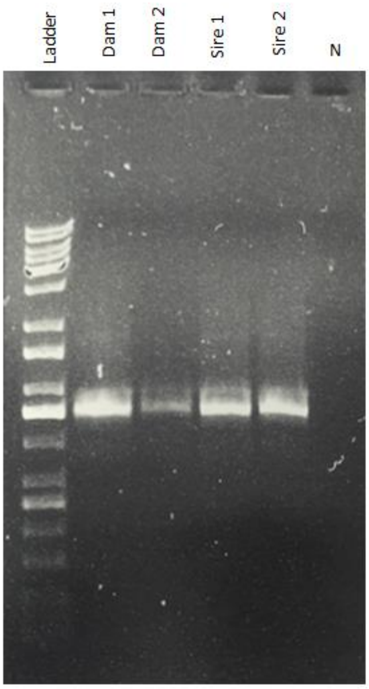
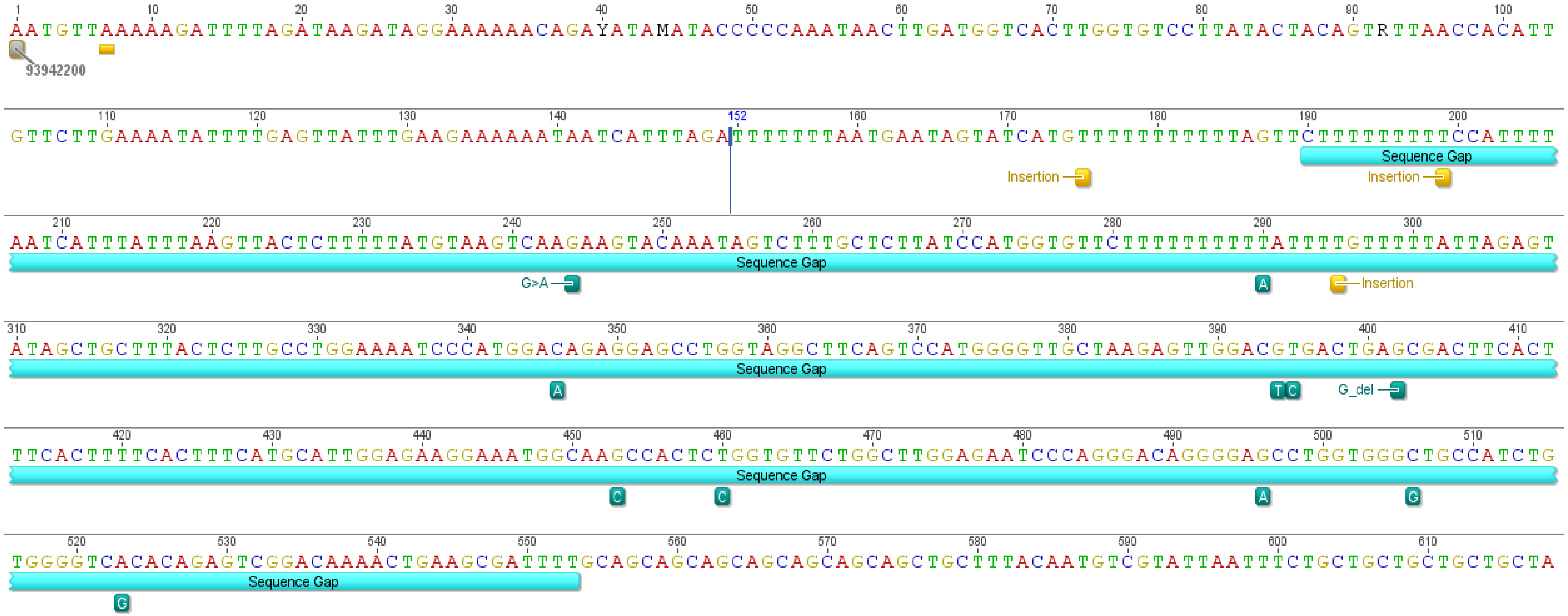


Figure 4.3 Gel electrophoresis of PCR products from two dams and two sires encompassing the *MGST1* intron 1 gap. Ladder = Kapa Universal DNA ladder. N = H_2O blank.



```
AATGTTAAAAAGATTTTAGATAAGATAGGAAAAAACAGAYATAMATACCCCCAAATAACTTGATGGTCACTTGGTGTCCTTATACTACAGTRTTAACCACATT
GTTCTTGAAAATATTTTGAGTTATTTGAAGAAAAAATAATCATTTAGATTTTTTTAATGAATAGTATCATGTTTTTTTTTTTAGTTCTTTTTTTTTCCATTTT
AATCATTTATTTAAGTTACTCTTTTTATGTAAGTCAAGAAGTACAAATAGTCTTTGCTCTTATCCATGGTGTTCTTTTTTTTTTATTTTGTTTTTATTAGAGT
ATAGCTGCTTTACTCTTGCCTGGAAAATCCCATGGACAGAGGAGCCTGGTAGGCTTCAGTCCATGGGGTTGCTAAGAGTTGGACGTGACTGAGCGACTTCACT
TTCACTTTTCACTTTCATGCATTGGAGAAGGAAATGGCAAGCCACTCTGGTGTTCTGGCTTGGAGAATCCCAGGGACAGGGGAGCCTGGTGGGCTGCCATCTG
TGGGGTCACACAGAGTCGGACAAAACTGAAGCGATTTTGCAGCAGCAGCAGCAGCAGCAGCTGCTTTACAATGTCGTATTAATTTCTGCTGCTGCTGCTGCTA
```

Figure 4.4 Consensus sequence for a gap in the UMD3.1 reference sequence in the promoter of *MGST1*
The genetic variants are shown in blue (13 SNP, 1 MNP and one deletion) and yellow (insertions). This sequence gap was established to be 410 bp in length, 46 nucleotides larger than anticipated.

4.4.2 Identification of CNV breakpoint sequences and genotyping candidate variant

Inspection of WGS alignments revealed a CNV that, given its proximity to the *MGST1* TSS, was a strong candidate mutation for the QTL. To investigate this CNV as a candidate variant, PCR and Sanger sequencing was conducted on 13 WGS sires to determine the exact size and structure of the CNV. Three PCR primer pairs were designed targeting each junction, with F1 and R1, and F1 and R6 producing single bands of the expected size when visualised by gel electrophoresis. These primer pairs were used to generate PCR products for two sires representing each CNVnator-called genotype class, such that at least three PCR products encompassing the breakpoints on either side of the CNV were purified, sequenced, and aligned to each other and the UMD3.1 reference sequence.

The alignment of the DNA sequences revealed an 8,202 bp deletion, in line with the size estimated from WGS alignments (Figure 4.2). Interestingly, this deletion appeared to abridge two ART2A RTE-BovB repeat fragments. An identical 129 bp sequence within these fragments prevented precise determination of the CNV breakpoints, instead providing a 'sliding window' of 129 bp at chr5:93951990-93952118 and chr5:93960192-93960320 from which the 8.2 kb segment was deleted (Figure 4.6).

These PCR assays were used to genotype 13 sires representing CNVnator-called genotype classes, with the presence and/or absence of amplification products for the F1/R1 and F1/R6 junction products indicative of the number of copies of the CNV. Both assays were carried out in duplicate for each of the 13 sires (as demonstrated by Figure 4.5). Notably, all 13 of the PCR-based genotype calls matched the CNVnator-called genotype classes, confirming concordance between PCR and sequence-based calls (Table 4.3).

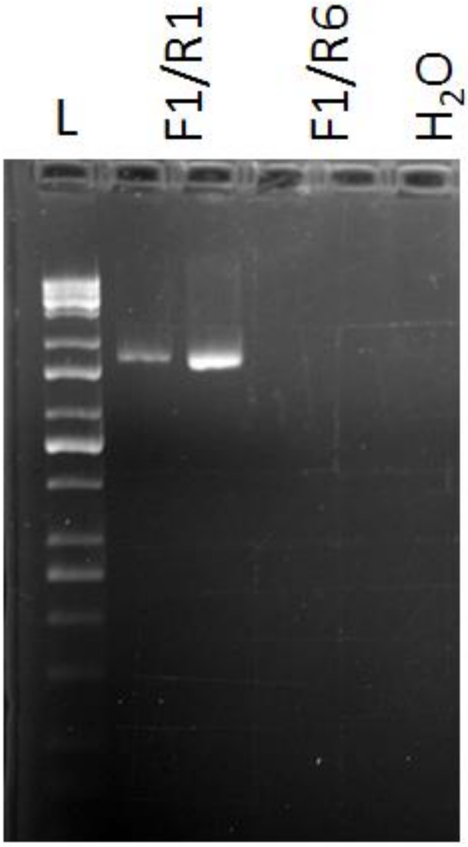


Figure 4.5 Gel electrophoresis image of F1/R1 and F1/R6 PCR products used to genotype the CNV adjacent to *MGST1* TSS in sire 22106318.
The animals were genotyped by PCR using KAPA2G Robust DNA polymerase and primers spanning the breakpoints for the CNV. PCR reactions were completed in duplicate for each sire. H_2O = water blank, L = KAPA Universal ladder.

Table 4.3 ***MGST1*** **CNV genotypes for LIC WGS bulls**
CN0 = zero copy number, CN1 = one copy number, CN2 = two copy numbers, of the CNV.

Animal Key	Name	F1/R1 PCR product	F1/R6 PCR product	Genotype
17034899	SCOTTS NORTHSEA	+	-	CN2
18278012	OKURA ACE ISAAQ ET	-	+	CN0
16052387	ERRLYN SS PRIDE GR	-	+	CN0
15729613	DAYSH'S LANDMARK GR	-	+	CN0
15462331	WILLIAMS ACE OF HEARTS	+	+	CN1
18183164	SRB GLENMEAD ROCKFEST-ET	+	-	CN2
15656114	CHRISTENSENS LIEGE	+	+	CN1
22990980	TIRONUI MEGANEV	+	+	CN1
19116294	VAN BYSTERVELDTS HOMERUN	+	+	CN1
17999932	MUDFORDS LEGENDAIRE	+	-	CN2
25330663	FOXTON NN FROSTY S3J	+	-	CN2
22253265	BLAKELOCK MD KNIGHT S3F	+	-	CN2
22106318	MAXWELLS DAN JAZZMAN S2F	+	-	CN2

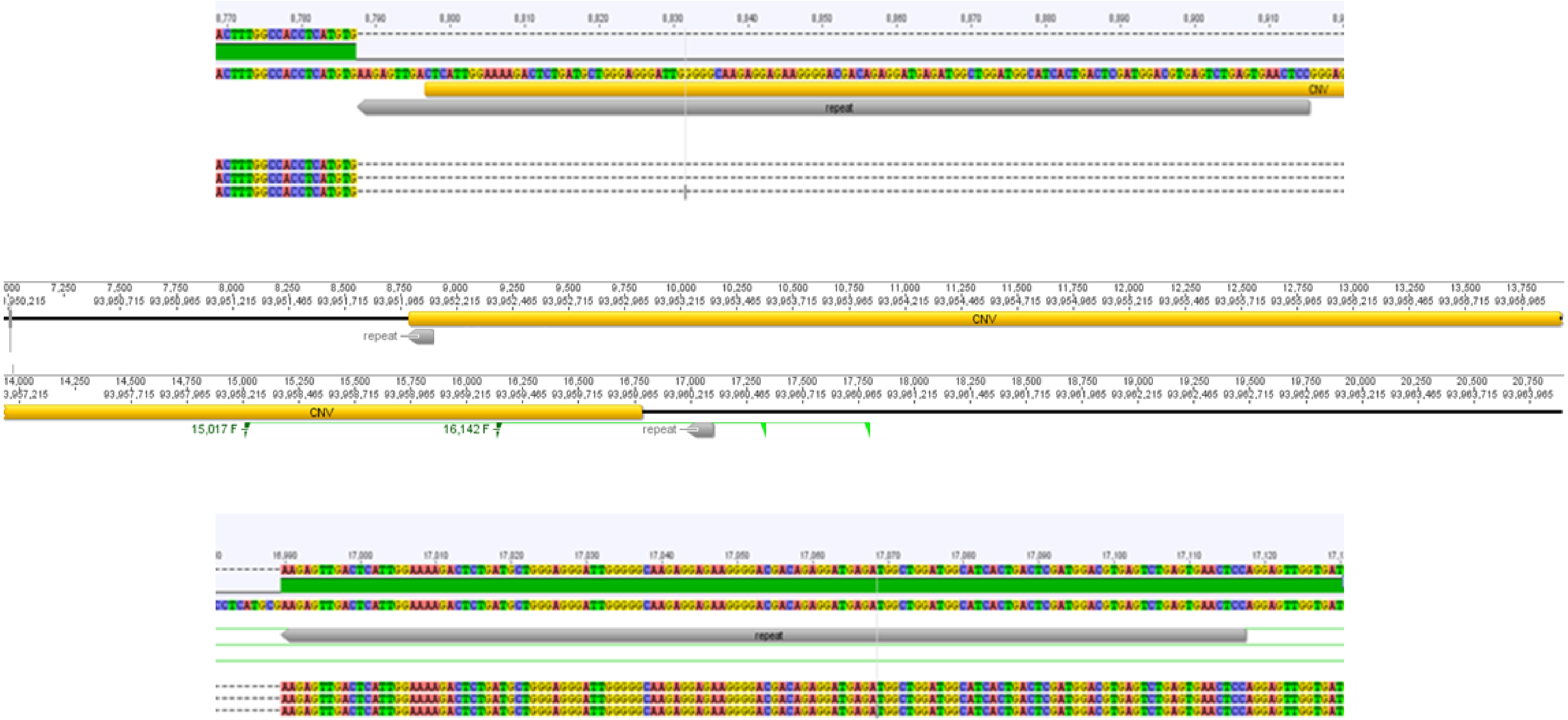


Figure 4.6 Consensus sequence for a 8,202 bp CNV adjacent to the TSS of *MGST1*
The identification of a 129 bp repeat motif (grey) at the breakpoints of the CNV prevented the precise delineation of the genomic coordinates of the left and right breakpoints for the CNV. These repeats were located at chr5:93951990-93952118 and chr5:93960192-93960320 (UMD3.1 genome build) with breakpoints residing within these boundaries.

4.4.3 Polymorphisms in *MGST1* do not associate with milk fatty acid composition

To investigate if other milk composition phenotypes may be impacted by the *MGST1* QTL, association analysis was conducted between the proportions of the individual fatty acids in conjunction with imputed genotypes for the chr5:93945738T>C SNP. After adjusting for multiple hypothesis testing, this analysis revealed no significant associations between the proportions of these fatty acids in milk and the chr5:93945738T>C SNP (Bonferroni significance threshold = P=0.0012; Table 4.4). Of those fatty acids tested, there were four that were significantly associated with the chr5:93945738T>C SNP using an alpha value of 0.05 in the absences of a multiple testing correction. These were: C20:0, C15:isoBr, C16:1 and C14:1 (P=0.0249, 0.0283, 0.0333, 0.0431, respectively; Table 4.4).

4.4.4 *Trans*-eQTL analysis

Trans-eQTL analysis was conducted to attempt to provide insight into the possible expression networks mediating the milk composition effects driven by differential expression of *MGST1*. Unfortunately, other than the highly significant *cis* effect on *MGST1*, association analysis between 9,348 mammary-expressed genes and the chr5:93945738T>C SNP failed to reveal any significantly differentially expressed genes when accounting for multiple hypothesis testing (Table 4.5).

However, differential expression of two genes (*ZNF593* and *DDX27),* approached the significance threshold of P=5.35x10^{-6} (P=4.41x10^{-05} and P=5.13x10^{-05}Table 4.5). The 93945738T>C SNP explained 5.4% and 5.3% of the phenotypic variance of the mammary expression of these two genes, respectively.

Table 4.4 Individual fatty acid association statistics for the chr5:939345738T>C SNP
'Phenotype' indicates the individual fatty acids tested. 'Parameter Estimate' indicates the per-allele parameter estimate and standard errors calculated from the restricted maximum likelihood models. 'Pheno var' indicates the proportion of phenotypic variance explained by the chr5:93945738 SNP for each fatty acid tested, with p-values of association indicated in the right-most column. Multiple testing threshold is P=0.012.

Phenotype	Parameter Estimate	Pheno var	P-value
C20_0	-0.122 (±0.055)	0.838	0.0248
C15_0isoBr	0.009 (±0.004)	0.882	0.0283
C16_1	0.035 (±0.016)	0.801	0.0333
C14_1	0.028 (±0.014)	0.742	0.0430
C18_0	-0.245 (±0.162)	0.423	0.1307
C20_1n9	-0.244 (±0.164)	0.412	0.1359
C15_0	0.018 (±0.012)	0.374	0.1414
C15_0anteisoBr	-0.071 (±0.049)	0.345	0.1504
C10_1	0.007 (±0.005)	0.362	0.1560
C18_2conjc9t11	0.029 (±0.021)	0.331	0.1658
UnkI	0.019 (±0.014)	0.304	0.1956
C20_5n3EPA	0.047 (±0.039)	0.245	0.2263
C20_3n3	0.033 (±0.027)	0.239	0.2318
C17_0isoBr	0.008 (±0.007)	0.246	0.2438
C13_0	0.181 (±0.170)	0.200	0.2868
UnkJ	-0.010 (±0.011)	0.158	0.3489
C14_0Br	-0.112 (±0.126)	0.133	0.3732
C20_4n6AA	-0.191 (±0.243)	0.111	0.4335
C16_0Br	0.002 (±0.002)	0.112	0.4338
C13_0Br	-0.196 (±0.254)	0.099	0.4409
C18_1n7	0.051 (±0.069)	0.096	0.4614
C18_2n6Linoleic	-0.008 (±0.011)	0.091	0.4652
C16_0	0.118 (±0.176)	0.075	0.5021

C6_0	0.013 (±0.020)	0.076	0.5165
C20_4n3	0.161 (±0.265)	0.069	0.5431
C12_1	0.101 (±0.169)	0.063	0.5521
C24_0	-0.021 (±0.038)	0.049	0.5875
C22_0	0.176 (±0.332)	0.051	0.5977
C18_1n9	-0.094 (±0.228)	0.029	0.6805
C17_0anteisoBr	0.002 (±0.004)	0.027	0.7004
C8_0	0.007 (±0.017)	0.025	0.7014
C4_0	0.010 (±0.028)	0.024	0.7125
C17_0	0.003 (±0.011)	0.011	0.8060
C17_1	0.001 (±0.004)	0.010	0.8098
C10_0	0.011 (±0.055)	0.006	0.8459
C12_0	0.012 (±0.066)	0.006	0.8551
C20_3n6	-0.062 (±0.366)	0.005	0.8653
C18_3n3	-0.002 (±0.012)	0.003	0.8918
C22_5n3	-0.008 (±0.112)	0.001	0.9420
C20_1n11	-0.017 (±0.306)	0.001	0.9555
C14_0	0.001 (±0.100)	0.000	0.9952

Table 4.5 Differentially expressed genes in the lactating mammary gland based on milk composition QTL/*MGST1* eQTL genotype.

'Parameter Estimate' indicates the per-allele parameter estimate and standard errors calculated from the restricted maximum likelihood models. 'Pheno var' indicates the proportion of phenotypic variance explained by the chr5:93945738 SNP for individual gene expression, with p-values of association indicated in the right-most column. Significance threshold = P=5.35x10^{-6}

Gene	Genomic Location	Parameter Estimate	Pheno Var	P-value
MGST1	Chr5:93926791-93950162	-0.4322(±0.0201)	60.73	3.22x10^{-66}
ZNF593	Chr2:127523710-127524790	-0.1079(±0.0261)	5.39	4.41x10^{-05}
DDX27	Chr13:78054502-78071705	-0.0550(±0.0134)	5.33	5.13x10^{-05}

4.5 Discussion

A QTL with a major impact on bovine milk fat percentage resides on chromosome 5. In a study of a Holstein-Friesian population, this locus represented the second largest effect on milk fat percentage, with only the chromosome 14 *DGAT1* QTL having a greater impact (X. Wang et al., 2012). Studies have proposed both *MGST1* and *EPS8* as the gene responsible for the QTL, however neither gene represents a strong biological candidate for milk fat modulation, and the most highly associated markers map to the non-coding sequences surrounding these genes (Raven, Cocks, Goddard, et al., 2014; X. Wang et al., 2012). We recently identified a strong mammary *cis*-eQTL for *MGST1* bearing the same genetic signal underpinning the milk fat percentage QTL, providing the first functional support for the gene and mechanism underpinning this effect (M. D. Littlejohn et al., 2016). The next step in characterising this QTL was to identify the cellular mechanism and specific genetic variant by which *MGST1* mediates its effect on milk fat percentage. To this end, this chapter describes the further investigation of this locus, by examining candidate variants for these effects, and by looking for further gene expression consequences to *MGST1* expression modulation.

4.5.1 Local sequence refinement at the *MGST1* locus revealed additional genetic variants

As part of the work reported in Littlejohn et al. (2016), we noted strong association between milk fat percentage and a cluster of whole-genome sequence-resolution variants in *MGST1* intron 1. Conspicuously, these variants were immediately adjacent a reference sequence gap, presenting the possibility that candidate causative variants could be 'hidden' in this gap. The approach was taken to fill the reference gap using PCR and Sanger sequencing of genomic DNA in NZ dairy cattle, thereby allowing variants to be catalogued, and subsequently used for imputation and association analysis. Sequencing of this interval revealed a 410 bp gap, slightly larger than the 364 bp run of N's present in the UMD3.1 genome build. The gap was found to contain 7 SNPs and 3 small insertions in the four Sanger-sequenced animals, and together with data generated using a purely *in silico* approach to gap extension using whole genome sequence data (M. D. Littlejohn et al., 2016), a variant panel comprising 17 variants representing 556 sequenced animals was used for imputation and association analysis (Littlejohn et al., 2016.) Although a subset of these variants were significantly associated with both milk fat percentage and *MGST1* gene expression (Littlejohn et al. 2016), they were not the most strongly associated variants. Although this is in effect a negative result, identification of the full complement of variation at a QTL is required to definitively sort candidate variants from those that are causal, and is particularly important in the context of a QTL that is underpinned by a gene expression-based mechanism, where non-coding variation likely underlies this effect.

4.5.2 Characterising the CNV breakpoint sequences and genotyping a candidate variant

Inspection of WGS alignments representing *MGST1* revealed a CNV that, given its proximity to the TSS, was a strong candidate variant for the expression and milk composition QTLs. To test this hypothesis, PCR and Sanger sequencing was undertaken to characterise the CNV breakpoint sequences and genotype the CNV in 13 WGS animals. The approach was taken to generate PCR products encompassing the breakpoints of the CNV, with the presence and/or absence of amplification products indicative of the number of copies of the CNV in each animal. Notably, all 13 PCR-based genotype calls matched the WGS-derived genotypes called by CNVnator software, providing reassurance on the use of the bioinformatics method to generate genotypes for the imputation reference population (Littlejohn et al 2016).

To determine the exact size and structure of the CNV, Sanger sequencing of the PCR products was conducted for two sires representing each genotype class (i.e. CN0, CN1, and CN2). This revealed an 8,202 bp deletion, which agreed with the size apparent from WGS information. However, the presence of a 129 bp ART2A RTE-BovB repeat fragment at each breakpoint prevented the precise determination of the CNV breakpoints. Instead, I was only able to determine that the CNV breakpoints resided within 'sliding windows' at chr5:93951990-93952118 and chr5:93960192-93960320 (UMD3.1).

The presence of the ART2A RTE-BovB repeat sequences flanking the CNV provides potential insight into how the polymorphism may have occurred. These ART2A RTE-BovB elements are a type of long interspersed nuclear element repeat sequence, that are proposed to be the result of horizontal transfer, and may represent up to 25% of the bovine genome (Adelson, Raison, & Edgar, 2009). It was hypothesised that the insertion of CNV copies close to the promoter of *MGST1* by this retrotransposon could alter *MGST1* expression as similar repeat elements residing in regulatory elements have been shown to be influence gene expression to contribute to complex traits in humans (Gymrek et al., 2016). However, the subsequent association analysis (conducted outside the remit of this thesis; see Littlejohn et al., 2016) suggested the CNV was unlikely to be the source of the expression and milk fat percentage QTL signals.

4.5.3 Milk fat percentage QTL genotype does not influence novel milk composition phenotypes or gene expression networks

Despite confirmation of the role of *MGST1* in milk fat composition regulation (as evidenced by the collocating, co-segregating *MGST1 cis*-eQTL; Littlejohn et al., 2016), the cellular explanation for this effect is unclear. To investigate the potential mechanism through which *MGST1* might be operating, impacts on intermediate phenotypes related to milk fat percentage were assessed. It was hypothesised that by de-convoluting the individual fatty acids that make up fats in milk, some further clue as to how *MGST1* was impacting fat production might be resolved (for e.g. observation of preferential modulation of long-chain versus short chain fatty acids). However, this analysis did not reveal any significant associations between the QTL tag-SNP (chr5:93945738) and the proportion of individual milk fatty acids.

One possible reason for the lack of association between the *MGST1* QTL tag-SNP and the proportions of individual fatty acids in this analysis is the low allele frequency for this marker in the FJXB population. Six F1 sires, none of which carried the alternative allele for this marker, were used to produce the F2 cows used in this analysis. As these sires contributed half of the alleles in these animals, this effectively halved the minor allele frequency (MAF) of this QTL. Combined with the already limited animal numbers, this low MAF considerably reduced the statistical power of this analysis to resolve an effect at this locus.

In a similar attempt to provide clues as to the functional roles of *MGST1*, genome-wide gene expression analyses were performed to look for co-associated gene networks in the mammary gland. It was hypothesised that other genes co-expressed by milk composition/*MGST1* eQTL genotype might give an indication of the pathways involved; however no other significantly differentially expressed genes were identified. This suggests that we were either underpowered to detect *trans*-eQTL effects, or that *MGST1* is operating at the terminal end of any possible gene expression networks. Given that *trans*-eQTLs imply molecular interactions, the power to resolve such effects is limited compared to detecting *cis* effects (Mackay et al., 2009). The testing of many thousands of genes also carries a multiple-testing burden in such genome-wide analyses (Westra et al., 2013). This is likely also true of the mammary RNAseq dataset, despite the data contributing to the genetic and functional characterisation of many *cis*-loci in this thesis (refer to Chapter 4, 5, 6, and 7).

It bears mentioning that two genes: *ZNF593* (chr2:127,523,710-127,524,790) and *DDX27* (chr13:78,054,502-78,071,705), were identified as differentially expressed in conjunction with *MGST1* QTL genotype prior to adjusting for multiple testing. *ZNF593* encodes zinc finger protein 593 which is predicted to function as an RNA-binding protein (P. L. Hayes, Lytle, Volkman, & Peterson, 2008). It is also thought to negatively modulate the DNA binding activity of Oct-2, and has a C2H2-like fold, which are extremely common in mammalian transcription factors (Terunuma, Shiba, & Noda, 1997). Additionally, *DDX27* encodes DEAD-Box Helicase 27, which is a component of the ribosomal RNA (rRNA) processing machinery (Kellner et al., 2015). Unfortunately, these functions are far too broad to provide clues as to the likely mode of *MGST1* milk composition regulation, and given the lack of statistical support for the genes, do not represent compelling candidates for further investigation.

4.5.4 Limitations and future directions

Despite extensive investigation at the chromosome 5 milk fat composition locus, and near unequivocal demonstration of the involvement of *MGST1* in this QTL, we were unable to definitively identify the causative variant(s) responsible. This is a reflection of the difficulty in 'proving' causality for non-coding sequence variants, where additional annotation resources, or direct functional testing, are required to differentiate clusters of variants in strong LD. Based on the identification of the *cis*-eQTL at this locus, potential functional experiments could include the use of CRISPR-Cas9 genome editing to introduce alternative, candidate alleles in mammary cells and measure their effects on gene expression (leveraging the work conducted in Chapter 3).

4.6 Summary and conclusions

In *Bos taurus* dairy cattle in NZ and abroad, a chromosome 5 locus at 93.9 Mbp has a large influence on milk fat percentage and other milk composition phenotypes. Work reported here has formed part of a detailed analysis of this locus. I report investigation of several, intractable, candidate causative variants for the QTL, including an 8.2 kb deletion that makes a near ideal functional candidate mutation. Additional datasets and approaches were also applied to attempt to shed light on the cellular mechanism of the QTL. These approaches did not definitively highlight a single causative variant, or provide insight as to the pathways involved, though taken together, contribute part of a complex story of how a gene with no previously demonstrated role in lactation can have major effects on milk phenotypes. In this respect, the current work forms a baseline for further mechanistic investigation of the role of *MGST1* in lactation, and future functional studies to directly test the non-coding candidate variants identified should help resolve the precise genetic elements responsible for these effects.

Chapter 5: *DGAT1* K232A; A familiar milk fat mutation with a new mechanism

5.1 Overview

In *Bos taurus*, a K232A amino acid substitution in the diacylglycerol O-acyltransferase 1 (*DGAT1*) gene has a major pleiotropic influence on milk composition traits, the most substantial being its impact on milk fat percentage (Grisart, Coppieters, Farnir, et al., 2002; Schennink et al., 2007). This lysine to alanine amino acid substitution results from an AA to GC dinucleotide substitution in exon eight of *DGAT1*, and likely constitutes the most widely studied and validated variant in association analyses of bovine milk composition (initially described by Grisart et al., 2002, with >800 Google Scholar citations to date). The *DGAT1* gene encodes an enzyme responsible for catalysing the terminal reaction in the mammary triglyceride synthesis pathway (Mayorek, Grinstein, & Bar-tana, 1989), and a paper by Grisart et al. (2004) has demonstrated that the *DGAT1* K allele synthesises more triglycerides *in vitro* when compared to the A allele. Aside from the *DGAT1* K232A mutation, an additional polymorphism 5′ of the transcription start site of the gene has also been shown to associate with milk fat percentage. This variant, a variable number tandem repeat (VNTR) expansion, is hypothesised to increase the number of putative transcription factor binding sites, and stimulate an increase in *DGAT1* expression (Kuhn et al., 2004). However, the functional testing of this VNTR variant was unable to show any differences in *DGAT1* expression between QTL genotypes in cell culture (Fürbass, Winter, Fries, & Kühn, 2006). This finding largely put the competing, gene expression-based hypothesis of the *DGAT1* milk fat effect to rest, with enzymatic differences deriving from the K232A mutation widely considered as the underlying mechanism.

Since these initial analyses, further functional characterisation of the K232A mutation has been largely absent. Having generated a large, mammary RNAseq dataset however, we had the opportunity to re-examine this locus for potential regulatory effects impacting *DGAT1*. This chapter describes the detailed investigation of the *DGAT1* locus and demonstrates, for the first time, a strong expression-QTL (eQTL) in the mammary gland. Importantly, the expression of *DGAT1* transcripts is associated with K232A genotype (and thus milk fat percentage). Based on this observation, functional investigation was also conducted to characterise a possible mechanism by which the K232A variant mediates this effect.

5.2 General aim

Investigate the gene expression effects at the *DGAT1* locus in the bovine lactating mammary gland.

5.3 Specific aims

1. Conduct eQTL analysis at the *DGAT1* locus in the mammary RNAseq dataset.
2. Quantify the splicing efficiency of multiple *DGAT1* intron/exon junctions in the mammary RNAseq dataset.
3. Conduct cell-based functional testing of the *DGAT1* K232A influence on the splicing efficiency of multiple junctions in *DGAT1*.

5.4 Methods

5.4.1 Animal cohort, genotypes and gene expression phenotype

The work described in this chapter was conducted using the high-depth mammary RNAseq dataset which was detailed in General Methods (Chapter 2). For eQTL analysis, the RNAseq reads mapping to *DGAT1* were transformed as described in General Methods.

The RNAseq animals were genotyped with the Illumina Bovine HD BeadChip. The 115 SNPs in the 1 Mbp interval centred on *DGAT1* K232A (chr14:1302265-2302265) from this panel were used in this chapter. In addition to these markers, RNAseq derived genotypes for K232A (chr14:1802265G>A SNP) were also included in the analyses, as this variant is the first base of the *DGAT1* K232A MNP (hereafter referred to as K232A). These genotypes were extracted using samtools (version 0.1.19), with missing genotypes called by manual interrogation of RNAseq reads overlapping the variant.

In addition to these genotypes, 3128 imputed whole-genome sequence (WGS) derived variants in the 1 Mbp interval of interest were used in this chapter. These markers were imputed into the RNAseq animals using a reference population of 556 animals as described in General Methods.

All genotypes were recoded using PLINK2 (version 1.90b2c; Chang et al., 2015) to 0, 1 or 2 to represent the number of alternative alleles for each marker (i.e. 0, 1, and 2 to represent the homozygous reference, heterozygous, and homozygous alternative genotypes, respectively).

5.4.2 Association analysis of *DGAT1* expression

Associations between K232A and the 115 Bovine HD SNPs in the 1 Mbp interval surrounding *DGAT1* K232A and *DGAT1* expression were quantified using pedigree-based mixed models in ASReml-R (A R Gilmour et al., 2009; Arthur R. Gilmour et al., 1995). Additionally, associations between the 3128 imputed WGS-derived variants in the 1 Mbp interval centred on *DGAT1* K232A and *DGAT1* expression were quantified in the same way. Each SNP was fitted in a separate sire-maternal grandsire single trait model, with SNP treated as a quantitative variable based on the number of copies of the alternative allele and variance components estimated in a restricted maximum-likelihood (REML) framework. Covariates for sequencing cohort, the proportions of NZ Holstein-Friesian ancestry, US Holstein-Friesian ancestry, Jersey ancestry and heterosis effects were also included in the models. The additive genetic variance, polygenic genetic variances, total genetic variance and phenotypic variance for *DGAT1* expression was calculated as described in General Methods. The proportion of phenotypic and genotypic variance explained by each SNP was calculated as described in General Methods.

5.4.3 Analysis of alternative splicing of *DGAT1* exon 8

The Bioconductor software package DEXSeq detects alternative splicing of transcripts in RNAseq data by using the genetic coordinates from all isoforms for a given gene to create 'counting bins' that correspond to one exon or part of an exon. The relative exon usage of an exon is defined as the number of transcripts from the gene that contain a particular exon divided by the number of all transcripts from the gene (Simon Anders, Reyes, & Huber, 2012).

This tool was used to investigate the alternative splicing of *DGAT1* exon 8, which has previously been demonstrated to be associated with *DGAT1* K232A genotype (Grisart et al., 2004). The genetic coordinates for the alternative form of *DGAT1* exon 8 were manually added to the Enseml gene transfer format (GTF) file which contained the information about the gene structure of all the genes in the reference genome (Table 5.1). Then, DEXSeq was used to count the relative usage of these two bins. A single factor ANOVA was used to test for the influence of K232A genotype on the alternative splicing of *DGAT1* exon 8.

Table 5.1 Genetic coordinates used by DEXSeq to define the reference and alternative forms of *DGAT1* exon 8

DGAT1 Exon 8	Start (bp)	End (bp)	Length (bp)
Alternative	1802251	1802259	9
Reference	1802260	1802325	66

5.4.4 Exon splicing enhancer motif search

The RESCUE-ESE analysis tool (http://genes.mit.edu/burgelab/rescue-ese/) annotates ESE hexamers in vertebrate exons and can be used to predict the potential consequences of sequence variation that disrupts or alters predicted ESEs (Fairbrother et al., 2004). This tool identifies ESEs in genomic sequences by searching for hexanucleotides that meet the following two criteria; they are significantly enriched in exons relative to introns, as well as at exons with non-consensus splice sites relative to consensus splice sites.

This tool was used to annotate *DGAT1* exon 8 for predicted ESEs. For both *DGAT1* alleles the first 23 nucleotides of the 5′ end *DGAT1* exon 8 were submitted to this tool (Table 5.2).

Table 5.2 The 5′ sequences of *DGAT1* exon 8 used as input for exon splicing enhancer motif analysis.
The chr14:1802265GC>AA MNP responsible for K232A is underlined in both alleles.

DGAT1 allele	Sequence
K allele	CTTTGGCAGGTAAGAAGGCCAAC
A allele	CTTTGGCAGGTAAGGCGGCCAAC

5.4.5 RNAseq-derived *DGAT1* splicing efficiency phenotypes and analysis

To investigate the influence of *DGAT1* K232A on *DGAT1* splicing efficiency, the number of reads mapping to each intron and exon of *DGAT1,* was determined by HTSeq 0.6.0 (S. Anders, Pyl, & Huber, 2015), with the intron and exon boundaries specified by the RefSeq annotation (NM_174693.2). The splicing efficiency phenotype for *DGAT1* intron 8 was calculated as the percentage of *DGAT1* RNAseq reads mapping to the intron. The splicing efficiency phenotypes for each individual RefSeq *DGAT1* junction were calculated as the ratio of exonic reads to intronic reads corresponding to the junction (of spliced and unspliced reads, respectively). Reads were considered exonic reads if they bridged the splicing junction i.e. mapped to the 3′ end of the preceding exon and the 5′ end of the following exon. Reads were considered intronic or unspliced if they mapped to the 3′ end of the preceding exon and through the intron-exon boundary into the intron.

To determine the effect of K232A on *DGAT1* splicing efficiency, these two scores were then used as phenotypes for association analysis using PLINK2, using the 115 Bovine HD SNPs and K232A genotypes derived from the RNAseq data. For each junction, data were analysed to include covariates for population structure and sequencing cohort using the linear model function in PLINK2. The covariates were used to remove the effects of genetic clusters in the data. Using the --genome, --cluster and --mds-plot functions in PLINK2, animals with similar genotypes were clustered together on the covariate axes, while animals with different genotypes will be separated from each other. Fitting the covariates adjusts the data for these genetic distances.

5.4.6 Functional testing of *K232A* influence on DGAT1 splicing efficiency

To test the effect of the K232A on *DGAT1* splicing efficiency *in vitro*, MAC-T cells (Huynh et al., 1991) were transfected with *DGAT1* mini-gene constructs containing either the K232 or the A232 allele. The *DGAT1* alleles were based on the reference sequence (Accession number AY065621) and were identical with the exception of the AA>GC MNP that causes the K232A amino acid substitution. The 5′ UTR was extended by 84 bp to represent the UTR apparent from mammary RNAseq data and the first two introns were removed due constraints on total insert size. Intron 1 is 3,616 bp and intron 2 is 1,943 bp, such that the collective 5,559 bp from these two introns is larger than the rest of the gene structure combined (which is 3,117 bp; Figure 5.1).

Plasmids containing the two *DGAT1* isoforms were generated by GenScript (New Jersey, USA) and single preparations were used for all experimental replicates reported here (see General Methods). Co-transfection of cells with pMAXGFP plasmid (Lonza) was conducted in a 1:1 ratio to provide a normalisation control for transfection efficiency.

Cells were plated in 24-well plates and grown for 24 hours in proliferation media to achieve approximately 70% confluency (see General Methods). For cell transfection, 0.5 μL Lipofectamine® LTX (Invitrogen) was gently mixed with 25 μL Opti-MEM reduced serum media (Invitrogen). Aliquots containing 375 ng of both DGAT1 and pMAXGFP plasmid DNA and 0.5 μL PLUS reagent were diluted in 25 μL Opti-MEM. The diluted plasmids were combined with the Lipofectamine® LTX, gently mixed and incubated at room temperature for 5 minutes, after which 50 μL transfection mix was added to each well. After 24 hours of incubation at 37^{0}C, the cells were visualised on a Nikon Ti-E inverted light microscope prior to RNA extraction. All experiments were repeated in triplicate in three separate cell preparations from passage numbers 9, 10, 11, and 12.

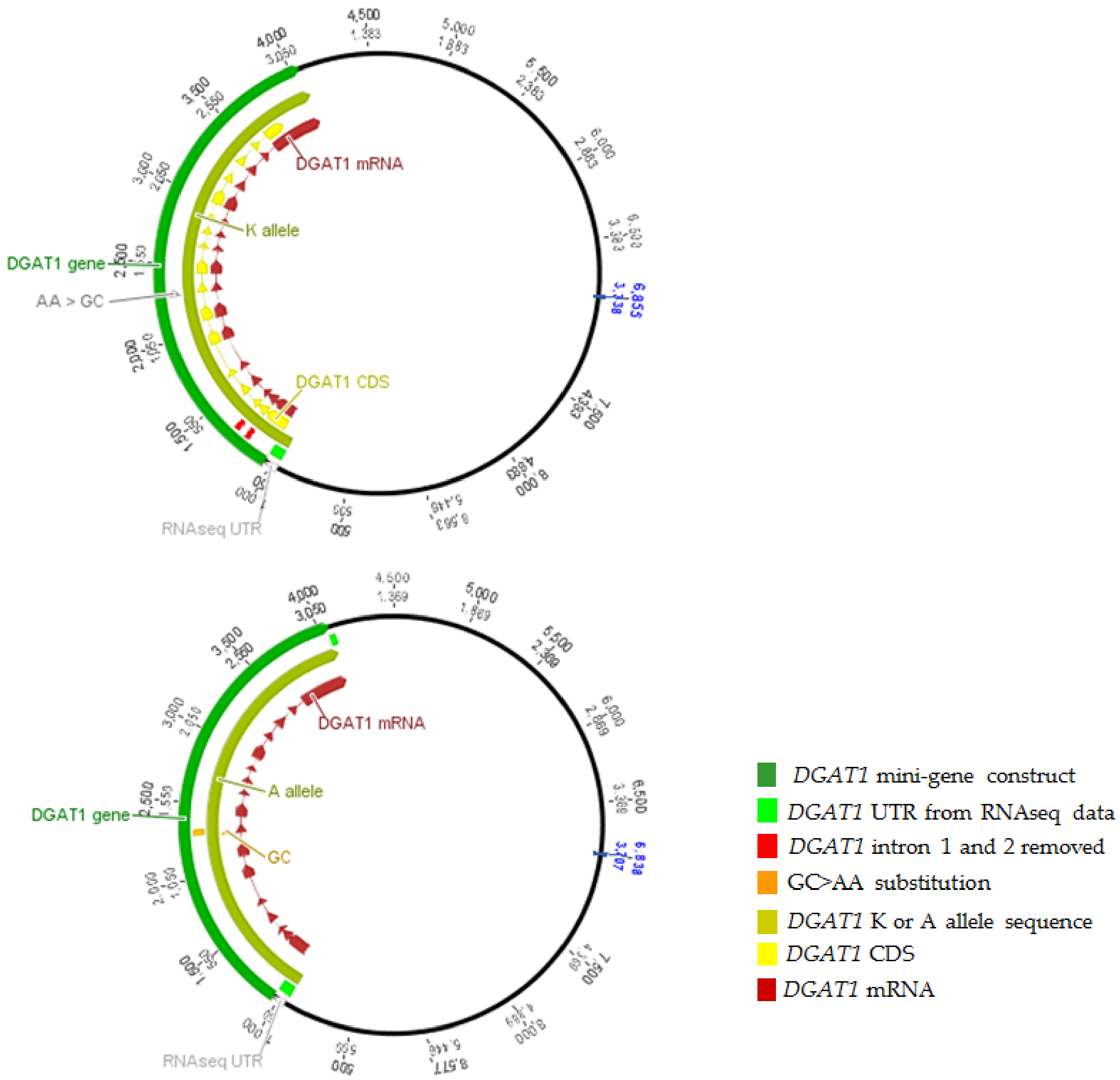


Figure 5.1 Schematic of the two *DGAT1* constructs inserted into pcDNA3.1
The only difference is the AA>GC MNP responsible for *DGAT1* K232A substitution which is indicated in the A allele construct (bottom).

5.4.6.1 RNA extraction and cDNA synthesis

RNA was extracted from each well of a 24-well plate using a TRIzol-based protocol and was subjected to two sequential DNase treatments before quantification as described in General Methods.

Following DNase treatment, cDNA synthesis was performed by reverse transcription-PCR (RT-PCR) using 2.5 µg of RNA as input for each 20 µL reaction. Complementary DNA (cDNA) was diluted 1:10 in Ultra-Pure water (Invitrogen) and used immediately for qPCR or stored at -20°C. Serial 5x cDNA dilutions were used to generate standard curves for each real-time PCR assay by pooling 4 µL from each experimental sample.

5.4.6.1 Real-time PCR experiments

Real-time PCR reactions were carried out in 10 µL volumes in 384-well plate format using standardised PCR cycling conditions and LightCycler480® Universal Probe System (described in General Methods).

Eukaryotic translation initiation factor 3K (*EIF3K*) was used as an endogenous control gene for normalisation of gene expression (Grala et al., 2011). In addition, an assay was designed for the pMAXGFP plasmid as a further control to normalise for transfection efficiency (Table 5.3).

To quantify splicing efficiency at the intron-exon junctions in *DGAT1*, assays were designed using Universal Probe Library and Primer 3 to generate two assays for each junction. These assays were designed such that they had a common primer (either forward or reverse), and used the same probe. The expression of spliced mRNA transcripts was measured using primers that bound to the two exons adjacent to the intron/exon junction. The expression of unspliced mRNA transcripts was measured using a primer that bound to one of the adjacent exons and a primer that bound across the intron/exon junction (Table 5.3 and Figure 5.2).

5.4.7 Assessment of relative expression of spliced and unspliced transcripts using qPCR

Relative quantification of spliced and unspliced transcripts was carried out as described in General Methods. Briefly, the average expression of each transcript across triplicate wells was calculated relative to the geometric mean of expression for the reference gene assays for each sample.

The average expression of the spliced transcripts was divided by the unspliced transcripts to get the splicing ratio for each junction for each sample. Student's t-test was used to determine the statistical significance of differences in the splicing ratio and mean spliced and unspliced transcript expression for each junction between the two alleles.

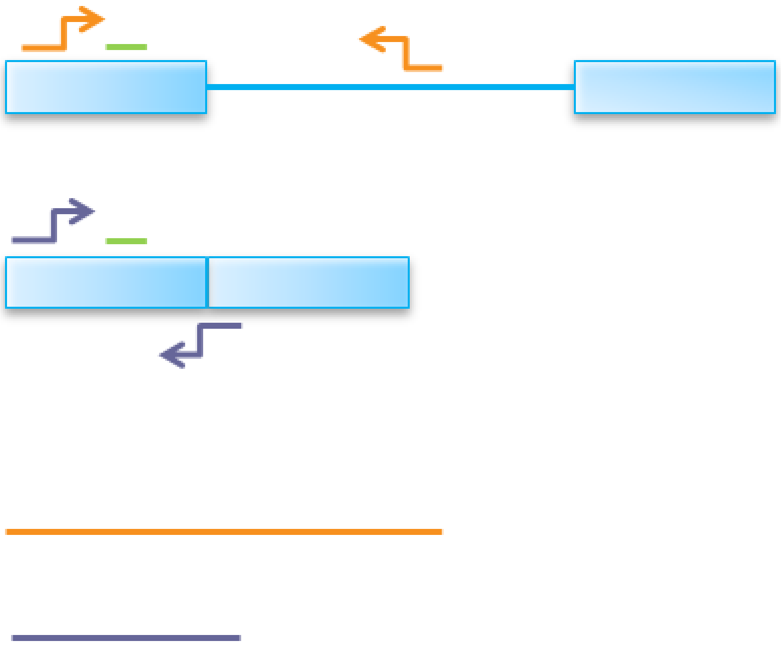


Figure 5.2 Schematic of the two RT-qPCR assays for each junction in *DGAT1*
The blue boxes represent exons while the blue line represents the intron. The green line represents the probe, while the orange and purple arrows represent the primers for unspliced and spliced mRNA transcripts, respectively. The first assay quantifies the intron containing pre-mRNA transcripts (orange) while the second assay quantifies the spliced mRNA transcripts (purple). The ratio of mRNA:pre-mRNA transcripts is used to generate a splicing efficiency phenotype for each junction.

Table 5.3 Primer sequences and assay design for RT-qPCR of *DGAT1* intron 3, 5, 7, and 13 junctions.

***DGAT1* Junction**	**Probe**	**Primers**	
3	9	F1	ACTACCGTGGCATCCTGAAT
		F2	CAGTTCTGACAGTGGCTTCAG
		R1	CACCAGGATGCCATACTTGAT
5	66	F1	CGTTCCAGGTGGAGAAGC
		F2	GTGGGAGCTCTGACGGAG
		R1	GAATGGTGGCCAGGTTGA
7	57	F1	TCAAGCTGTTCTCCTACCGG
		R1	CGAGGCAGCCCTCACCAG
		R2	CTTACCTGCCAAAGCAGC
13	71	F1	CACTTCTACAAGCCCATGCTC
		R1	CTTCACCGGCATGATGGC
		R2	CACCAGGTACTCGTGGAAGAA
Control Genes			
EIF3K	1	F	AAGTTGCTCAAGGGGATCG
		R	TTGGCCTGTGTCTCCACATA
GFP	5	F	CGACGGCGGCTACTACAG
		R	GTGGATGGCGCTCTTGAA

5.5 Results

5.5.1 *DGAT1* K232A associates with *DGAT1* transcript abundance in the lactating mammary gland

Association analysis between mammary *DGAT1* expression and the 115 SNPs from the BovineHD panel and K232A variant revealed a significant eQTL for *DGAT1*. Curiously, K232A was one of the top associated variants (P=1.59x10^{-25}; Figure 5.3), explaining 29.7% of the phenotypic variance in mammary *DGAT1* expression (Table 5.4).

While K232A was significantly associated with *DGAT1* expression, the most highly associated marker for this signal was BovineHD1400000216, which is located at chr14: 1736599 (P=1.29x10^{-27}; Figure 5.3). This marker is highly correlated with K232A, with an R^2 value of 0.92 (Table 5.4). Notably, the milk fat percentage-increasing K allele was the same allele associated with increased *DGAT1* expression in this analysis. Those animals in the K allele genotype class had a mean transformed read count for *DGAT1* of 9.628 (±0.024) whereas those animals in the A allele genotype this value was 9.245 (±0.026). The heterozygous animals had a mean transformed read count of 9.436 (±0.019), intermediate between the two opposing genotype classes (Table 5.4). The frequency of the *DGAT1* K allele was 0.51 in the RNAseq population.

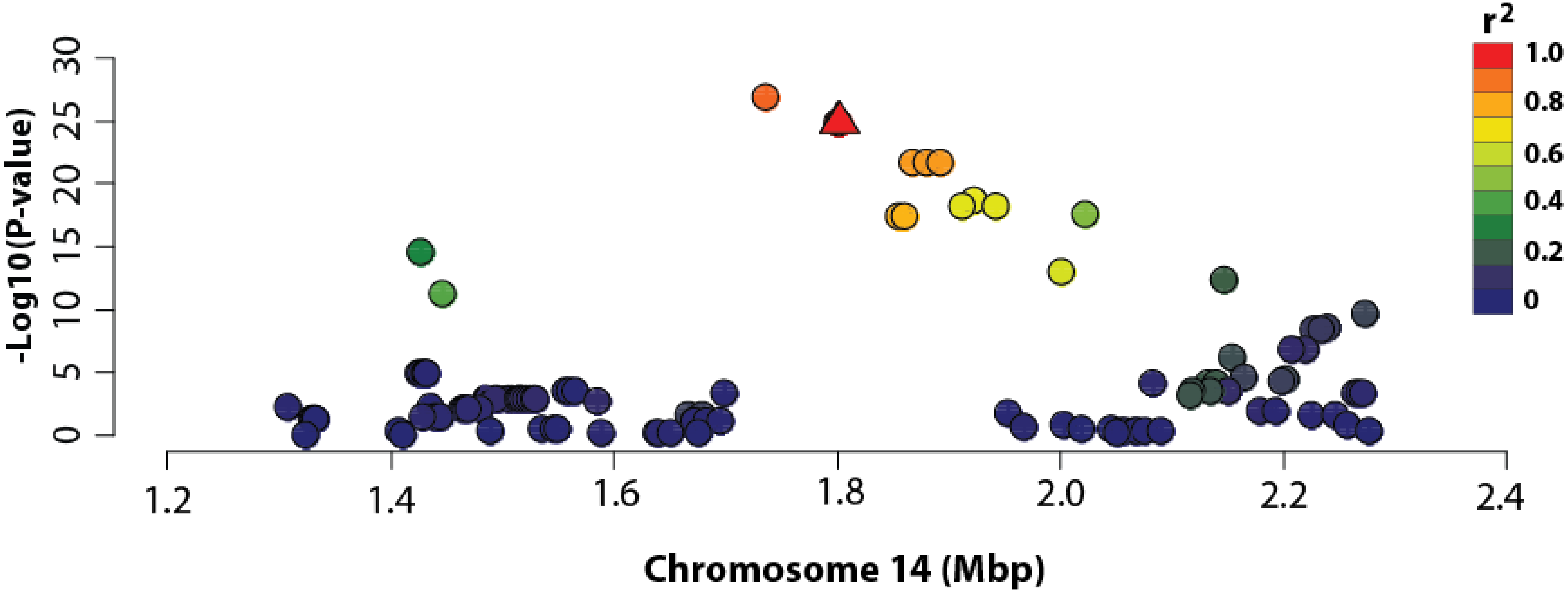


Figure 5.3 Expression QTL analysis at the *DGAT1* locus in bovine lactating mammary gland
The X-axis shows bp on chromosome 14, the Y-axis shows –log10 P-values of marker association for the 115 SNPs from the BovineHD panel and K232A in the 1 Mbp interval centred on *DGAT1* K232A. The *DGAT1* K232A marker is denoted as a triangle.

Table 5.4. Mammary *DGAT1* expression association statistics for the top 10 BovineHD variants
The positions of these SNP variants on chromosome 14 are indicated (Chr14 pos); with parameter estimates shown with standard errors in units of VST-transformed RNAseq read counts. The genetic and phenotypic variance explained by each SNP, along with parameter-adjusted means for each of the three genotypes classes is indicated. The linkage disequilibrium R^2 values for each SNP relative to K232A variant is shown, with the P-values indicated in the right most column.

			Adjusted means						
Variant	**Chr14 pos**	**Parameter Est**	**Geno 0**	**Geno 1**	**Geno 2**	**Pheno var**	**Geno var**	**R^2 with K232A**	**P-value**
BovineHD1400000216	1736599	0.196 (±0.016)	9.244	9.440	9.637	31.228	99.999	0.922	1.29x10^{-27}
ARS-BFGL-NGS-4939	1801116	0.192(±0.017)	9.245	9.436	9.628	29.714	99.999	1	1.59x10^{-25}
K232A	1802265	0.192(±0.017)	9.245	9.436	9.628	29.714	99.999	1	1.59x10^{-25}
BovineHD1400000243	1868636	0.178(±0.017)	9.247	9.425	9.603	25.645	99.999	0.832	2.08x10^{-22}
BovineHD1400000246	1880378	0.178(±0.017)	9.247	9.425	9.603	25.645	99.999	0.832	2.08x10^{-22}
BovineHD1400000249	1892559	0.178(±0.017)	9.247	9.425	9.603	25.645	99.999	0.832	2.08x10^{-22}
Hapmap52798-ss46526455	1923292	0.164(±0.017)	9.253	9.416	9.580	21.432	99.999	0.661	2.62x10^{-19}
BovineHD1400000251	1911696	0.162(±0.017)	9.255	9.416	9.580	20.938	99.999	0.650	6.32x10^{-19}
BovineHD1400000256	1943598	0.162(±0.017)	9.255	9.416	9.578	20.938	99.999	0.650	6.32x10^{-19}
BovineHD1400000276	2022413	-0.177(±0.019)	9.654	9.477	9.300	23.385	78.068	0.485	2.75x10^{-18}

5.5.2 *DGAT1* K232A associates with alternative splicing of *DGAT1* exon 8

Based on a previous report of the influence of *DGAT1* K232A on the generation of the alternative *DGAT1* isoform, I wondered if the alternative splicing of exon 8 could give rise to the observed eQTL in the RNAseq dataset. The relative usage of the reference and alternative exons was counted by DEXSeq, and used to create a *DGAT1* exon 8 alternative splicing ratio phenotype by dividing the reads the DEXSeq counts for the alternative exon by the DEXSeq counts the reference exon. Analysis of the splicing ratio phenotype in conjunction with K232A genotype revealed a significant difference in the ratio of reads splicing at the alternative splice site and the constituent splice site for exon 8 based on K232A genotype (P=4.60x10^{-06}; Table 5.5).

Notably, the *DGAT1* expression-increasing K allele was the same allele associated with increased alternative splicing of *DGAT1* exon 8. The animals in the K allele genotype class had an alternative splicing ratio for *DGAT1* exon 8 of 1.470 (±0.023) whereas those animals in the A allele genotype class had a mean ratio of 1.606 (±0.048). The heterozygous animals had a mean *DGAT1* exon 8 alternative splicing ratio of 1.562 (±0.036), intermediate between the two opposing genotype classes (Table 5.5). However, this is unlikely to explain the eQTL as the alternative isoform differs through the 'intronification' of the majority of exon 8, and is therefore expected to result in a reduction in reads mapping to the exons of the gene, versus the increase observed in the animals bearing the K allele.

Table 5.5 Alternative splicing of *DGAT1* exon 8 in RNAseq animals based on K232A genotype
A single factor anova was conducted to test the difference between the means of the three genotype classes

Allele	Animal N	Ratio	Standard Error	P-value
KK	98	1.470	0.023	
KA	189	1.562	0.036	4.60x10^{-06}
AA	91	1.606	0.048	

5.5.3 *DGAT1* K232A disrupts a putative consensus exon splice enhancer

Given the strong statistical association of the K232A variant with mammary *DGAT1* expression, we wondered whether this variant might be somehow functionally implicated in that effect. The mutation sits within 15 bp of the intron/exon boundary, highlighting other splicing-based mechanisms as a possible explanation for the observed eQTL. To examine this possibility, the genomic sequence surrounding K232A was examined for the presence of an exon splice enhancer (ESE) motif. This analysis was performed using the online tool RESCUE-ESE, and revealed the presence of two predicted ESE motifs in the 5′ end of *DGAT1* exon 8 which both overlap K232A (Figure 5.4).

Notably, the ESE motifs, AG<u>AA</u>GG and AAG<u>AA</u>G were located at chr14:1802263-1802268 and chr14:1802262-1802267 (UMD3.1 genome build) respectively, and were only encoded by the K allele. These ESEs were disrupted by the AA>GC MNP (the corresponding AA nucleotides for responsible for K232A are underlined in these motifs below). As such, this polymorphism removes these two ESE motifs when the 232A allele is present (GC), which is the same allele associated with decreased mammary *DGAT1* expression and milk fat percentage.

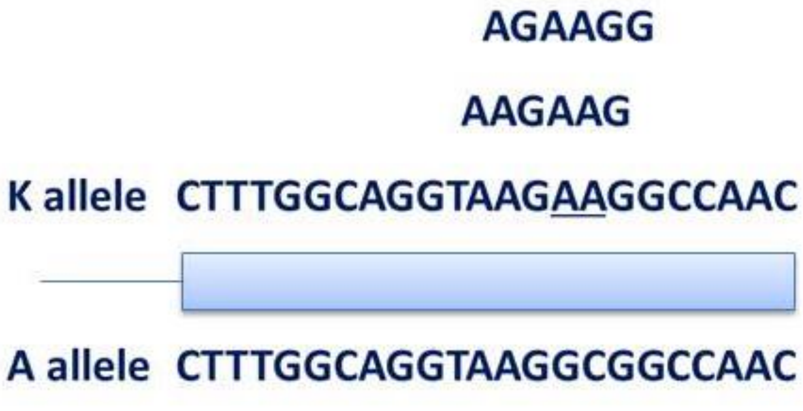

AGAAGG
AAGAAG
K allele CTTTGGCAGGTAAGAAGGCCAAC
A allele CTTTGGCAGGTAAGGCGGCCAAC

Figure 5.4. Schematic of the 5′ end of *DGAT1* exon 8 with ESE motifs overlapping the K232A amino acid substitution
The AA>GC MNP responsible for the K232A substitution is underlined in the *DGAT1* K allele.

5.5.4 *DGAT1* K232A associates with splicing efficiency of *DGAT1* intron 8

Having observed differential expression of *DGAT1* and the possible disruption of ESEs based on K232A genotype, the next step was to investigate if this variant associated with the efficiency of splicing at the neighbouring splicing junction. To this end, a splicing efficiency 'phenotype' was derived for intron 8, which involved the quantification of the percentage of total *DGAT1* RNAseq reads that mapped to this intron junction (see Methods).

Association analysis between the percentage of RNAseq reads mapping to intron 8 and K232A and the 115 SNPs from the BovineHD panel revealed a strong splicing efficiency effect, with K232A the most significantly associated variant ($P=6.29\times10^{-18}$, Figure 5.5; Table 5.6). Importantly, the direction of effect for the splicing efficiency effect was consistent with a mechanism that might explain the *DGAT1* eQTL, such that the K allele was associated with an increased percentage of completely spliced transcripts at the exon 8 junction. Of the three genotype classes, the KK animals had the highest percentage of completely spliced transcripts with only 1.16% of *DGAT1* RNAseq reads mapping to intron 8, while the AA animals had the lowest percentage of completely spliced transcripts with 1.94% of reads mapping to intron 8, with the heterozygous AK animals intermediate of the two (1.53% of reads mapping to intron 8; Table 5.7; Figure 5.5). As the K allele is also associated with increased *DGAT1* expression, this suggests splicing efficiency as a potential limiting mechanism for the production of fully spliced mRNA (Figure 5.5).

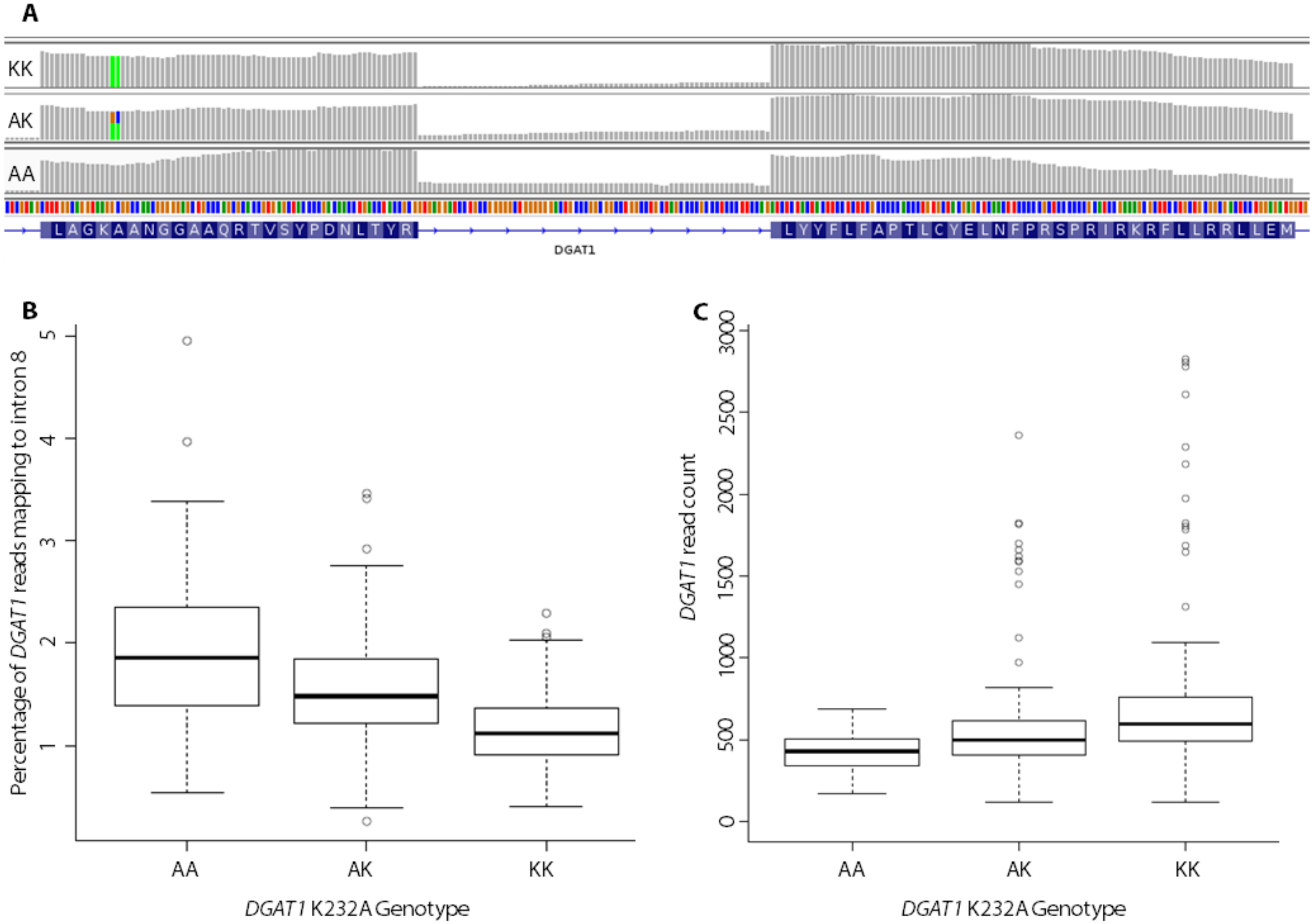


Figure 5.5 Splicing efficiency analysis of *DGAT1* intron 8 in bovine lactating mammary gland

A) Alignment of RNAseq reads at *DGAT1* intron 8 junction in animals of the three *DGAT1* K232A genotype classes. B) Boxplot of percentage *DGAT1* RNAseq reads mapping to *DGAT1* intron 8 in animals of the three *DGAT1* K232A genotype classes. C) Boxplot of *DGAT1* read count in animals of the three *DGAT1* K232A genotype classes in mammary RNAseq dataset.

Table 5.6 Association analysis for *DGAT1* intron 8 splicing efficiency
The positions of the top 10 SNP variants on chromosome 14 are indicated, with the P-value of association. Bonferroni threshold = P=4.31x10^{-4}

Marker	Position	P-value
K232A	1802265	5.95x10^{-20}
ARS-BFGL-NGS-4939	1801116	2.85x10^{-19}
BovineHD1400000243	1868636	1.12x10^{-18}
BovineHD1400000246	1880378	1.12x10^{-18}
BovineHD1400000249	1892559	1.12x10^{-18}
BovineHD1400000216	1736599	2.88x10^{-18}
BovineHD1400000239	1855090	4.28x10^{-18}
BovineHD1400000241	1861799	4.28x10^{-18}

Table 5.7 Percentage of mammary RNAseq reads mapping to *DGAT1* intron 8 in animals representing the three *DGAT1* K232A genotype classes.

Genotype	Animal N	% Reads mapping to *DGAT1* intron 8
AA	89	1.9489
AK	185	1.5330
KK	101	1.1612

5.5.5 *DGAT1* K232A associates with splicing efficiency of multiple junctions of *DGAT1*

Following the observation of association between K232A and the splicing efficiency of the intron 8 junction, we wondered whether the variant could have similar effects at other intron/exon junctions in *DGAT1*. To investigate this possibility, splicing efficiency phenotypes were generated from the ratio of RNAseq reads that were spliced vs. unspliced for all 14 junctions in *DGAT1* using read count data from the RNAseq dataset. The subsequent association analyses were conducted using the same subset of 116 markers as used for junction 8 revealing significant associations at six additional *DGAT1* junctions (Table 5.8; Figures 5.6-5.10). The junctions with effects were introns 1-3, 7, 8, 11, and 12, with the most significant being splicing of the intron 2 junction ($P=6.13x10^{-41}$; Table 5.8). For the seven significantly impacted junctions, K232A was the lead SNP for the association at 3 junctions (intron 2, 7 and 8). Interestingly, many of these junctions are physically distant to the K232A variant and the proposed ESE motif. For example, in base position terms, intron 2 is several kb from K232A yet was the most significant splicing efficiency effect in this analysis (Table 5.8).

Table 5.8 *DGAT1* junction splicing efficiency association statistics for the top BovineHD panel variants and K232A
This table presents the top SNP and the significance of its association for each of the *DGAT1* splicing junctions. In the fourth column, the rank of K232A amongst the markers is given, and the significance of its association is presented in the fifth columns well as the P-value of its association (in brackets) if not already presented in the previous columns. Bonferroni threshold = P=4.31x10^{-4}

***DGAT1* Junction**	**Top SNP**	**P-value**	**K232A Rank**	**K232A P-value**
Intron 1	BovineHD1400000216	6.47x10^{-21}	3	**1.75x10^{-17}**
Intron 2	K232A	6.13x10^{-41}	1	**6.13x10^{-41}**
Intron 3	ARS-BFGL-NGS-4939	8.58x10^{-18}	5	**1.42x10^{-17}**
Intron 4	BovineHD1400000271	0.02562	10	0.6644
Intron 5	BovineHD1400000216	0.00181	7	0.0628
Intron 6	BovineHD1400000175	0.245	40	1
Intron 7	K232A	4.01x10^{-16}	1	**4.01x10^{-16}**
Intron 8	K232A	6.29x10^{-18}	1	**6.29x10^{-18}**
Intron 9	BovineHD1400000180	0.6564	59	1
Intron 10	BovineHD1400000325	1	95	1
Intron 11	BovineHD1400000216	5.70x10^{-8}	3	**4.22x10^{-7}**
Intron 12	BovineHD1400000276	0.0017	12	0.1968
Intron 13	BovineHD1400000276	0.4633	6	1
Intron 14	BovineHD1400000276	6.62x10^{-5}	6	0.0025

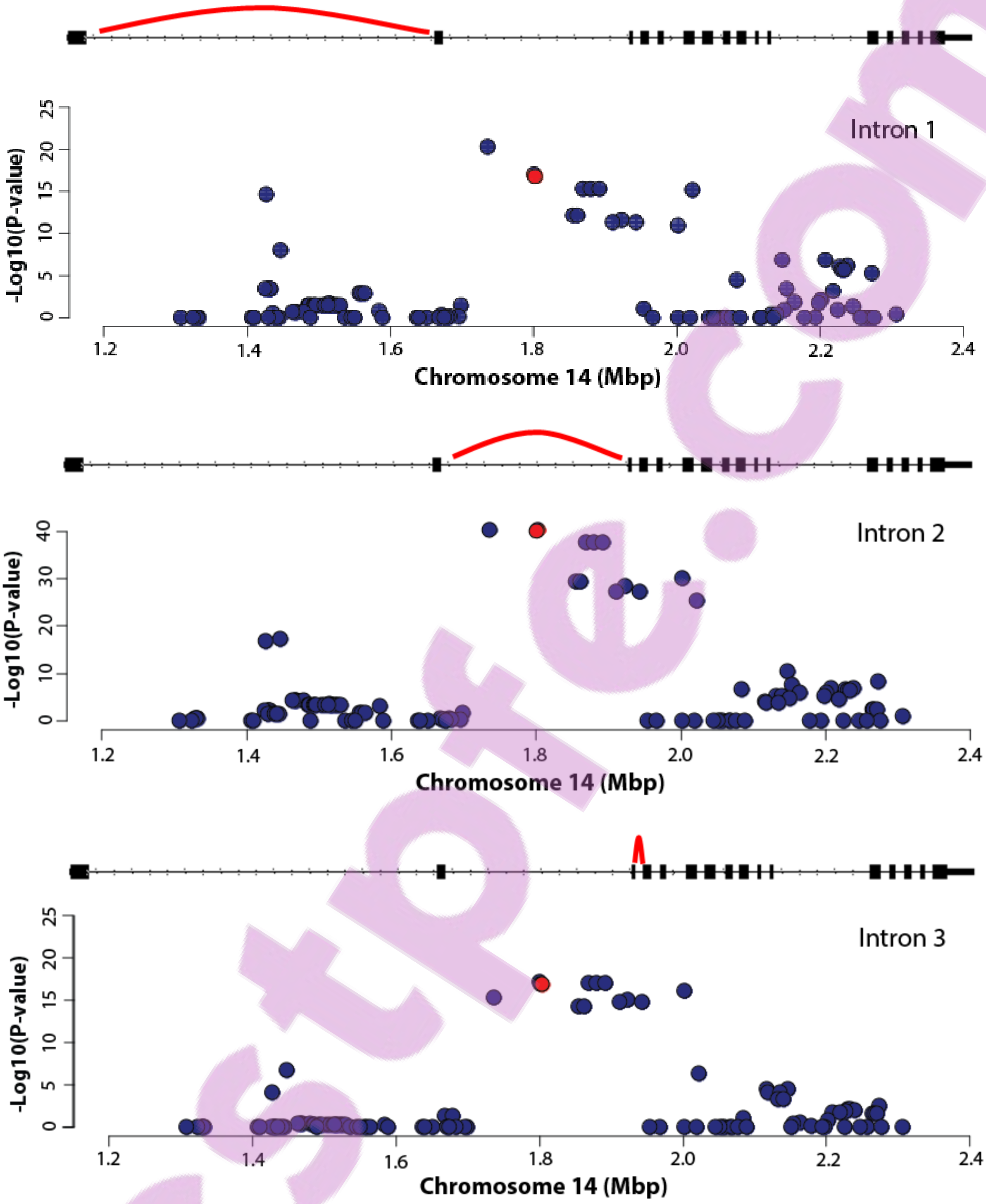


Figure 5.6 Manhattan plots for splicing efficiency analysis for *DGAT1* intron 1-3 junctions

The blue dots represent the 116 markers in the 1 Mbp interval centred on *DGAT1* K232A. *DGAT1* K232A is coloured red. The gene structure of *DGAT1* (not to scale) is shown above each Manhattan plot with the intron included in the analysis indicated by a red line.

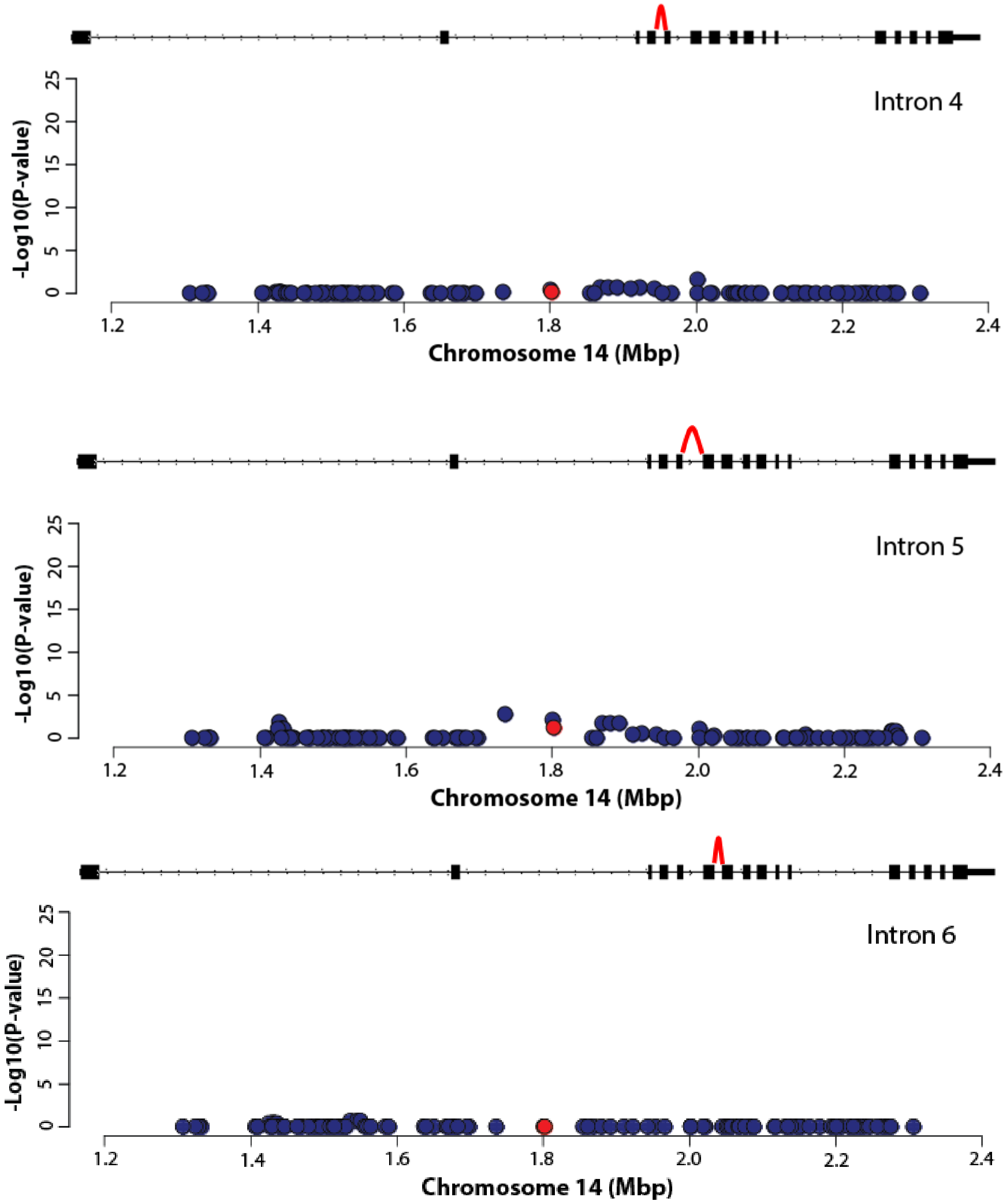


Figure 5.7 Manhattan plots for splicing efficiency analysis for *DGAT1* intron 4-6 junctions
The blue dots represent the 116 markers in the 1 Mbp interval surrounding *DGAT1* K232A. The K232A marker is coloured red. The gene structure of *DGAT1* (not to scale) is shown above each Manhattan plot with the intron included in the analysis indicated by a red line.

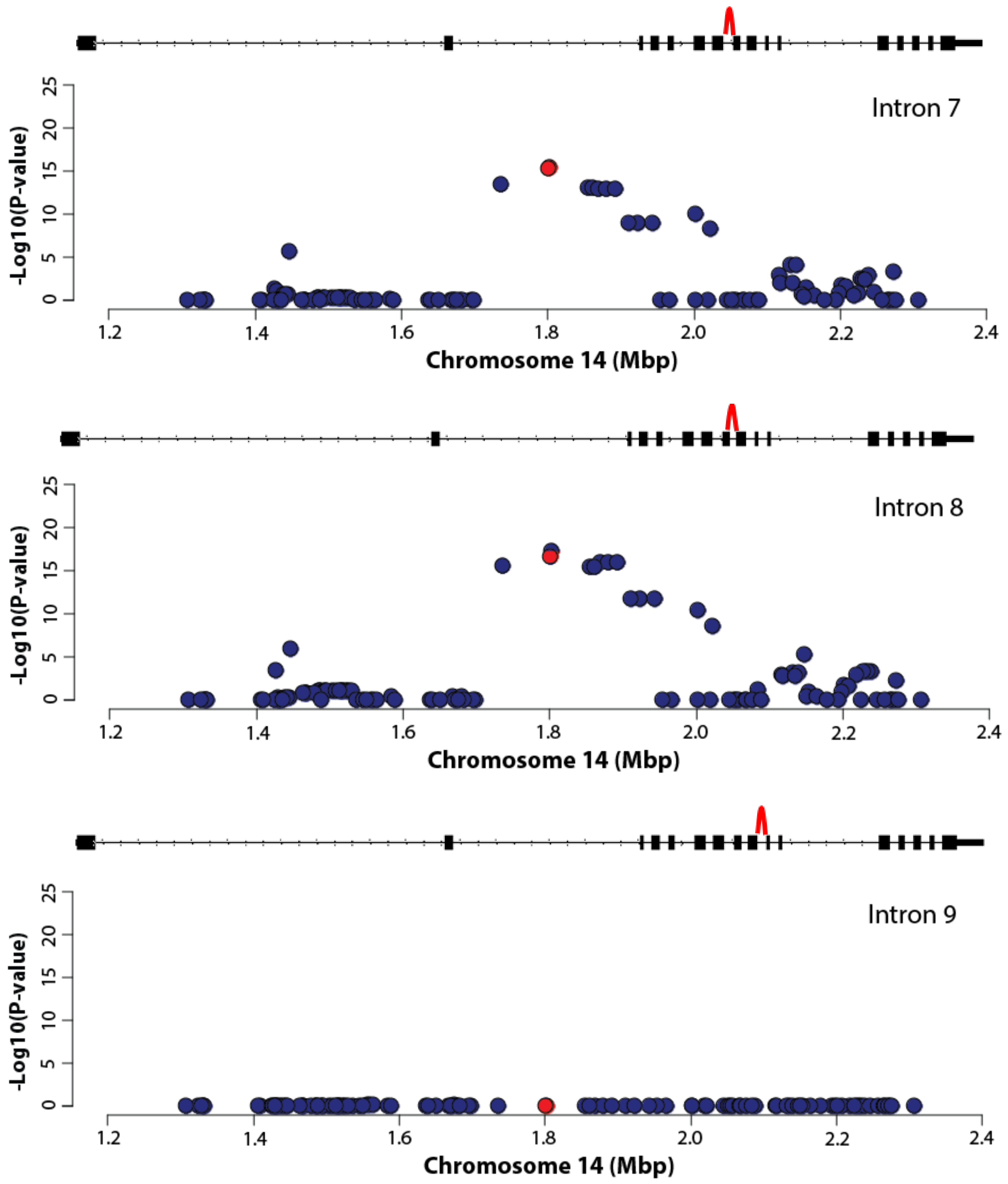


Figure 5.8 Manhattan plots for splicing efficiency analysis for *DGAT1* intron 7-9 junctions
The blue dots represent the 116 markers in the 1 Mbp interval surrounding *DGAT1* K232A. The K232A marker is coloured red. The gene structure of *DGAT1* (not to scale) is shown above each Manhattan plot with the intron included in the analysis indicated by a red line.

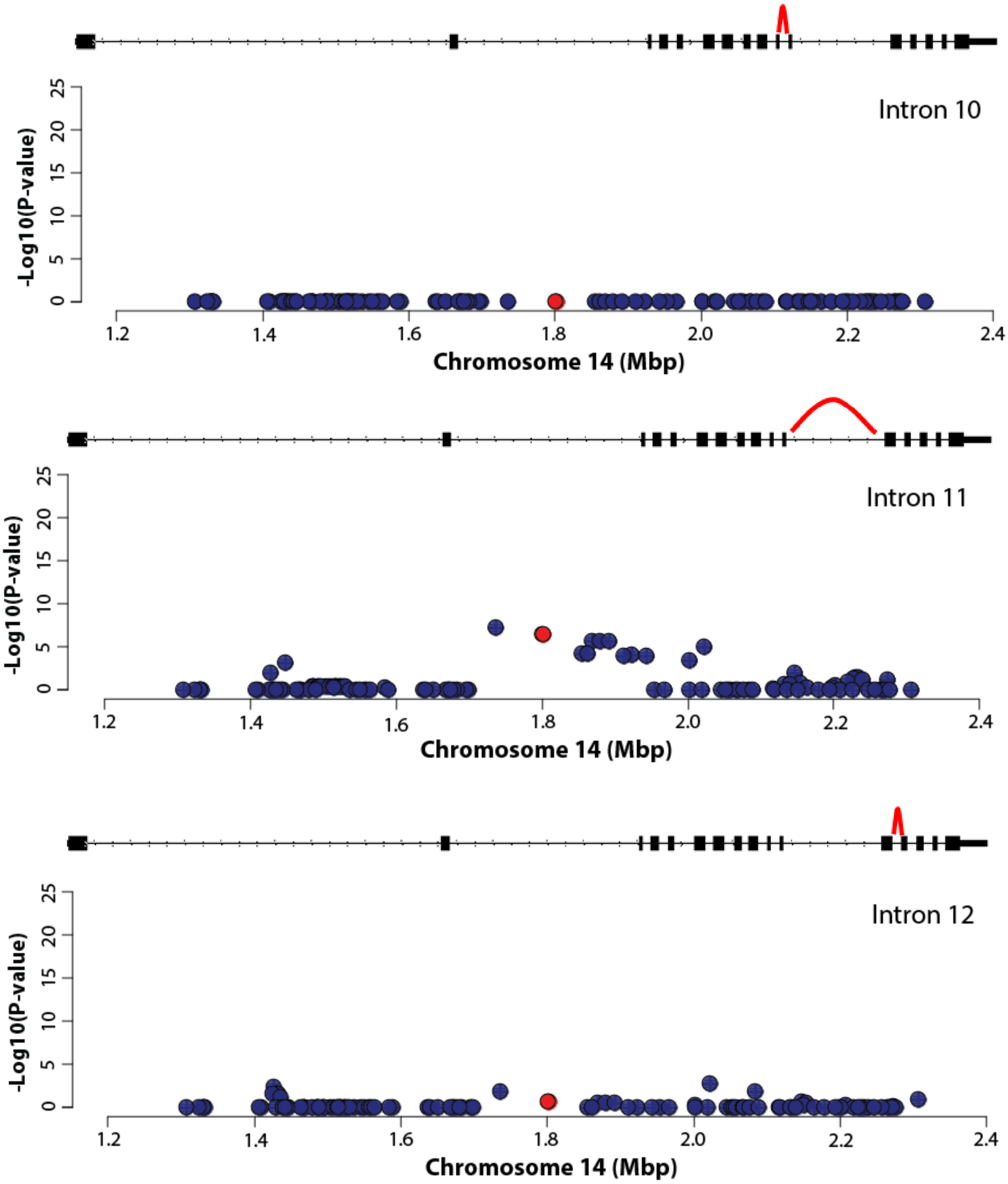


Figure 5.9 Manhattan plots for splicing efficiency analysis for *DGAT1* intron 10-12 junctions
The blue dots represent the 116 markers in the 1 Mbp interval surrounding *DGAT1* K232A. The K232A marker is coloured red. The gene structure of *DGAT1* (not to scale) is shown above each Manhattan plot with the intron included in the analysis indicated by a red line.

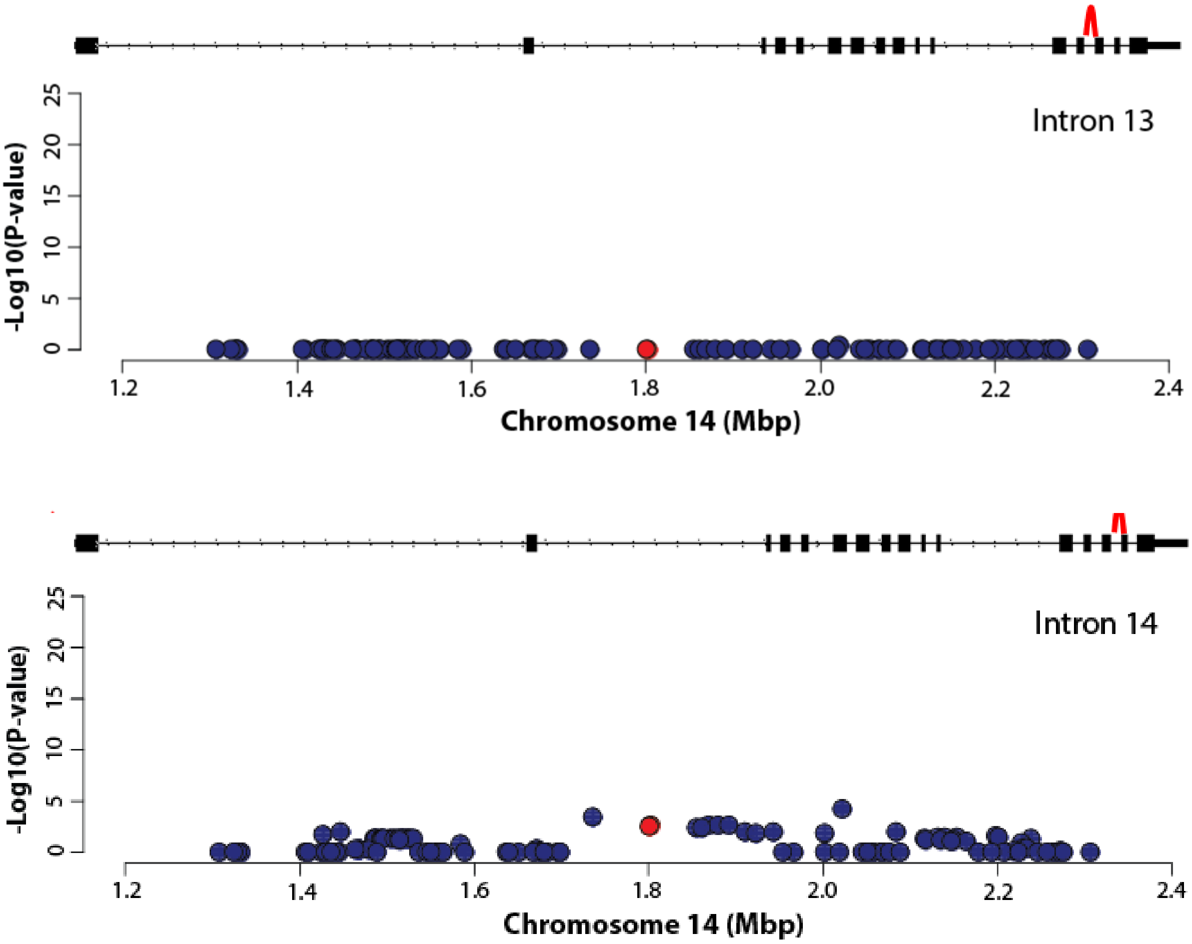


Figure 5.10 Manhattan plots for splicing efficiency analysis for *DGAT1* intron 13 and 14 junctions
The blue dots represent the 116 markers in the 1 Mbp interval surrounding *DGAT1* K232A. The K232A marker is coloured red. The gene structure of *DGAT1* (not to scale) is shown above each Manhattan plot with the intron included in the analysis indicated by a red line.

5.5.6 *DGAT1* K232A alters *DGAT1* splicing efficiency *in vitro*

While *DGAT1* K232A is associated with the expression of *DGAT1* and splicing efficiency of multiple *DGAT1* junctions, the observation of a higher ranking SNP-chip marker in the eQTL analysis presents an alternative, equally plausible hypothesis that that effect is driven by an unknown promoter or other *cis*-regulatory variant. To test the possibility of K232A simply being in LD with another regulatory mutation, cell-based experiments were undertaken to remove the two K232A transcript isoforms from their genomic context.

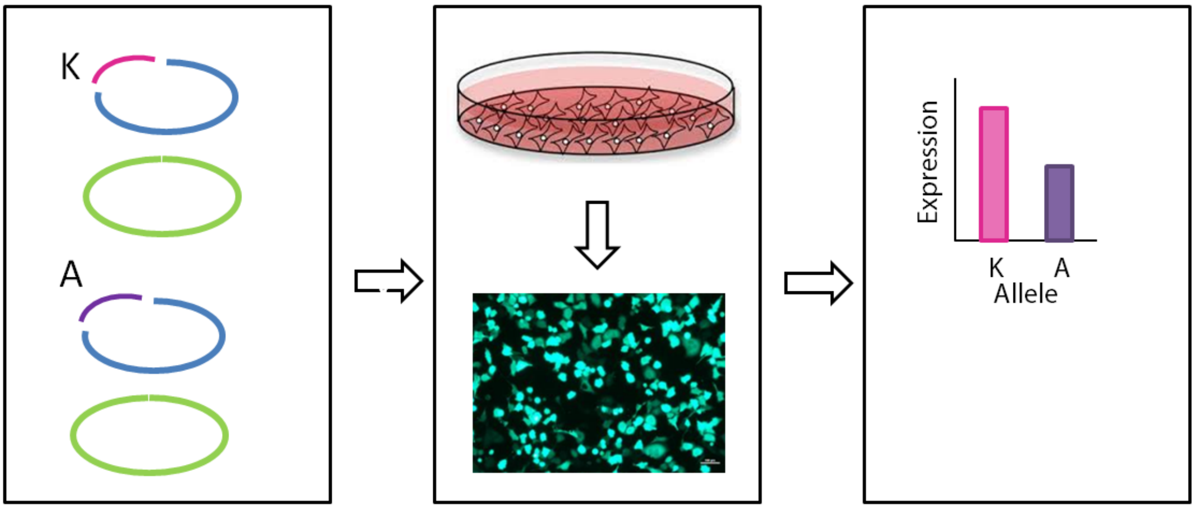


Figure 5.11 Schematic of the *in vitro DGAT1* splicing efficiency experiment
DGAT1 mini-gene constructs for the K or A allele (pink and purple, respectively) were synthesised in pcDNA3.1 plasmids (blue) and co-transfected with pMAXGFP plasmid into mammary cells (green; first panel). After 24 hours, transfection was checked by visualising GFP fluorescence (second panel) and RNA was extracted and used as input for cDNA synthesis and q-PCR measuring the expression of spliced and unspliced mRNA transcripts at four of the junctions in *DGAT1* (third panel).

Plasmids containing *DGAT1* mini-gene constructs for either the K or A allele were generated and transfected into a bovine mammary cell line. The relative expression of the two alleles was tested in cell culture, with the splicing efficiency phenotype generated from the relative expression of the spliced and unspliced transcripts as measured by qPCR (Figure 5.11). Twelve of 14 junctions were initially targeted for analysis, with four of these (introns 3, 5, 7 and 13) yielding robust intron and exon-targeting assays as required for association analysis of each junction.

Mammary cells transfected with the *DGAT1* K allele construct had a higher splicing ratio at the intron 3 and intron 7 junctions compared with those cells transfected with the A allele (P=9.37x10^{-4} and P=9.05x10^{-11}, respectively; Table 5.9; Figure 5.12 and 5.13). At the intron 7 junction there was also a corresponding increase in the mean expression of the spliced transcripts in the cells transfected with the K allele compared to those cells transfected with the A allele (P=0.009, respectively; Table 5.10; Figure 5.13). Additionally, there was also an increase in the mean expression of the spliced transcripts at intron 3 junction, however this difference was non-significant (P=0.015; Bonferroni threshold P=0.0125; Table 5.10; Figure 5.12). Conversely, there were no significant differences in the mean expression of the unspliced transcripts at these two junctions in cells transfected with either of the two alleles (P=0.376 and P=0.383, respectively; Table 5.11; Figure 5.12 and 5.13). As such, this splicing efficiency difference between the two alleles for both intron 3 and intron 7 junctions appears result in increased spliced mRNA expression, rather than increased expression *per se* as there was no concomitant increase in unspliced transcript expression at either of these junctions in cell culture.

There was no significant difference in the splicing ratio of *DGAT1* intron 5 and 13 junctions with mammary cells transfected with either *DGAT1* allele (P=0.256 and P=0.497, respectively; Table 5.9; Figure 5.14 and 5.15). Further, there were no significance differences in the mean expression of the spliced mRNA transcripts at both the intron 5 and 13 junctions (P=0.097 and P=0.071, respectively; Table 5.10; Figure 5.14 and 5.15). There were also no significance differences in the mean expression of the unspliced mRNA transcripts at both the intron 5 and 13 junctions (P=0.069 and P=0.044, respectively; Table 5.11; Figure 5.14 and 5.15). Notably, the splice enhancement effects for the *DGAT1* intron 3 and intron 7 junctions were two of the junctions also highly associated with K232A in the RNAseq dataset (Table 5.7, Figures 5.5 and 5.7). Similarly, *DGAT1* intron 5 and intron 13 junctions were two of the junctions in the gene that did not exhibit a splicing efficiency effect associated with K232A in the RNAseq dataset (Table 5.7, Figures 5.6 and 5.9).

Table 5.9 Average splicing ratio at *DGAT1* intron 3, 5, 7 and 13 junctions in mammary cell culture for the 232K and 232A *DGAT1* plasmids. Bonferroni threshold P=0.0125

DGAT1 Junction	*DGAT1* K allele	*DGAT1* A Allele	P-value
Junction 3	0.253(±0.009)	0.173(±0.005)	$9.37\text{x}10^{-4}$
Junction 5	143.18(±10.23)	133.52(±9.68)	0.256
Junction 7	28.49(±2.77)	5.55(±0.74)	$9.05\text{x}10^{-11}$
Junction 13	92.90(±5.86)	90.92(±5.83)	0.411

Table 5.10 Average spliced mRNA expression at *DGAT1* intron 3, 5, 7 and 13 junctions in mammary cell culture for the 232K and 232A *DGAT1* plasmids. Bonferroni threshold P=0.0125

DGAT1 Junction	*DGAT1* K allele	*DGAT1* A Allele	P-value
Junction 3	0.111(±0.002)	0.082(±0.002)	0.015
Junction 5	0.946(±0.3238)	0.421(±0.160)	0.097
Junction 7	0.545(±0.117)	0.255(±0.047)	0.009
Junction 13	0.984(±0.219)	0.549(±0.233)	0.071

Table 5.11 Average unspliced mRNA expression at *DGAT1* intron 3, 5, 7 and 13 junctions in mammary cell culture for the 232K and 232A *DGAT1* plasmids. Bonferroni threshold P =0.0125

DGAT1 Junction	*DGAT1* K allele	*DGAT1* A Allele	P-value
Junction 3	0.092(±0.012)	0.0842(±0.227)	0.376
Junction 5	0.286(±0.0415)	0.180(±0.0350)	0.069
Junction 7	0.216(±0.030)	0.234(±0.059)	0.383
Junction 13	0.821(±0.180)	0.503(±0.124)	0.044

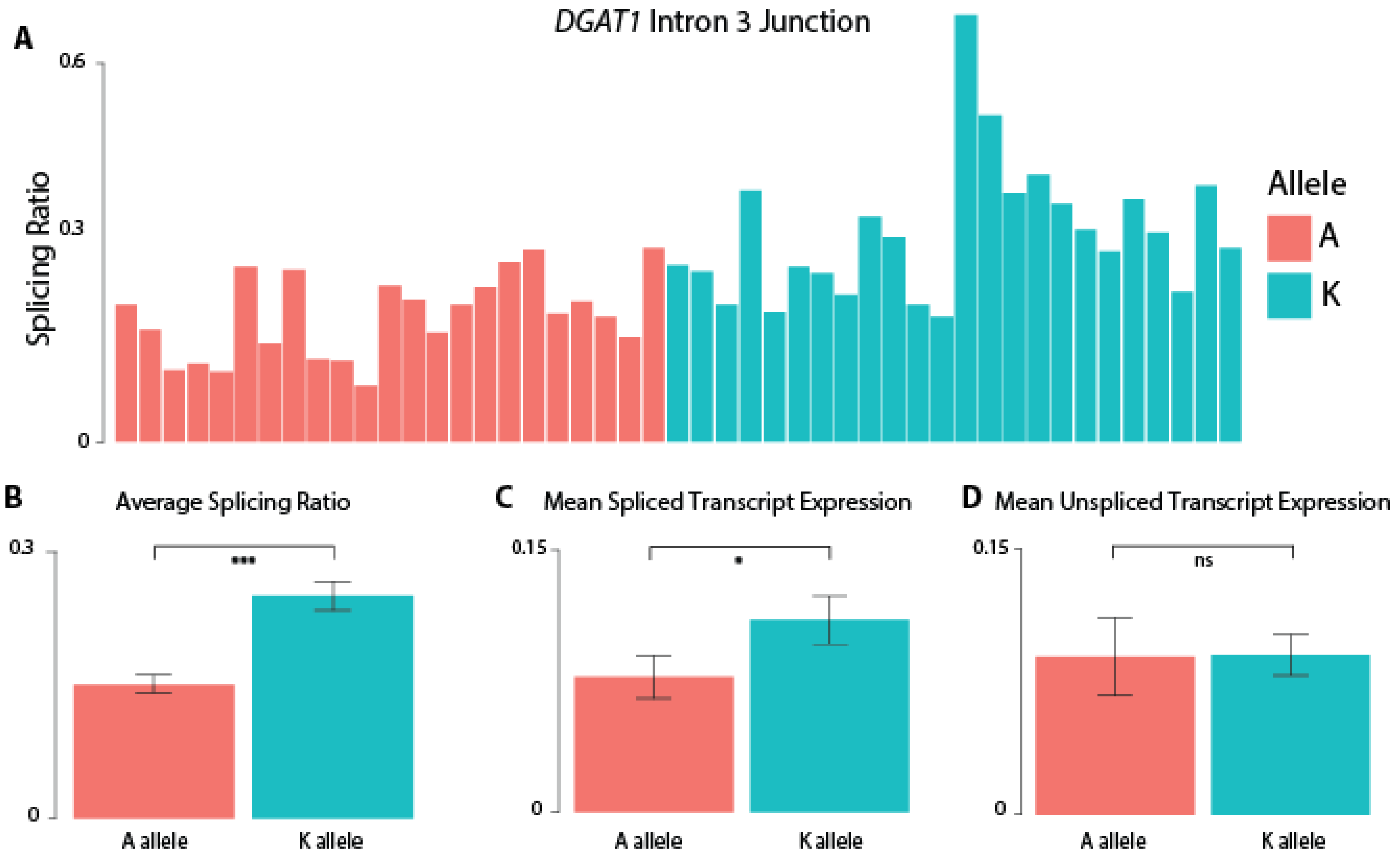


Figure 5.12 Cell-based functional testing of *DGAT1* K232A influence on splicing efficiency at the *DGAT1* intron 3 junction

Figure A represents the splicing ratio (spliced transcripts:unspliced transcripts) measured by qPCR in each of the individual replicates for the intron 3 junction. Figure B represents the average splicing ratio for the intron 3 junction for the two *DGAT1* K232A alleles. The error bars represent the standard deviation across all samples. Figure C and D represent the mean spliced and unspliced transcripts for the intron 3 junction, respectively. The error bars represent the standard error of the difference between means. *** = P ≥ 0.001, ** = P ≥ 0.01, * = P ≥ 0.05, ns = P > 0.05

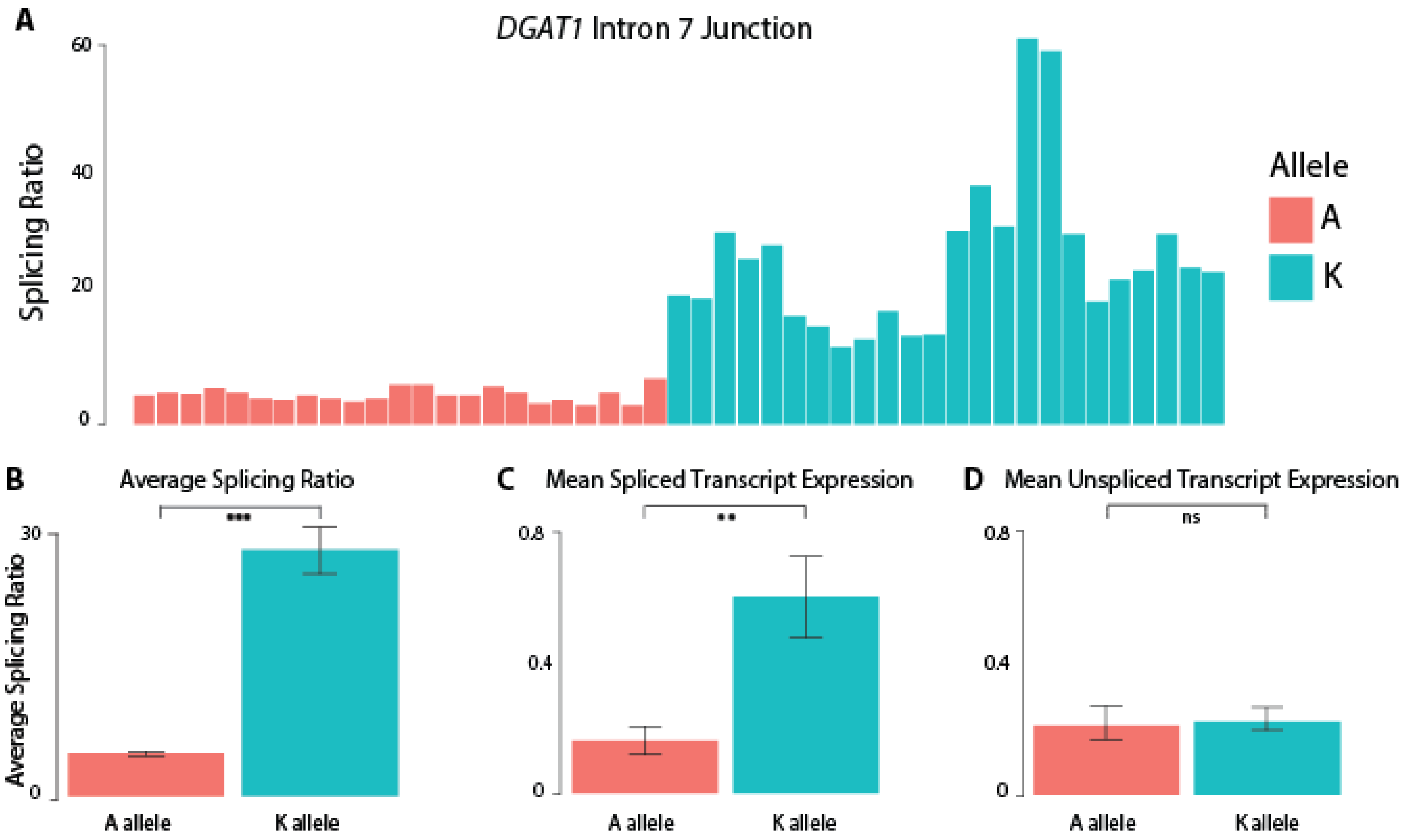


Figure 5.13 Cell-based functional testing of *DGAT1* K232A influence on splicing efficiency at the *DGAT1* intron 7 junction

Figure A represents the splicing ratio (spliced transcripts:unspliced transcripts) measured by qPCR in each of the individual replicates for the intron 7 junction. Figure B represents the average splicing ratio for the intron 7 junction for the two *DGAT1* K232A alleles. The error bars represent the standard deviation across all samples. Figure C and D represent the mean spliced and unspliced transcripts for the intron 7 junction, respectively. The error bars represent the standard error of the difference between means. *** = P ≥ 0.001, ** = P ≥ 0.01, * = P ≥ 0.05, ns = P > 0.05

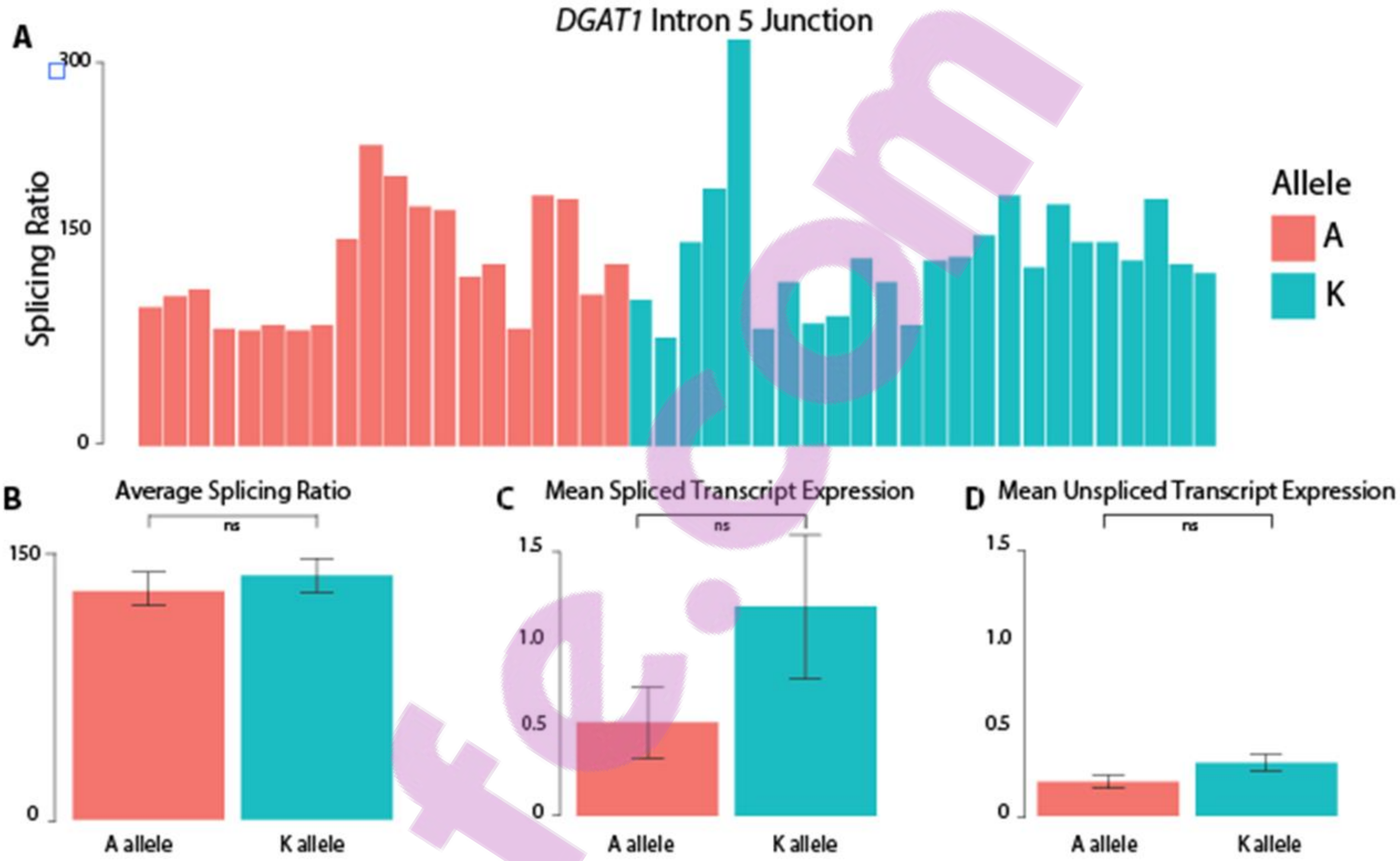


Figure 5.14 Cell-based functional testing of *DGAT1* K232A influence on splicing efficiency at the *DGAT1* intron 5 junction
Figure A represents the splicing ratio (spliced transcripts:unspliced transcripts) measured by qPCR in each of the individual replicates for the intron 5 junction. Figure B represents the average splicing ratio for the intron 5 junction for the two *DGAT1* K232A alleles. The error bars represent the standard deviation across all samples. Figure C and D represent the mean spliced and unspliced transcripts for the intron 5 junction, respectively. The error bars represent the standard error of the difference between means. *** = P ≥ 0.001, ** = P ≥ 0.01, * = P ≥ 0.05, ns = P > 0.05

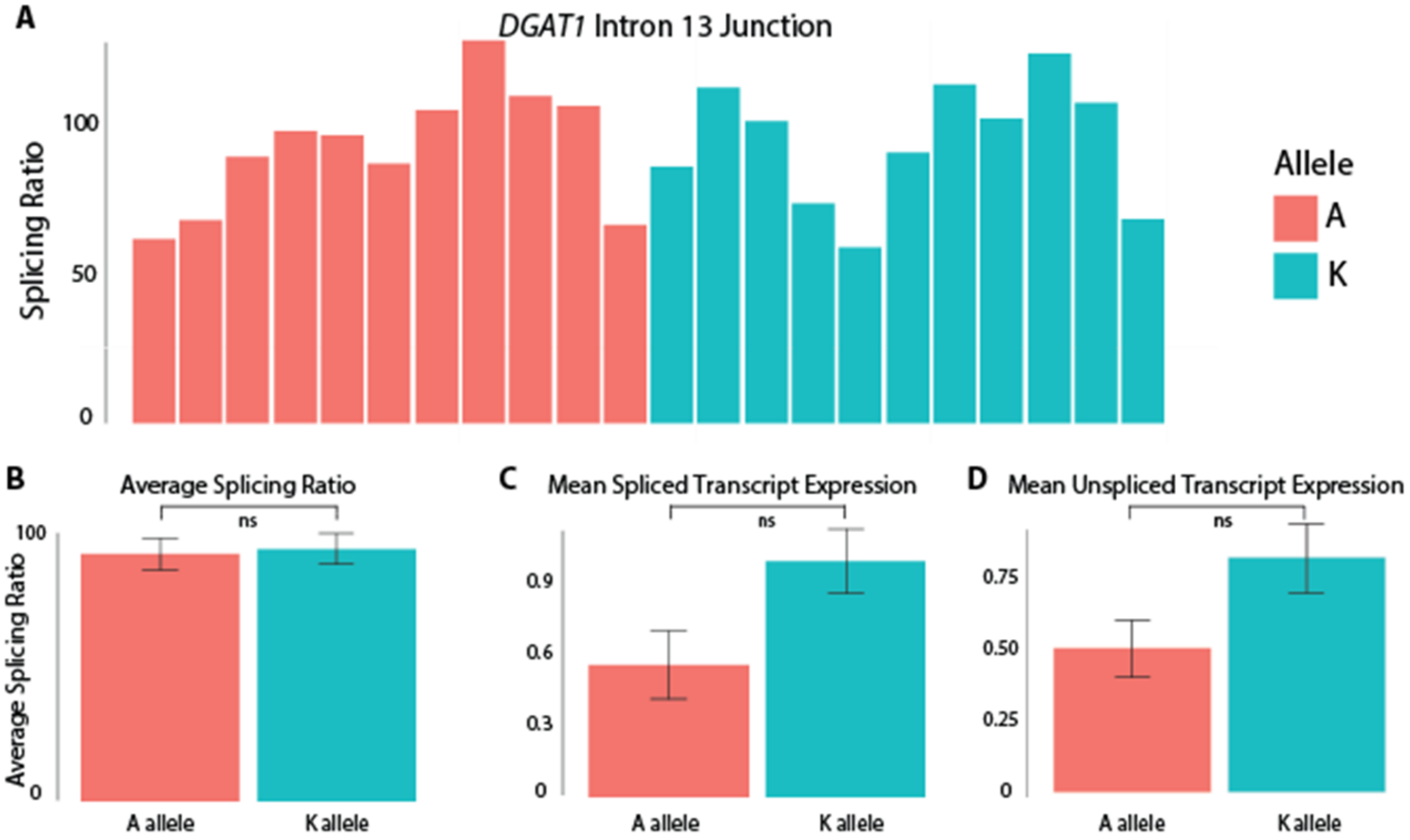


Figure 5.15 Cell-based functional testing of *DGAT1* K232A influence on splicing efficiency at the *DGAT1* intron 13 junction
Figure A represents the splicing ratio (spliced transcripts:unspliced transcripts) measured by qPCR in each of the individual replicates for the intron 13 junction. Figure B represents the average splicing ratio for the intron 13 junction for the two *DGAT1* K232A alleles. The error bars represent the standard deviation across all samples. Figure C and D represent the mean spliced and unspliced transcripts for the intron 13 junction, respectively. The error bars represent the standard error of the difference between means. *** = P ≥ 0.001, ** = P ≥ 0.01, * = P ≥ 0.05, ns = P > 0.05

5.5.7 *DGAT1* eQTL analysis using genetic markers derived from WGS

Given the demonstration that K232A was associated with splice enhancement and mean mRNA expression *in vitro*, we re-visited association mapping of the locus using imputed whole genome sequence data. In this analysis, we aimed to assess the relative contribution of the K232A variant to the *DGAT1* eQTL in the context of full sequence, examining whether other sequence variants were substantially more associated with expression, or whether the K232A wholly explained this effect. Association analysis was conducted using 3128 genetic markers imputed from WGS in the 1 Mbp centred on *DGAT1* K232A in conjunction with mammary *DGAT1* expression. Like the previous association analysis using the BovineHD markers, this analysis revealed a strong eQTL for *DGAT1* in the mammary gland, with K232A remaining one of the top associated variants ($P=1.59x10^{-25}$; Figure 5.15; Table 5.10).

The most highly associated markers for this signal were rs209328075 and rs209929366 which are located at chr14:1730455 and chr14:1747132, respectively ($P=2.31x10^{-28}$; Figure 5.15). These markers exhibited identical association statistics with mammary *DGAT1* expression. They were also highly correlated with K232A, exhibiting an R^2 value of 0.88 (Table 5.11). When K232A was fitted as a covariate in the association model, the association of the two lead variants was greatly reduced. These associations were non-significant when applying a Bonferroni correction, though significant in the absence of this correction (P=0.000263; Bonferroni threshold $P=1.60x10^{-05}$; Table 5.15). Of greater note, a cluster of 39 variants in perfect linkage disequilibrium (LD) with each other, but only modestly correlated with K232A ($R^2=0.548$) were significant ($P=3.58x10^{-05}$) in these models. These variants explained 10.32% of the residual phenotypic variance in mammary *DGAT1* expression, suggesting the possibility of another, functionally independent regulatory effect at the locus. Interestingly, some of the markers most highly associated with the residual *DGAT1* eQTL signal reside several kb upstream of the transcription start site of the gene (chr14:1428907-1754446; Table 5.13), representing candidates variants overlaying an additional upstream promoter or other regulatory feature impacting mammary *DGAT1* expression.

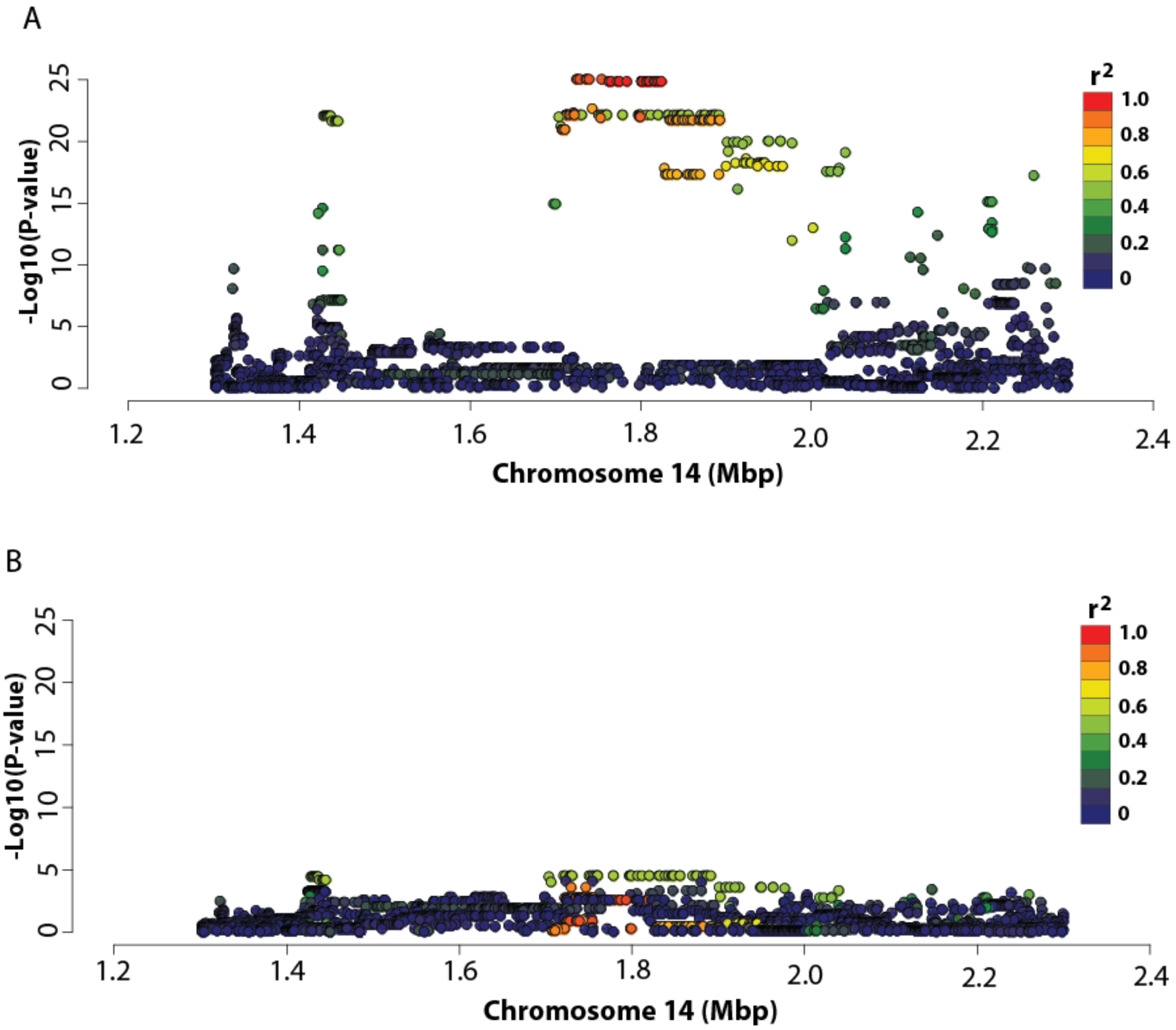


Figure 5.16 Expression QTL analysis at the *DGAT1* locus in bovine lactating mammary gland.
Figure A shows the Manhattan plot for *DGAT1* expression at the *DGAT1* locus in the RNAseq animals. The X-axis shows bp on chromosome 14, the Y-axis shows –log10 P-values of marker association for the 3218 WGS-derived SNPs in the 1 Mbp interval centred on *DGAT1* K232A. Figure B represents the Manhattan plot for *DGAT1* expression at the *DGAT1* locus in the RNAseq animals conditioned on K232A. Markers are coloured based on their correlations (R^2) with K232A in both Figure A and B.

Table 5.12. Mammary *DGAT1* expression association statistics for top WGS-derived variants
The positions of these SNP variants are indicated, with parameter estimates shown with standard errors in units of VST-transformed RNAseq read counts. The genetic and phenotypic variance explained by each SNP, along with parameter-adjusted means for each of the three genotypes classes is indicated. The linkage disequilibrium R^2 values for each SNP relative to *DGAT1* K232A variant is shown, with the P-values indicated in the right most column.

				Adjusted means					
Variant	**Chr14 pos**	**Parameter Est**	**Geno 0**	**Geno 1**	**Geno 2**	**Pheno var**	**Geno var**	**R^2 with K232A**	**P-value**
rs209328075	1730455	0.1927(±0.0161)	9.251	9.446	9.642	31.32	99.99	0.881	2.31x10^{-28}
rs209929366	1747132	0.1927(±0.0161)	9.251	9.446	9.642	31.32	99.99	0.881	2.31x10^{-28}
rs208091850[*]	1722033	0.1961(±0.0164)	9.244	9.440	9.637	31.23	99.99	0.922	1.29x10^{-27}
rs208417762[^]	1756075	0.1969(±0.0166)	9.240	9.437	9.634	31.33	99.99	0.968	2.38x10^{-27}
rs135458711[+]	1724688	0.1906(±0.0167)	9.250	9.440	9.631	29.46	99.99	0.952	1.10x10^{-25}
K232A[&]	1802265	0.1919(±0.0169)	9.244	9.436	9.628	29.71	99.99	1	1.59x10^{-25}

[*]31, [^]27, [+]7 and [&]20 additional genetic variants, respectively had the same association signal for *DGAT1* expression and are included in Appendix II. These variants were statistically indistinguishable from each other and are not included in this table in the interest of size.

Table 5.13 Mammary *DGAT1* expression association statistics for top sequence variants conditioned on *DGAT1* K232A
The positions of these SNP variants are indicated, with parameter estimates shown with standard errors in units of VST-transformed RNAseq read counts. The genetic and phenotypic variance explained by each SNP, along with parameter-adjusted means for each of the three genotypes classes is indicated. The linkage disequilibrium R^2 values for each SNP relative to *DGAT1* K232A variant is shown, with the P-values indicated in the right most column.

			Adjusted means						
Variant	**Chr14 pos**	**Parameter Est**	**Geno 0**	**Geno 1**	**Geno 2**	**Pheno var**	**Geno var**	**R^2 with K232A**	**P-value**
rs472613236*	1721117	0.1013(±0.0242)	9.361	9.463	9.564	10.32	99.99	0.548	3.58x10^{-05}
rs383105805^	1428907	0.1002(±0.0240)	9.363	9.463	9.564	10.13	99.99	0.539	3.88x10^{-05}
rs109448144	1704351	-0.1004(±0.0240)	9.564	9.464	9.364	10.06	99.99	0.544	4.14x10^{-05}
rs137587412&	1438890	0.0975(±0.0243)	9.366	9.463	9.561	9.62	99.99	0.546	7.08x10^{-05}
rs445906781	1723278	0.1165(±0.0295)	9.424	9.540	9.657	4.95	99.99	0.090	9.71 x10^{-05}
rs476272800	1754446	0.1165(±0.0295)	9.424	9.540	9.657	4.88	99.99	0.853	9.71 x10^{-05}

*39, ^9 and &5 additional genetic variants, respectively had the same association signal for *DGAT1* expression and are included in Appendix II. These variants were statistically indistinguishable from each other and are not included in this table in the interest of size.

5.6 Discussion

A pleiotropic QTL with a large influence on milk composition resides on the centromeric end of bovine chromosome 14, underpinned by the *DGAT1* gene. Although there has been some speculation as to the specific genetic variant and mechanism responsible for the QTL (Grisart et al., 2004; Kühn et al., 2004), *in vitro* functional evidence showing that two protein isoforms of *DGAT1* differed in their ability to synthesise triglycerides, led to the now dominant hypothesis that the K232A amino acid substitution is the causative variant underpinning this signal. However, we have generated a large mammary RNAseq dataset that provides the opportunity to revisit the idea of a transcriptionally regulated mechanism at this locus. To this end, this chapter describes the detailed investigation of the *DGAT1* locus using RNAseq data from lactating cows, and provides functional evidence for an expression-based mechanism by which *DGAT1* K232A may influence milk composition.

5.6.1 *DGAT1* K232A associates with *DGAT1* transcript abundance in the lactating mammary gland

Association analysis using RNAseq-derived expression data revealed a strong eQTL for *DGAT1* in lactating mammary tissue. This would appear to be the first report of a *cis*-eQTL for *DGAT1* in the bovine mammary gland. Importantly, the mammary *DGAT1 cis*-eQTL identified in this study bears a similar genetic signal underpinning the milk production QTLs reported for this locus, with the K232A highly associated with the gene expression effect. Animals bearing the K allele for *DGAT1* K232A possess greater mammary *DGAT1* expression compared to those animals bearing the A allele. This is of particular note given the K allele is the same allele associated with increased milk fat percentage (Grisart, Coppieters, Farnir, et al., 2002), and more *DGAT1* enzyme as a consequence of increased expression could be expected to increase triglyceride synthesis.

The finding that the K232A variant is one of the most highly associated genetic markers with *DGAT1* expression is surprising, since previous *in vitro* investigations using RT-PCR found no difference in *DGAT1* mRNA expression based on K232A genotype, albeit with limited numbers of animals (N=24; Grisart et al., 2004). Therefore, the effect of *DGAT1* K232A on milk fat production has been attributed to the enzymatic difference between the two DGAT1 isoforms, as the K allele had the greatest enzymatic activity *in vitro*, and associates with increased milk fat percentage. Based on this study, it was widely assumed that this enzymatic difference was the sole mechanism driving the effect of K232A on milk composition.

5.6.2 *DGAT1* K232A is associated with alternative splicing of *DGAT1*

Previously, *DGAT1* K232A has been shown to influence alternative splicing of *DGAT1* mRNA transcripts (Grisart et al., 2004). The alternative splice form of *DGAT1* differs based on its utilisation of a splice site 6 bp upstream of K232A, resulting in the 'intronification' of the majority of exon 8. The protein encoded from this isoform has an internal deletion of 22 amino acids, and is assumed to be non-functional based on its inability to synthesise triacylglyercides *in vitro* (Grisart et al., 2004). The proportion of this alternative isoform is approximately 10% of the total *DGAT1* transcripts and using RT-PCR, Grisart et al., (2004) illustrated that the K allele results in an increase in the amount of the alternatively spliced mRNA transcripts *in vitro*. In line with this study, the proportion of the alternative isoform to the full length-form differed by K232A genotype in the RNAseq dataset, with the animals bearing the K allele producing more of this isoform compared with those animals carrying the A allele.

Originally, it was hypothesised that the mammary *DGAT1* eQTL might be the result of increased alternative *DGAT1* isoform production. As the alternative isoform results in the intronification of the majority of exon 8, however, this is unlikely to be the case, since the K allele was associated with increased production of the alternative *DGAT1* isoform, which is also the allele associated with increased mean *DGAT1* expression.

5.6.3 *DGAT1* K232A disrupts a conserved exon splice enhancer and associates with efficiency of splicing *in vivo*

Gene expression can be influenced by polymorphisms within regulatory elements, which are most commonly attributed to non-coding sequences. However, regulatory elements can be part of the coding sequence, and coding variants, such as the dinucleotide substitution underlying *DGAT1* K232A, can influence gene expression through the modulation of auxiliary splicing elements. It was hypothesised that the *DGAT1* K232A mutation may overlap one of these elements to influence gene expression. The data presented in this chapter supports this hypothesis, demonstrating that the dinucleotide substitution responsible for K232A results in the disruption of a consensus ESE motif. Use of the RESCUE-ESE tool to annotate the exonic sequence around K232A suggested two predicted ESE motifs that overlapped the K232A polymorphism. These ESE motifs, AGAAGG and AAGAAG, were located at chr14:1802263-1802268 and chr14:1802262-1802267 (UMD3.1 genome build) respectively, and were only present with the K allele. Importantly, the K allele is the same allele associated with increased mammary *DGAT1* expression, which suggested a possible mechanism by which *DGAT1* K232A might exert its effect on *DGAT1* splicing, and hence mRNA expression.

The AAGAAG ESE motif has been proposed as the second most common ESE hexamer in vertebrates (Mersch, Gepperth, Suhai, & Hotz-Wagenblatt, 2008). Given the importance of ESEs for promoting splicing, I wondered if the motif could influence the splicing efficiency of *DGAT1* pre-mRNA to influence mRNA expression. It has been previously shown in humans that SNPs in ESEs can inhibit affinity for splicing factors and affect splicing, leading to altered mRNA and protein translation sequences that contribute to genetic disorders (Dvinge & Bradley, 2015). Additionally, the disruption of splicing has recently been reported for a novel *DGAT1* mutation in dairy cattle, whereby a non-synonymous A>C transversion in exon 16 disrupts a putative ESE motif and causes the skipping of this exon (Lehnert et al., 2015). This polymorphism results in an enzymatically inactive DGAT1, which in the homozygous state results in a severe phenotype characterised by scouring and slow growth (Lehnert et al., 2015).

Given the identification of a putative consensus ESE overlapping K232A and its association with enhanced production of the alternative isoform of *DGAT1*, the next step was to investigate the hypothesis that this polymorphism may influence *DGAT1* pre-mRNA processing more generally. To investigate this idea, a novel molecular phenotype was derived by quantifying the percentage of *DGAT1* RNAseq reads mapping to intron 8 and other junctions of the gene. The approach taken was to use this splicing efficiency phenotype in conjunction with the SNPs in the 1 Mbp interval centred on *DGAT1* K232A to conduct association analysis, similar to the method used for eQTL investigation. This analysis revealed strong splice enhancement for 5 *DGAT1* introns, providing evidence supporting the mechanism by which this variant might influence mammary *DGAT1* mRNA expression. Critically, the mammary *DGAT1* intron splicing efficiency effects appeared to bear the same genetic signature underpinning the eQTL and milk production QTLs reported this locus, that is, the association rankings for SNPs were similar for all QTLs. The direction of effects is also consistent with this hypothesis, where animals bearing the K allele have increased milk fat percentage, *DGAT1* expression and efficiency of splicing. Conversely, the RNAseq animals bearing the A allele had decreased *DGAT1* expression and splicing efficiency, where this allele is also associated with decreased milk fat percentage.

Splicing efficiency is dependent on a number of factors, with the likelihood of an intron being retained in mature mRNA a reflection of the strength of the splice site, intron length, GC content, splicing factor expression and changes in chromatin structure (Wong, Au, Ritchie, & Rasko, 2016). As such, polymorphisms in ESEs and other splicing elements can influence transcription levels by modifying the strength of the recruitment of the splicing machinery to the junctions in the pre-mRNA transcript (Ge & Porse, 2014). The splicing efficiency effect and increased *alternative* splicing for *DGAT1* suggests the possibility that there are a number of weak splice sites in the gene, and the presence of the ESE in the K allele enhances the recruitment of the splicing machinery to increase their usage, resulting in increased splicing of these junctions.

Interestingly, the junctions that had a splicing efficiency phenotype associated with K232A were distributed throughout the gene and included intron 1-3, and 11, which are several kb from the polymorphism and ESE motif. Similar to a recent study by Ni et al., (2016), the approach taken in this study accounted for any bias size and read coverage differences across the gene and subsequently revealed no relationship between the size of the intron and the splicing efficiency at the junction. The *DGAT1* intron 1 and 2 junctions, which contain the two largest introns, both exhibit a strong splicing efficiency effect, with the intron 2 junction exhibiting the most significant effect in this analysis. Reasons why particular *DGAT1* junctions appear to be influenced by K232A genotype while others remain unaffected are unclear at this stage. It is possible that during pre-mRNA processing, the *DGAT1* junctions are processed in an order such that some junctions become rate-limiting steps in the process. If such a bottleneck exists, then the presence of the ESE could influence the efficiency of the processing of this junction and the junctions that are subsequently processed. This would result in certain junctions exhibiting a splicing efficiency difference based on the presence or absence of the ESE, while the junctions prior to the bottleneck would remain unaffected. Ultimately, further research is required to understand the relationship between the activation of the exon 8 ESE in *DGAT1* and its influence on the splicing efficiency at multiple junctions in the gene.

5.6.4 *DGAT1* K232A influences *in vitro* splicing efficiency

Despite the strong association between *DGAT1* K232A and the expression and splicing efficiency phenotypes, there was still some possibility that one or more of these associations were due to LD effects exerted by an unknown genetic variant. To more directly probe the function of K232A, mini-gene constructs were generated for the K and A alleles in the absence of native promoter sequence. Differing only by the dinucleotide substitution responsible for K232A, expression testing of these constructs replicated the splicing efficiency effect for a subset of the same junctions implicated *in vivo*, unequivocally assigning an expression-based mechanism to this variant.

Interestingly, the splicing efficiency effects appeared to result in an increase in spliced mRNA expression, rather than increased expression *per se* as there was no concomitant increase in unspliced transcript expression at the two junctions exhibiting the splicing efficiency phenotype. The lack of increased expression of the unspliced pre-mRNA transcripts may be the result of an increased rate of pre-mRNA processing and supports the hypothesis that splicing is directly impacting mammary mRNA expression. Boutz et al., (2015) used RNAseq and RT-PCR to demonstrate that many transcripts retain specific introns. This study also reported that transcripts containing introns were retained in the nucleus and did not undergo degradation via NMD, and *in vitro* these introns appear to be eventually spliced out at a much slower rate than other introns in the same transcript, exhibiting a mean half-life of 29 minutes, compared with the 11 minute half-life of normally spliced introns (Boutz, Bhutkar, & Sharp, 2015). These observations suggest that incompletely spliced *DGAT1* transcripts may also eventually be spliced, albeit at a much slower rate.

5.6.5 Expression analysis conditioned on K232A revealed additional effects at *DGAT1* locus

While it is not the first time an expression-based effect of *DGAT1* has been proposed as the mechanism by which this gene influences milk composition (Lehnert et al., 2015), our study is the first to provide evidence supporting an expression-based effect associated with K232A. Association mapping at the *DGAT1* locus using imputed WGS variants showed that *DGAT1* K232A remains as one of the most highly associated variants, even in the context of full sequence resolution data. However, K232A was not the marker with the smallest P-value, so the possibility remains that additional effects reside at the locus, or that imperfect sequence imputation or sampling error may have impacted the relative association rankings of the variants in this interval.

To attempt to address these possibilities, further association analysis was conducted to include K232A genotype as an additional covariate in the models. This analysis removed the majority of the association signal for *DGAT1* expression, suggesting that the *cis*-eQTL is derived, for the most part, from *DGAT1* K232A. The clusters of highly significant markers in the previous analysis were no longer associated with *DGAT1* expression in these models, suggesting these variants were tagging the signal from K232A. However, a seemingly distinct, statistically significant eQTL remained, signifying there may be additional effects on mammary *DGAT1* expression. A number of these highly associated markers are located upstream of the transcription start site of the gene, suggesting there may be an additional promoter driven effect on mammary *DGAT1* expression. One possibility is the VNTR polymorphism proposed by Kuhn et al., (2004) which was hypothesised to increase the number of putative SP1 transcription factor binding sites, and stimulate an increase in *DGAT1* expression.

5.6.6 Limitations of this study

Although the current study definitively assigns an expression based effect to the *DGAT1* K232A polymorphism, the described work has several limitations. Cell culture experiments were conducted using the MAC-T cell line, which, while a reasonable model of the lactating mammary gland, has a heterogeneous morphology. The effect of this is unclear but may have added noise to gene expression measurements. Additionally, the plasmids used in this study contained mini-gene *DGAT1* constructs under the control of a CMV promoter. Removing the constitutive promoter was critical to establish any expression-based effect outside of a bovine genomic context; however having the expression of these genes driven by a viral promoter provides an artificial context to measure *DGAT1* expression. It is therefore possible that the splicing efficiency effect may 'bottle-neck' mRNA expression when extreme levels of pre-mRNA are produced, but the impact is less pronounced when the transcript is expressed at a more physiological level. An alternative approach that would circumvent some of the shortcomings of the work would be to use genome editing, to introduce different alleles of the variant in their native genomic background.

While we were able to identify the *cis*-eQTL and splicing efficiency effects at the *DGAT1* locus, an unresolved question is what proportion of the K232A impacts on milk composition are derived from differences in enzymatic activity and expression based effects. One possible option to delineate these two mechanisms would be to use redundant codons to create cell lines that encode identical DGAT1 proteins, yet have alternate ESE-encoding genomic sequences. Unfortunately however, lysine and alanine amino acids have limited redundancy, precluding design of such constructs.

5.7 Summary and Conclusion

DGAT1 represents the most well-known and validated gene influencing bovine milk composition and production. Given its importance to lactation traits, the genetic regulation of mammary *DGAT1* expression was investigated in this chapter. The approach taken was to use high-depth mammary RNAseq data in conjunction with the BovineHD panel markers to conduct *cis*-eQTL mapping in NZ dairy cattle. This analysis revealed a strong eQTL that appeared to be the result of an ESE which overlaps the K232A polymorphism previously demonstrated to alter triglyceride synthesis *in vitro*. Further analysis revealed that the disruption of this splice enhancer influences the splicing efficiency at numerous junctions of the genes. This effect was confirmed in cell-based *in vitro* experiments, whereby the amounts of spliced and unspliced RNA transcripts were quantified for four junctions of the gene. In line with the RNAseq data, the ratio of these transcripts differed between the K and A allele at the intron 3 and 7 junctions and was unaffected at intron 5 and 11 junctions. Taken together, the data presented in this chapter suggests that the effect of *DGAT1* K232A on milk production, at least in part, be due to this expression-based mechanism.

Chapter 6: Detailed investigation of the *AGPAT6* milk fat percentage QTL

6.1. Overview

In *Bos taurus*, genetic studies have highlighted a quantitative trait locus (QTL) on chromosome 27 with a substantial impact on milk fat percentage (Khatkar, Thomson, Tammen, & Raadsma, 2004b; X. Wang et al., 2012). The gene implicated in this effect is glycerol-3-phosphate acyltransferase 4 (A*GPAT6*, also known as *GPAT4*), which represents a strong positional candidate, as this gene encodes the enzyme that catalyses the second acylation step in the triacylglycerol synthesis pathway (Bionaz & Loor, 2008). Further support for *AGPAT6* includes the observation that it is the most abundantly expressed *AGPAT* in the mammary gland, and its transcription rate is highly correlated with the concentration of diacylglycerols and triacylglycerols in milk (Takeuchi & Reue, 2009). Additionally, mice with *AGPAT6* deficiency have impaired lactation, reduced size and number of alveoli, fewer fat droplets, and reduced diacylglycerols and triacylglycerols in their milk (Beigneux et al., 2006; Vergnes et al., 2006).

Despite the strong candidacy of *AGPAT6* underpinning the chromosome 27 milk fat percentage QTL, the causal variant through which this gene mediates its effect is yet to be established. This reflects the difficulty in deciphering the causal variant amongst the large number of associated sequence polymorphisms, particularly when the most highly associated variants reside in non-coding gene regions of *AGPAT6*. Wang et al., (2012) re-sequenced this locus using Sanger sequencing to propose rs208314235 (chr27:36211252GA>T; UMD3.1 genome build) as the causative variant for the QTL, with the variant proposed to locate to the 5′ flanking region of *AGPAT6*. The polymorphism was hypothesised to influence the binding of SREB, CREB, RXR-*α* and RAR-β transcription factors as part of a bioinformatically predicted regulatory sequence upstream of the transcription start site (X. Wang et al., 2012). However, the gene model used by Wang et al., (2012) did not include the first exon notable from human and other mammalian gene models (Beigneux et al., 2006; Yan et al., 2008), suggesting the variant locates to *AGPAT6* intron 1, versus the putative promoter of the gene. This opened the possibility that alternative, and potentially more biologically plausible candidate causative variants exist in unexplored regions of the gene. Subsequently, through visualisation of mammary RNA sequence (RNAseq) alignments, we confirmed the existence of a previously unannotated *AGPAT6* exon 1, consisting of 5'UTR sequence seemingly present in all mammary *AGPAT6* transcripts (Littlejohn et al. 2014). Further examination of whole genome sequence (WGS) and RNAseq data suggested the presence of an indel mutation (chr27:36198118GGC(4_5); hereafter referred to as VNTR) in this exon, where this variant appeared to be in linkage disequilibrium (LD) with milk fat percentage-associated variants in the small number of animals for which WGS and RNAseq data was available.

This chapter describes the detailed investigation of the *AGPAT6* milk fat percentage QTL, with a particular focus on developing a genotyping assay to directly interrogate the candidate causal VNTR variant. Results of this analysis also form part of a journal article examining effects at the *AGPAT6* locus in broader detail (Littlejohn et al. 2014). Additionally, I report work conducted to attempt to validate the association results of Littlejohn et al., (2014) in a separate population of 37,236 dairy cows, and conduct CRISPR-Cas9 mediated experiments to attempt to provide functional evidence for the causality of the VNTR.

6.2. General aim

Investigate the causative status of the *AGPAT6* VNTR candidate variant as responsible for the chromosome 27 milk fat percentage QTL.

6.3. Specific aims

1. Develop a robust PCR-based assay to genotype the *AGPAT6* VNTR candidate causal variant in the F1 sires and F2 dams from the FJXB population.
2. Conduct association analysis of milk composition and production phenotypes using imputed WGS derived variants for the *AGPAT6* locus.
3. Conduct CRISPR-Cas9 mediated genome editing of the *AGPAT6* locus to generate clonal bovine mammary cell lines containing three candidate causative variants.
4. Conduct *in vitro* functional testing of three *AGPAT6* candidate causative variants by measuring *AGPAT6* expression in genome edited bovine mammary cell lines.

6.4. Methods

6.4.1. Animal cohorts, genotypes and milk composition phenotypes

The work described in this chapter was conducted using two independent animal populations; the FJXB pedigree, and mixed ancestry dairy cow population. Both of these datasets are described in General Methods (Chapter 2). The specific numbers of animals studied, and details of phenotypes and genotypes pertaining to this work are stated below.

6.4.2. Genotyping assays

6.4.2.1. PCR and Sanger sequencing

Genomic DNA (DNA had already been extracted; details beyond the remit of this thesis) from six F1 sires and 889 F2 dams from the FJXB cohort was used as input for the PCR-based genotyping assays.

Genomic DNA flanking the *AGPAT6* VNTR was amplified by PCR in the six F1 FJXB sires using the KAPA Robust PCR system (Kapa Biosystems) and primers in Table 6.1. Amplification was performed in 25 µL reactions containing KAPA 2G Robust enzyme in conjunction with the KAPA GC-rich buffer, and 50 ng genomic DNA (see General Methods for PCR reaction mix). Amplification conditions included an initial denaturation at 95°C for 3 minutes followed by 35 cycles of 95°C for 30 seconds, 60°C for 30 seconds, and 72°C for 30 seconds.

PCR products were visualised following separation by gel electrophoresis (2% w/v agarose) for 30 minutes. Following this, the PCR products were cleaned using ExoSAP-IT (Affymetrix) following the manufacturer's instructions. Briefly, 10 µL of each PCR product was combined with 4 µL ExoSAP-IT and incubated for 15 minutes at 37°C. Then, the samples were incubated for 15 minutes at 80°C and quantified by Nanodrop. The PCR products were then sequenced using the primers in Table 6.1 by the Genomics Centre, Auckland Science Analytical Services, The University of Auckland (Auckland, NZ).

Table 6.1 Primer sequences and PCR of *AGPAT6* VNTR in F1 FJXB sires

Primer	Sequence	Product size
Short_For1	GACGAGAGGGTCACGTCAAG	192 bp
Short_Rev1	AGCCCCGCTAGAGGTTCAT	

6.4.2.2. GeneScan genotyping assay

Genomic DNA flanking the *AGPAT6* VNTR was amplified by PCR in 889 FJXB F2 dams using the primers in Table 6.2. The forward primer was fluorescently labelled with FAM to enable GeneScan analysis. Amplification was conducted in 10 µL reactions using 20 ng genomic DNA, with reaction and PCR cycling conditions the same as described above (see General Methods for PCR reaction mix).

After completion of PCR, 90 µL of Ultra-Pure water (Invitrogen) was added to each reaction and gently mixed. A 2 µL aliquot was transferred to a 96-well plate for GeneScan analysis at the Genomics Centre, Auckland Science Analytical Services, The University of Auckland (Auckland, NZ), who also provided the ROX 400HD Ladder.

GeneScan fragment traces were visualised and analysed using Geneious (version 6.0.3). Raw traces were trimmed and the ladder was inspected to ensure all points in the ROX 400HD ladder were called. FAM locus information was set for each independent GeneScan sequencing run, whereby two peaks with a repeat unit of 3 bp were expected at position 181 bp and 184 bp within the PCR fragment, representing the two VNTR alleles. When a peak was observed in the first bin it was judged to represent the VNTR T allele, while a peak observed in the second bin was judged to represent a VNTR TGGC allele. Genotypes were manually called from the presence and/or absence of peaks within these bins.

Table 6.2 Primer sequences used for the GeneScan *AGPAT6* VNTR genotyping assay
The forward primer AGPAT6_FAM was fluorescently labelled with FAM.

Primer	Sequence
AGPAT6_rev1	ABDTAILAGCCCCGCTAGAGGTTCAT
AGPAT6 _FAM	6FAMCAAGGCGGCGTAGACAAA

6.4.3. Association analysis of milk composition

Association analysis was conducted in the mixed ancestry dairy cow population, consisting of 12,605 Holstein-Friesians, 5,652 Jerseys, and 18,979 NZ Holstein-Friesians x Jersey cross breeds. Phenotypes in the form of milk fat percentage and yield, protein percentage and yield, lactose percentage and yield, and milk yield were obtained from the herd testing records for these animals (described in General Methods). Differences in animal numbers quoted for each analysis is a reflection of the quality filtering performed on each trait. Briefly, animals were only included for analysis if they had at least two herd-test records that met the following criteria: more than 5 litres of milk volume, less than 250,000 somatic cells, and a minimum lactation duration of 30 days. Additionally, records were removed if they were more than 5 standard deviations from the mean.

The 3519 SNPs in the 1 Mb interval centred on the *AGPAT6* VNTR (chr27:25905136-36404713) in the imputed WGS dataset were used for association analysis in the mixed ancestry population. These markers were imputed using a reference population of 556 animals as described in General Methods. Genotypes were extracted using samtools (version 0.1.19) and recoded using PLINK2 (version 1.90b2c) to 0, 1 or 2 to represent the number of alternative alleles for each marker.

Associations between the SNPs in the 1 Mbp interval surrounding the *AGPAT6* VNTR and the seven lactation phenotypes were quantified using pedigree-based models in ASReml-R (A R Gilmour et al., 2009; Arthur R. Gilmour et al., 1995). For each phenotype, each SNP was fitted in a separate sire-maternal grandsire single trait model, with SNP treated as a quantitative variable based on the number of copies of the alternative allele and variance components estimated in a restricted maximum-likelihood (REML) framework. Covariates for cohort, the proportions of NZ Holstein-Friesian ancestry, US Holstein-Friesian ancestry, Jersey ancestry and heterosis effects were also included in the models. The additive genetic variance, polygenic genetic variances, total genetic variance and phenotypic variance for each milk composition phenotype was calculated as described in General Methods. The proportion of phenotypic and genotypic variance explained by each SNP was also calculated as described in General Methods.

6.4.4. CRISPR-Cas9 mediated genome editing of *AGPAT6* locus

To test the effect of the VNTR variant on *AGPAT6* expression *in vitro*, CRISPR-Cas9 mediated editing of the *AGPAT6* locus was conducted in the bovine mammary cell line (MAC-T; Huynh, Robitaille, & Turner, 1991) based on the optimised protocol established in Chapter 3. In conjunction, the chr27:36211257GA>T and chr27:36212352G>A variants were also targeted as these represent other candidate causative variants identified previously (Daetwyler et al., 2014; X. Wang et al., 2012). To obtain specific, targeted edits, single-stranded oligonucleotide (ssODN) HDR templates were co-transfected with CRISPR RNP, containing alternate alleles for each of these three variants. As the cell line is heterozygous for these variants (refer to Chapter 3), both alleles were targeted for HDR by transfecting two forms of HDR template for each variant (refer to Appendix I).

Cells were plated in 12-well plates and grown for 24 hours in complete proliferation media to achieve ~70% confluency (see General Methods). For cell transfection, RNP complexes were formed using the gRNAs for each target and Cas9 protein (as described in Chapter 3; sequences in Appendix I), and incubated with 5 µL Lipofectamine® RNAiMAX (ThermoFisher) in a final volume of 250 µL Opti-MEM (ThermoFisher) for 20 minutes. After 15 minutes of the incubation, 1.5 nM ssODN and 1500 ng pMAXGFP were added to the transfection mix. During this incubation, each well was washed with 1 mL pre-warmed PBS and replaced with 1 mL antibiotic-free proliferation media. Following the 20 minute incubation, 250 µL transfection complexes were added to each well and incubated at 37°C for 48 hours.

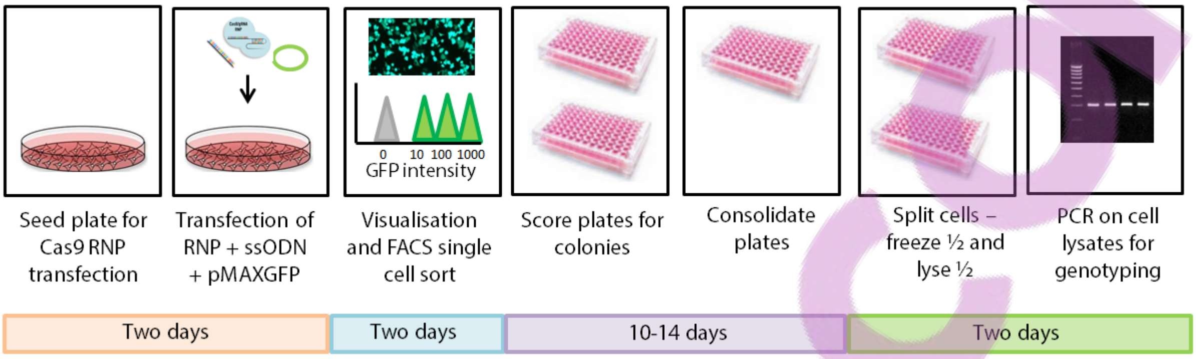

0
10 100 1000
GFP intensity
Seed plate for Cas9 RNP transfection
Transfection of RNP + ssODN + pMAXGFP
Visualisation and FACS single cell sort
Score plates for colonies
Consolidate plates
Split cells – freeze ½ and lyse ½
PCR on cell lysates for genotyping
Two days
Two days
10-14 days
Two days

Figure 6.1 Schematic of the generation of CRISPR-Cas9 edited clonal cell lines
First, cells are transfected with CRISPR-Cas9 RNP, ssODN and pMAXGFP plasmid. Two days later, GFP-positive cells are sorted and plated onto 96-well plates. When the cells reach seeding density, the plates are scored for the presence of colonies and transferred to new plates. When the cells reach ~90% confluence, they are split such that half the cells are frozen and half are lysed directly and the crude lysate is used to amplify the genomic regions encompassing the target sites.

6.4.4.1. FACS sorting and clonal expansion of cells

After 48 hours, transfection efficiency was assessed by visualising GFP fluorescence using a Nikon Ti-E inverted light microscope. Following visualisation, the media was removed, replaced with 250 µL of trypsin and incubated for 5 minutes. After the incubation, the trypsin was deactivated by adding 750 µL of full proliferation media and gently mixed. The cell suspensions were transferred to a 15 mL falcon tube and centrifuged for 5 minutes at 1300 g. Then, the media was carefully removed and the cell pellet was washed in 100 µL pre-warmed PBS and resuspended in 500 µL FACS Pre-sort buffer (BD Biosciences) supplemented with 50 ng DAPI immediately prior to cell sorting.

The fluorescence of GFP was detected by a FACSAria™ II flow cytometer (BD Biosciences). FACS was used to plate low and high GFP expression cell populations into 96-well plates such that individual cells were sorted into each well and further grown for clonal isolation. Two plates from each of the high and low GFP expression cell populations were plated, along with one plate of mixed GFP expression cells, for each RNP complex. Cells were sorted into wells containing 100 µL pre-warmed conditioned proliferation media.

The clonal cell expansions were visually monitored for almost two weeks, and wells with single colonies were marked and grown until they were at the approximate seeding density for a 96-well plate. At this stage, colonies were stripped from wells using 40 µL trypsin (using the protocol described above), and consolidated in new 96-well plates.

Once cells were ~90% confluent, an aliquot was taken for genotyping while the remaining cells were frozen. Then, 40 µL trypsin was used to strip the cells from the surface of the well (using the protocol described above). Then, 20 µL of the cell suspension was transferred to a 96-well PCR plate for genotyping, with 100 µL freezing media added to the remaining cells and frozen at -1°C/second and stored at -80°C.

The 96-well plate containing 20 µL cell suspensions was centrifuged at 1200 rpm for 5 minutes, before trypsin was removed and replaced with 20 µL of cell lysis buffer (10 mM Tris pH 8.0, 10% Triton-X 100). After gently mixing, the cell lysates were heated at 55°C for 10 minutes to ensure complete disruption of the cells before input into PCR for genotyping.

6.4.4.2. Genotyping of clonally derived CRISPR-Cas9 edited cell lines

The genomic DNA flanking the target site was amplified by PCR using the KAPA Robust PCR system (Kapa Biosystems) directly on the cell lysates from each clonal cell line. A 5 µL aliquot of the cell lysate was transferred directly to 20 µL PCR reaction mixture containing 0.5 units KAPA 2G Robust enzyme. Each target had slightly different reaction mixtures and PCR annealing temperatures (as per Table 6.3), but had standard PCR cycling conditions (95°C for 3 minutes, 35 cycles of 95°C for 15 seconds, 56°C or 60°C for 15 seconds and 72°C for 15 seconds, with a final extension at 72°C for 5 minutes).

Table 6.3 PCR cycling conditions for amplification of genomic DNA encompassing three *AGPAT6* variants

Target variant	Kapa buffer	Annealing Temp (°C)
VNTR	GC-rich	60
Chr27:36211257	Buffer A	56
Chr27:36212352	Buffer A and Enhancer 1	60

PCR products were purified using AMPure Bead purification (Agencourt), by adding a 1.6 volume of AMPure beads to the sample and following the manufacturer's protocol. The sequencing strategy for this work was 2 x 150 bp paired-end sequencing with an Illumina Miseq. PCR products were barcoded and purified DNA samples were quantified using a Qubit 2.0 Fluorometer and were pooled in an equimolar ratio. Sequencing libraries were then sequenced with the Illumina MiSeq Sequencer using a Miseq 300 cycle Nano kit (Life Technologies), by New Zealand Genomics Limited (NZGL; Auckland, NZ).

Alignments were performed for each sample using the mem command of the Burrows-Wheeler Aligner (BWA mem; default parameters; version 0.7.12), referencing the UMD3.1 assembly. The mapped sequence data were sorted using samtools version (1.3.1), with genotypes subsequently called for all variants using haplotype caller (The Genome Analysis Toolkit (GATK); v2015.1.1-3.4.46-0-ga8e1d99).

6.5. Results

6.5.1. PCR-based genotyping of the *AGPAT6* VNTR variant in the FJXB cohort

The 5′ region of *AGPAT6* is highly GC-rich, and an initial attempt to genotype the VNTR variant using the Sequenom platform was unsuccessful (prior to the onset of this thesis; data not shown). To develop a more robust approach, a number of PCR-based assays were designed to genotype this variant in the FJXB population. The primer pair in Table 6.1 produced a single, clean 190 bp band encompassing the VNTR. The PCR products from the six FJXB F1 sires were subsequently, amplified, purified, and sequenced. Sanger sequencing revealed that three sires were heterozygous for the VNTR (T/TGGC; sire 004, 114 and 405), two sires were homozygous reference (T/T; sire 685 and 837), while sire 740 was homozygous alternate for the VNTR (TGC/TGGC; Table 6.4).

While Sanger sequencing can be considered the gold standard for mutation detection, its cost is prohibitive to genotyping large numbers of individuals. To address this, PCR-based GeneScan was used to genotype the F2 dams of the FJXB population for the VNTR variant. Of the 889 F2 cows for which DNA samples were available, 826 (92.9%) were successfully genotyped using this platform, with only 63 animals (7.1%) failing either PCR and GeneScan. Of the 826 genotyped animals, 234 were homozygous reference (T/T), 414 animals were heterozygous (T/TGGC), while 178 animals were homozygous alternate (TGGC/TGGC; Table 6.5). The allele frequency of the *AGPAT6* VNTR T allele was 0.53 in the genotyped F2 animals. The genotyping results for the 889 F2 daughters of the FJXB population for the *AGPAT6* VNTR are presented in Appendix III.

Following genotyping of the VNTR variant in the F2 cows, this variant was included in the association analysis at the *AGPAT6* locus as described in Littlejohn et al., (2014). The association analysis between milk fat percentage and the 332 SNPs in the 1 Mbp interval centred on *AGPAT6* revealed a significant impact on milk fat percentage and, notably, the VNTR was the most highly associated variant in this analysis ($P=4.81x10^{-10}$; Littlejohn et al., 2014). The VNTR explained 5.9% and 7.8% of the phenotypic and genotypic variance, respectively, in milk fat percentage in the FJXB F2 animals.

Table 6.4 ***AGPAT6*** **VNTR genotypes for the six F1 sires of the FJXB population**

Sire	Genotype	
004	Heterozygous	T/TGGC
114	Heterozygous	T/TGGC
405	Heterozygous	T/TGGC
685	Homozygous Reference	T/T
740	Homozygous Alternate	TGGC/TGGC
837	Homozygous Reference	TGGC/TGGC

Table 6.5 GeneScan genotypes for the ***AGPAT6*** **VNTR in the FJXB F2 dams**

Genotype Class		Observed Number
Homozygous Reference	T/T	234
Heterozygous	T/TGGC	414
Homozygous Alternate	TGGC/TGGC	178
Unable to genotype	N/A	63

6.5.2. *AGPAT6* VNTR associates with milk composition in NZ dairy cows

Subsequent to the Littlejohn et al., (2014) paper, I conducted further work at this locus to validate these results and provide further evidence for the causal status of the *AGPAT6* VNTR. I conducted association mapping in an independent NZ population using imputed WGS data (with VNTR genotypes called from sequence alignments). In this analysis, I aimed to assess the relative contribution of the variants to the milk composition QTLs at sequence resolution. Association analysis was conducted using 3519 genetic markers imputed from WGS in the 1 Mbp interval centred on the *AGPAT6* VNTR in conjunction with milk fat percentage. This analysis replicated the strong QTL for milk fat percentage ($P=1.89x10^{-96}$; Table 6.7, Figure 6.2), with the VNTR one of the top-ranking variants ($P=3.98x10^{-90}$; Table 6.6, Figure 6.2). The VNTR variant accounted for 2.22% of the genetic variance and 1.26% of the phenotypic variance in milk fat percentage for this population (Table 6.7).

While the VNTR variant was significantly associated with milk fat percentage ($P=3.98x10^{-90}$), the most highly associated marker for this signal was rs208396531, which is located at chr27:36223974 ($P=1.89x10^{-96}$; Table 6.8, Figure 6.2). This marker is located 3′ of the VNTR variant and is in moderately strong LD with the VNTR, exhibiting an R^2 value of 0.81 (Table 6.8). Despite the VNTR variant not being the most significant variant in this analysis, it still represents a strong candidate. Indeed, using the relative ranking of the VNTR as an indication of its likely causality requires acknowledgement of the potential inaccuracy of genotype imputation (and therefore association statistics). This is particularly relevant in the current context where, due to the *AGPAT6* 5′UTR exon being encoded by a highly GC-rich sequence, read depth representation for the VNTR was low or absent for a substantial proportion of the WGS reference samples (38 of 556 animals had zero read depth).

To further investigate if other lactation phenotypes may be impacted by the *AGPAT6 VNTR*, association analysis was conducted for milk protein, fat and lactose yield, milk volume, milk protein and lactose percentage in conjunction with the same set of genetic markers imputed from WGS. Significant associations were demonstrated for all lactation traits except protein yield (Table 6.7, Figure 6.2). Specifically, the VNTR was significantly associated with milk yield ($P=1.79x10^{-06}$), milk fat yield ($P=3.16x10^{-14}$), protein percentage

($P=4.28x10^{-07}$), and lactose percentage and yield ($P=3.40x10^{-51}$ and $P=1.94x10^{-11}$, respectively; Table 6.7 and Figure 6.2). When the VNTR variant as fitted as a covariate in the association models, the association was greatly reduced for milk fat yield and lactose percentage (P=0.0181 and 0.3994, respectively; Table 6.8, Figure 6.3). However, significant associations remained for milk yield, protein percentage and lactose yield ($P=5.32x10^{-08}$, $5.26x10^{-15}$ and $6.67x10^{-07}$, respectively; Table 6.8 and Figure 6.3). The top markers for these residual signals were modestly correlated with the VNTR variant, with the lead SNPs exhibiting an R^2 value of 0.41, 0.61 and 0.41 for milk yield, protein percentage and lactose yield, respectively (Table 6.8). Taken together, these results suggest the VNTR is not fully capturing the genetic variance for some traits at this locus. This might be the result of inaccurate imputation of the variant as discussed above, or alternatively, multiple genetic effects may be at play at this chromosomal region.

The frequency of the milk fat percentage increasing VNTR 'T' allele was 0.59 in this population, and was associated with increased milk fat and protein percentage and milk fat yield, and decreased lactose percentage and yield and total milk volume (Table 6.7).

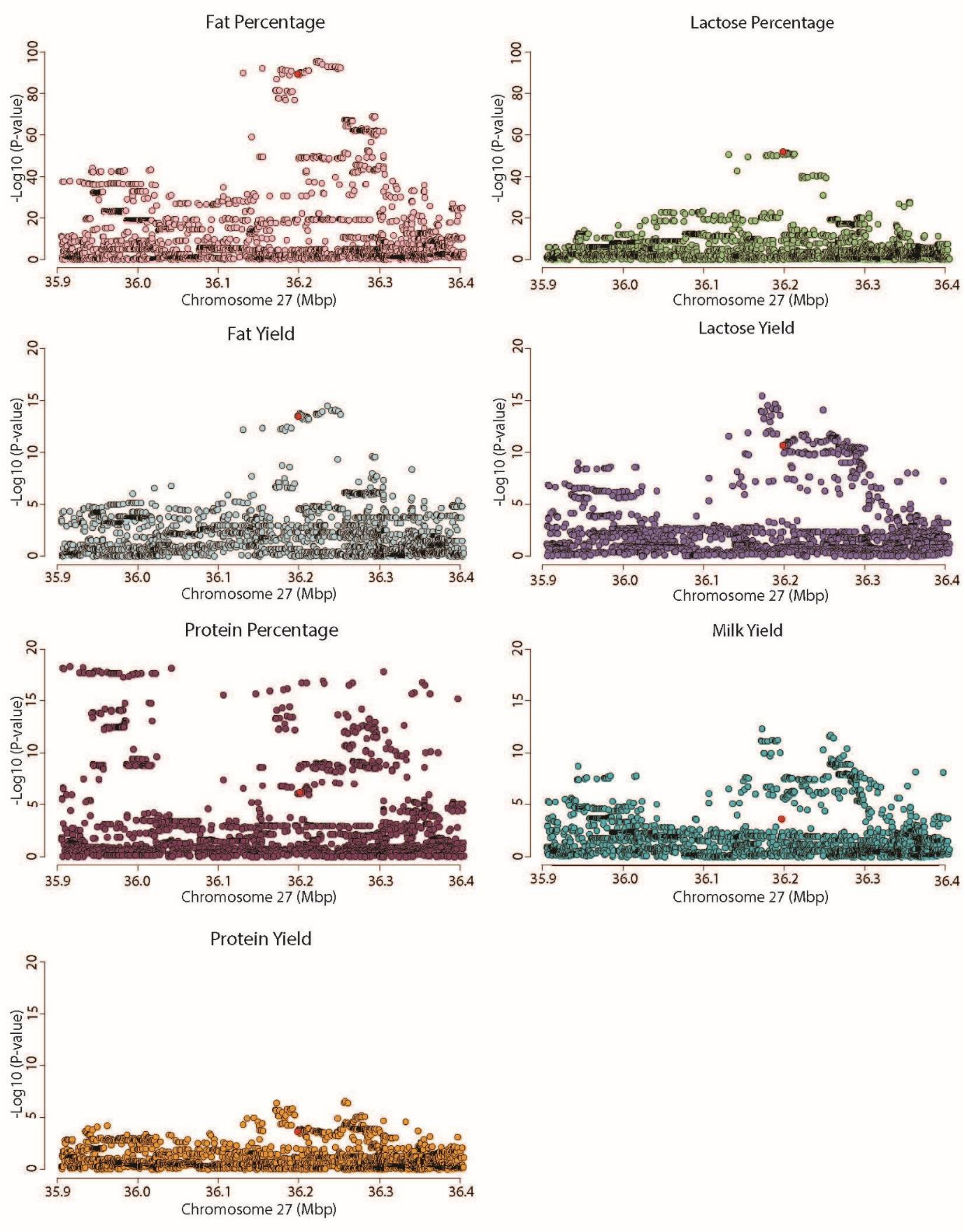


Figure 6.2 Milk composition QTLs at the chromosome 27 locus
Manhattan plots for the seven milk composition and yield phenotypes, with the X-axis showing Mbp position on chromosome 27, and the Y-axis showing –log10 P-values of marker association. The *AGPAT6* VNTR polymorphism is coloured red.

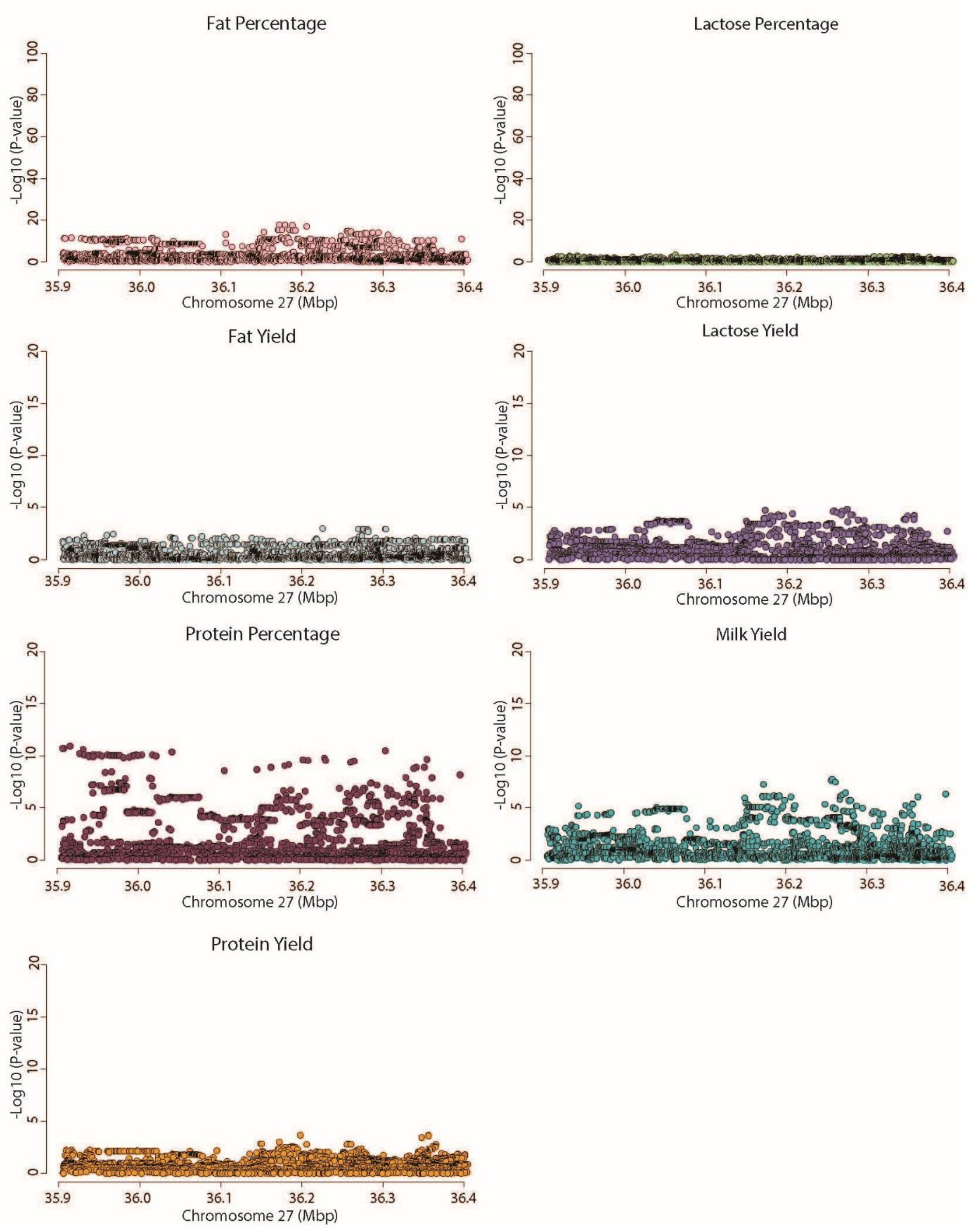


Figure 6.3 Milk composition QTLs at the chromosome 27 locus conditioned on *AGPAT6* VNTR genotype Manhattan plots for the seven milk composition and yield phenotypes where *AGPAT6* VNTR genotype has been fitted as an additional covariate in otherwise identical association models, with the X-axis showing Mbp position on chromosome 27, and the Y-axis showing –log10 P-values of marker association.

Table 6.6 Association between *AGPAT6* VNTR and milk composition traits in NZ dairy cows

Parameter adjusted estimates are given with standard errors in units of grams for yield traits and litres for milk volume, with adjusted means in the same units. The percentage of the total phenotypic and genotypic variance of milk composition explained by the rs381105171 SNP for each trait is in the 'Pheno var' and 'Geno var' columns, respectively. P-values of genetic association are indicated in the right-most column. Bonferroni threshold P=1.42x10^{-05}

Phenotype	Animal N	Effect ±SE	Pheno var	Geno var	P-value
Milk yield	37218	0.0983 ± 0.0206	0.075	0.265	**1.79x10^{-06}**
Fat %	37220	-0.0903 ± 0.0045	1.264	2.219	**3.98x10^{-90}**
Fat yield	37191	-0.0071 ± 0.0009	0.190	0.003	**3.16x10^{-14}**
Protein yield	37202	0.0027 ± 0.0007	0.191	0.002	0.0002
Protein %	37266	-0.0103 ± 0.0020	0.080	0.144	**4.28x10^{-07}**
Lactose %	30302	0.0182 ± 0.0012	0.889	1.911	**3.40x10^{-51}**
Lactose yield	30501	0.00801 ± 0.0012	0.181	0.686	**1.94x10^{-11}**

Table 6.7 Chromosome 27 milk composition association statistics for top WGS derived variants
This table presents the top variant and the significance of its association for each milk composition trait. In the fifth column, the rank of VNTR amongst the markers is given. The significance of its association conditioned on the VNTR is presented in the sixth column well as the R^2 of the variant with the VNTR. Bonferroni threshold $P=1.42x10^{-05}$

Phenotype	Top SNP	Position	P-value	VNTR Rank	Cond. P-value	R^2 with VNTR
Milk yield	rs108992692	36172308	$4.50x10^{-13}$	304	**$5.32x10^{-08}$**	0.41
Fat %	rs208396531	36223974	$1.89x10^{-96}$	35	**$4.86x10^{-10}$**	0.81
Fat yield	rs210746953	36235730	$3.51x10^{-15}$	17	0.0181	0.81
Protein yield	rs109021925	36257230	$0.282x10^{-07}$	225	0.0004	0.11
Protein %	Chr27_35915889	35915889	$9.46x10^{-16}$	717	**$5.26x10^{-15}$**	0.61
Lactose %	rs211401126	36203904	$5.25x10^{-52}$	14	0.3994	0.11
Lactose yield	rs108992692	36172308	$3.38x10^{-16}$	85	**$6.67x10^{-07}$**	0.41

6.5.3. Targeted CRISPR-Cas9 editing of AGPAT6 locus

As part of the detailed investigation of the *AGPAT6* locus conducted by Littlejohn et al., (2014), a strong expression-QTL (eQTL) was identified in the lactating mammary gland. Importantly, the expression of *AGPAT6* transcripts was associated with VNTR genotype, with the VNTR 'T' allele associated with both increased milk fat percentage and increased mammary *AGPAT6* expression (M. D. Littlejohn, Tiplady, et al., 2014). Based on this observation, CRISPR-Cas9 genome editing in bovine mammary cells was attempted to determine which of the high-priority candidate causative variants; VNTR, chr27:36211257GA>T, and chr27:36212352G>A, proposed by Littlejohn et al., (2014), Wang et al., (2012), and Daetwyler et al., (2014), are responsible for the gene expression phenotype. As these variants are in near perfect linkage disequilibrium (LD), the approach was taken to generate clonal CRISPR-Cas9 mammary cell lines to assess the effect of individual alleles outside of their native haplotypic context.

To generate CRISPR-Cas9 edited clonal lines, the first step was to establish and screen large numbers of colonies for the presence of the targeted edits. Five 96-well plates of single-cells were sorted for each target variant (N=480 cells per variant). The sorting of GFP positive cells based on high, low and a mix of GFP fluorescence did not appear to affect the proportion of cells that survived the sort and were able to establish a colony (Table 6.9). Of the 480 FAC-sorted single cells for each target, 179 (37.3%), 190 (39.6%), 203 (42.3%), 107 (20.2%) and 183 (38.1%) colonies were derived for the 36198117GGC, 36198117T, 36211257GA, 36211257T and 32212352G/A targets, respectively (Table 6.9).

6.5.3.1. Testing direct lysis conditions for PCR amplification

The extraction of genomic DNA from individual clones presents a significant bottleneck prohibiting the high-throughput screening of genome edited cell lines. To address this, the approach was taken to use a cell lysis buffer amenable to performing PCR directly on the crude cell lysates. Preliminary tests to compare the optimal concentration of cells revealed that 5 μL of the crude cell lysates from 1,000, 5,000, 10,000 and 20,000 mammary cells were sufficient for PCR amplification of the genomic DNA encompassing the VNTR locus (Figure 6.4). Based on these results, 5 μL crude cell lysate from each of the individual colony was used as input for PCR amplification of the genomic DNA encompassing the three target loci.

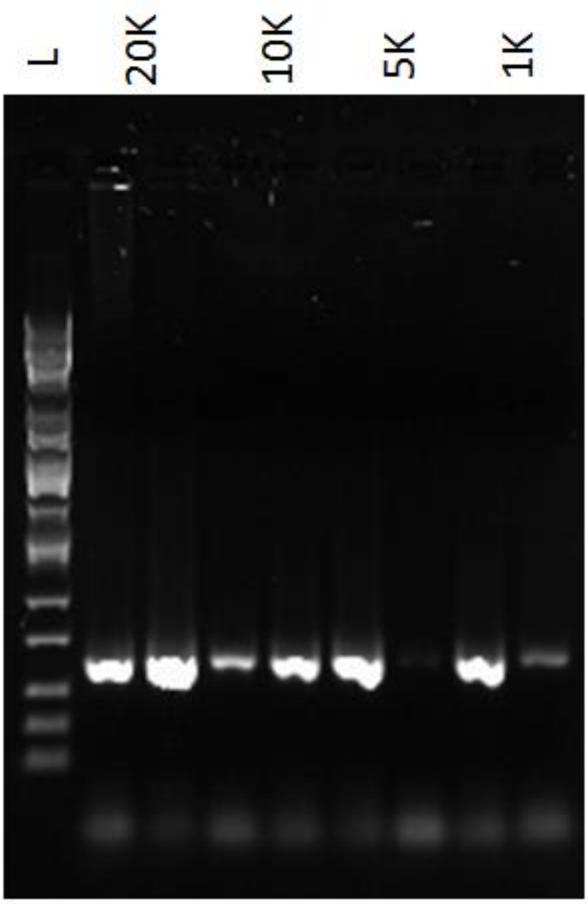


Figure 6.4 PCR screens on cell lysates from 20,000 (20K), 10,000 (10K), 5,000 (5K), and 1,000 (1K) cells
Each PCR was conducted in duplicate and used 5 µL of cell lysate as input in a 25 µL reaction. L = Kapa Universal Ladder**.**

6.5.3.2. Genotyping clonal cell lines with high depth sequencing reveals no HDR editing at *AGPAT6* locus

Following the clonal expansion of these cells, the next step was to genotype each colony for the target variants. Genomic DNA flanking the target variants was amplified and subjected to high-throughput sequencing, yielding an average of 513.2X mapped read depth, with a median 581X for the 862 samples (with depth ranging from 0X to 1080X). Genotypes called directly from these alignments revealed no colonies carrying alternative genotypes at any of the three target sites, suggesting no CRISPR-Cas9 mediated HDR (or even NHEJ) editing had occurred.

A second attempt was made to generate CRISPR-Cas9 edited cell lines by conducting a single cell sort from cell pools previously shown to contain CRISPR-Cas9 edited cells for each of the target variants (described in Chapter 3). Surprisingly, the survival rate of these cells was just 3.13% following the sorting protocol as previously described (without sorting on GFP fluorescence). Based on these limited colony numbers, and the low expected rate of HDR, genotype screening of these cells was not conducted.

Table 6.8 Clonal CRISPR-Cas9 edited mammary cell lines

CRISPR	Target	Plate	# of wells	GFP exp	Target	Plate	# of wells	GFP exp
VNTR	TGGC	1	36	Low	T	1	42	Low
		2	34	High		2	33	Low
		3	40	Mix		3	37	High
		4	37	Mix		4	41	High
		5	32	Mix		5	37	Mix
Total			179 (37.3%)				190 (39.6%)	
36211257	GA	1	40	Low	T	1	21	Low
		2	43	Low		2	25	Low
		3	41	High		3	25	High
		4	41	High		4	16	High
		5	38	Mix		5	20	Mix
Total			203 (42.3%)				107 (20.2%)	
36212352	G/A	1	35	Low				
		2	37	Low				
		3	41	High				
		4	34	High				
		5	36	Mix				
Total			183 (38.1%)					

6.6. Discussion

As part of a detailed study at the chromosome 27 milk fat percentage QTL, we reported strong association of *AGPAT6* locus polymorphisms, and further demonstrated a strong *AGPAT6* eQTL in the lactating mammary gland (M. D. Littlejohn, Tiplady, et al., 2014). This study provided the first evidence for the mechanism by which this gene mediates its effect on milk fat percentage, namely expression promotion impacted by non-coding regulatory variants. This chapter describes aspects of the work reported in Littlejohn et al., (2014), with a particular focus on investigating an *AGPAT6* exon 1 VNTR polymorphism as a functional candidate for the expression and milk composition QTLs.

6.6.1. Genotyping of the *AGPAT6* VNTR in the FJXB population

The high GC content (>80%) of the DNA sequence surrounding the *AGPAT6* VNTR prevented the genotyping of this variant using the high-throughput Sequenom platform. As a result, this variant was interrogated in the FJXB cohort using a custom, PCR-based GeneScan assay. To provide baseline genotype information, and given the importance of the six F1 sires to the FJXB population, these animals were targeted for PCR and Sanger sequencing. This approach can be considered the gold-standard for mutation screening and genotyping (Fitarelli-Kiehl et al., 2016), with this information used to optimise the GeneScan assay, and subsequently apply this cheaper, higher throughput method to genotyping the F2 dams of the FJXB cohort (N= >800). Using this method, this otherwise problematic region was able to be genotyped in 92.9% of the F2 animals.

Ultimately, the direct genotyping of this variant facilitated association analysis between the VNTR and milk fat percentage in the F2 animals. Of the 332 genetic variants in the 1 Mbp window centred on *AGPAT6* reported in Littlejohn et al., (2014), the VNTR was the most highly associated genetic marker with milk fat percentage. These results, and the status of the variant as an indel in a highly conserved 5'UTR sequence in *AGPAT6* exon 1, strongly implicate the variant as driving the eQTL for that gene, and consequent changes in milk composition traits (Littlejohn et al., 2014).

6.6.2. *AGPAT6* VNTR associates with milk fat percentage

Association analysis using 3915 imputed WGS variants in the 1 Mbp surrounding *AGPAT6* VNTR revealed a strong effect on milk fat percentage in a large, independent NZ dairy population. The *AGPAT6* VNTR was one of the most highly associated polymorphisms in this analysis, though was not the most significant. This finding does not preclude the VNTR as being causally involved in the QTLs, but leaves the door open regarding the role of other linked variants, or the potential for multiple genetic effects. Regarding the latter, residually significant association signal remained when the VNTR was fitted as a fixed effect in the association models for some traits. This might indeed suggest the presence of multiple QTLs, though alternatively might suggest an issue with the accuracy of the VNTR genotypes (i.e. that they fail to fully capture the genetic variance at the locus due to problems with genotype imputation). The latter scenario seems plausible, since as mentioned previously, the genomic sequence encompassing the VNTR variant is highly GC-rich, and a paper by Kemper et al., (2015) failed to type this variant in an independent sequence-based analysis that had otherwise detected the *AGPAT6* QTL. The variant had limited depth of sequence representation in many of our WGS samples, and 38 of 556 animals had no calls at this position. Together, these observations suggest the VNTR may not have been accurately imputed for this study, making benchmarking of the variant with other linked variants difficult.

Possibilities to address this issue for future analysis would be to incorporate the manual VNTR genotypes from the FJXB population into the imputation strategy, or physically genotype the variant in large numbers of samples directly. Re-running the association analysis incorporating higher quality genotype calls for the VNTR should increase the power of the analysis, though given long range LD effects in cattle, it is unlikely the variant can be definitively implicated for the QTL using statistical genetics approaches alone. For this reason, experiments to directly test the function of the variant *in vitro* were pursued (summarised below).

6.6.3. Towards the generation of clonal CRISPR-Cas9 edited cell lines

The CRISPR-Cas9 experiment was undertaken with the primary purpose of establishing bovine mammary cell lines representing each genotype for the three candidate causative variants proposed to underpin the *AGPAT6* QTL.

Prior to this experiment, the conditions for the amplification of genomic DNA directly from crude cell lysates were optimised. This was a lysis solution that did not inhibit the polymerase enzyme (i.e. with no EDTA; previously used by other members of our laboratory) and the cell numbers and the amount of lysate required for amplification was determined as 5 µL of lysate from as little as 1,000 cells. The establishment of these conditions was important to circumvent the bottleneck of genotyping clones for high-throughput purposes (Ramlee, Yan, Cheung, & Chuah, 2015).

Following the establishment of these conditions, the approach was taken to generate clonal lines from cell pools transfected with CRISPR-Cas9 RNPs, ssODN HDR templates and a GFP plasmid. The majority of publications report the use of a selectable marker (e.g. antibiotic resistance or fluorescent marker) which is on the same plasmid as Cas9 (Patrick et al., 2014) to increase the editing efficiency up to 50% (Böttcher et al., 2014; Richardson et al., 2016). However, based on previous work establishing the conditions for the efficient CRISPR-Cas9 editing of the *AGPAT6* locus using the Cas9 protein (refer to Chapter 3), the co-transfection of a separate GFP plasmid was used in this experiment. Those cells positive for GFP fluorescence were clonally expanded and screened for the presence of HDR at three target variant sites. Analysis of genomic DNA isolated from individual colonies by PCR and sequencing revealed that there had been no HDR for the three *AGPAT6* candidate causative variants. Reasons for this remain unclear at this stage, particularly given I previously demonstrated a high CRISPR-Cas9 mediated editing efficiency at these sites (Chapter 3). It is possible that technical aspects of the particular transfection in this experiment contributed to this result; and unfortunately, the absence of unsorted controls in the experiment make it impossible to tell whether the lack of edited cells derive from a problem with the transfection step, or are a consequence of cell sorting. In this respect, future experiments to troubleshoot the lack of editing would consist of retaining and sequencing an aliquot of CRISPR-treated cells not subjected to single cell sorting, and given the routine use of this method to isolate isogenic cell lines in the literature (Grobarczyk et al., 2015; F Ann Ran et al., 2013; T. Wang et al., 2014), could be expected to lead to the successful generation of cell lines for *AGPAT6* expression testing.

6.6.4. Limitations and future directions

The work presented in this chapter provides further evidence of the influence of the *AGPAT6* locus on milk composition. However, despite extensive genetic and functional experiments to probe the identity the causative variant at this locus, these analyses were inconclusive. Despite this, these findings do provide information that can be leveraged to conduct further functional experiments, with strategies to revise the CRISPR-Cas9-based approach, or utilisation of other new technologies to the same end. These include methods such as the massively parallel reporter assay, which is comparatively high throughput, yet also allows the direct testing of potential regulatory variants (Tewhey et al., 2016).

In retrospect, the design of cross-repairing HDR templates may have meant the Cas9 endonuclease continued to cleave the genomic DNA once the DSB was introduced and repaired, contributing to the lack of HDR detected in this experiment. As it has recently been reported that modifying the PAM site can reduce this re-editing to increase the rate of HDR editing by 10-fold per allele (Paquet et al., 2016), the use of HDR templates that blocks the PAM site may yield improved editing accuracy. Given the importance of testing regulatory variants in their native genomic context, future experiments could leverage two rounds of gene editing whereby the initial round introduces a 'blocking mutation' (modifying the PAM site) along with the targeted sequence change and the subsequent clones are screened to find the ones with the intended change. The following round of editing would then edit these cells using a second repair template that corrects the blocking mutation and again screening of these clones would identify those containing 'scarless' targeted HDR editing (Kwart et al., 2017).

6.7. Summary and conclusions

In *Bos taurus*, a chromosome 27 locus, underpinned by an *AGPAT6* eQTL, has a large influence on milk fat percentage and other milk composition phenotypes. Work reported here describes the investigation of the locus, with the aim of characterising the causative variant responsible for the QTL. While this work made steps towards this goal, the causative variant was not definitively confirmed. The work does, however, provide proof-of-principle for the generation of CRISPR-Cas9 edited mammary cell lines, with methods optimised that should allow direct interrogation of the candidates proposed by the various research groups working with this QTL (Littlejohn et al., 2014; Daetwyler et al., 2014, Wang et al., 2012).

Chapter 7: Functional confirmation of *PLAG1* as the causative gene underlying major pleiotropic effects on liveweight and milk characteristics

The following work was published as: Fink, T. et al. (2017). Functional confirmation of *PLAG1* as the candidate causative gene underlying major pleiotropic effects on body weight and milk characteristics. *Scientific Reports*. 7(44793), 1-8. http://doi: 10.1038/ srep44793). Because this chapter is based on this publication, there is some repetition of methods already presented in Chapter 2.

The co-authors of this work contributed to conception and design of the experiments as well as the analysis tools I used for this project (as outlined in the co-authorship form included with this thesis). I designed and performed all experiments and conducted the data analysis.

7.1 Overview

Stature and body weight represent economically important traits in cattle. In beef animals, maximizing growth and development holds obvious importance for meat production. Despite modest positive correlations of body size with milk volume, protein, and fat yield (Brotherstone, 1994), smaller or larger animals may be desirable in dairy farming contexts, depending on management considerations. Stature and body weight are both highly heritable (Kathryn E Kemper et al., 2012), and large scale genetic studies have identified chromosomal regions impacting these traits in both *Bos taurus* and *indicus* species. Of these, a quantitative trait locus (QTL) on chromosome 14 with a major effect on stature and body weight was reported by Karim et al. in 2011, and has since been observed in many independent populations (Fortes, Kemper, et al., 2013; Hoshiba et al., 2013; M. Littlejohn et al., 2012; Nishimura et al., 2012).

In a detailed analysis of this chromosome 14 locus (Karim et al., 2011), fine mapping yielded 13 candidate polymorphisms as potentially underlying the QTL, none of which mapped to protein coding sequences. Further functional and genetic analysis reduced the number of candidate variants to two polymorphisms in the bidirectional promoter of the *PLAG1* and *CHCHD7* genes; rs209821678, a (CCG) repeat of 9 or 11 copies and rs210030313 an A to G nucleotide substitution (Karim et al., 2011). In this analysis, both variants were associated with foetal expression of seven of the nine genes within a ~780 kilobase (kb) interval representing the stature QTL. Of these genes, *PLAG1*, *RPS20*, and *SDR16C5* were plausible biological candidates for these effects, with demonstrated roles in growth and oncogenesis (Lettre et al., 2008; McGowan et al., 2008; Voz et al., 2000). In particular, *PLAG1* represented an obvious candidate given *plag1* knockout mice suffer from slow growth rates and dwarfism (Hensen et al., 2004). This gene encodes a transcription factor that regulates several growth factors including IGF2 (Van Dyck, Declercq, Braem, & Van de Ven, 2007; Voz et al., 2000), a key modulator of growth in both dogs and humans (De Vos et al., 2008; DeChiara, Efstratiadis, & Robertson, 1990). Despite these observations, given the differential expression of multiple genes at this locus, the causative status of these genes remains to be resolved.

In the current study, we have used a mammary RNAseq dataset (N = 375) to perform expression QTL (eQTL) mapping in a 2 megabase (Mbp) interval encompassing the previously implicated chromosome 14 locus. We report genetic effects for a subset of the nine genes of interest, and further report investigation of milk composition and body weight effects in a separate population (N = 39,391) of lactating cows.

7.2 Results

7.2.1 eQTL analysis of the chromosome 14 locus

To quantify the expression levels of the nine genes of interest at the chromosome 14 body weight locus, high-depth, mammary RNAseq data from 375 lactating cows was assessed. Five of these nine genes were appreciably expressed, consisting of *CHCHD7, IMPAD1, LYN, PLAG1* and *RPS20*. Of these, *RPS20* was by far the most abundantly expressed (109.6 fragments per kilobase of exon model per million mapped; FPKM), and *PLAG1* was relatively lowly expressed (0.8 FPKM).

Association analysis of these five genes was then conducted to test for eQTL effects. In these analyses, transformed read counts for each gene were tested in conjunction with 432 Illumina BovineHD SNPs located in a 2 Mbp interval encompassing the previously published QTL interval. This region was centred on the BovineHD panel variant rs109815800 (Chr14:25015640, UMD3.1 genome build), selected as a tag-SNP of the two putative causative variants, since this SNP has been shown to be in complete LD with these polymorphisms in (New Zealand) NZ Holstein-Friesians, Jerseys and their crosses (M. Littlejohn et al., 2012). Restricted maximum likelihood analysis using pedigree-based mixed models revealed significant eQTLs for three of five expressed genes in the QTL interval (Figure 1a; Table 1). However, only *PLAG1* and *LYN* were significantly differentially expressed by rs109815800 genotype. For *PLAG1*, rs109815800 was the most significant SNP ($P=1.33\times10^{-23}$; Figure 1a and b; Table 2), which was 17 orders of magnitude more significant than the association with *LYN* expression ($P = 1.15\times10^{-6}$; Fig. 1a and c Table 2).

Notably, the genetic signature of the *LYN* eQTL appeared to differ from the *PLAG1* eQTL, with a different lead SNP ($P = 1.71\times10^{-7}$; rs109116062; chr14:24909247; Fig. 7.1a,b and c; Table 7.1), and with the rs109815800 variant appearing lower in the association rankings for *LYN*. The rs109815800 SNP explained 46.6% of the genetic variance and 32.6% of the phenotypic variance in *PLAG1* expression, and 45.2% and 9.6% respectively for expression of *LYN* (Table 2). This contrasted with 61.6% and 9.9% respectively for the lead *LYN* SNP rs109116062, again suggesting that the *PLAG1* and *LYN* eQTLs might be being driven by different or overlapping (i.e. functionally independent, yet genetically linked) effects. The frequency of the body weight-increasing 'G' allele was 0.68 in the mostly Holstein-Friesian RNAseq population.

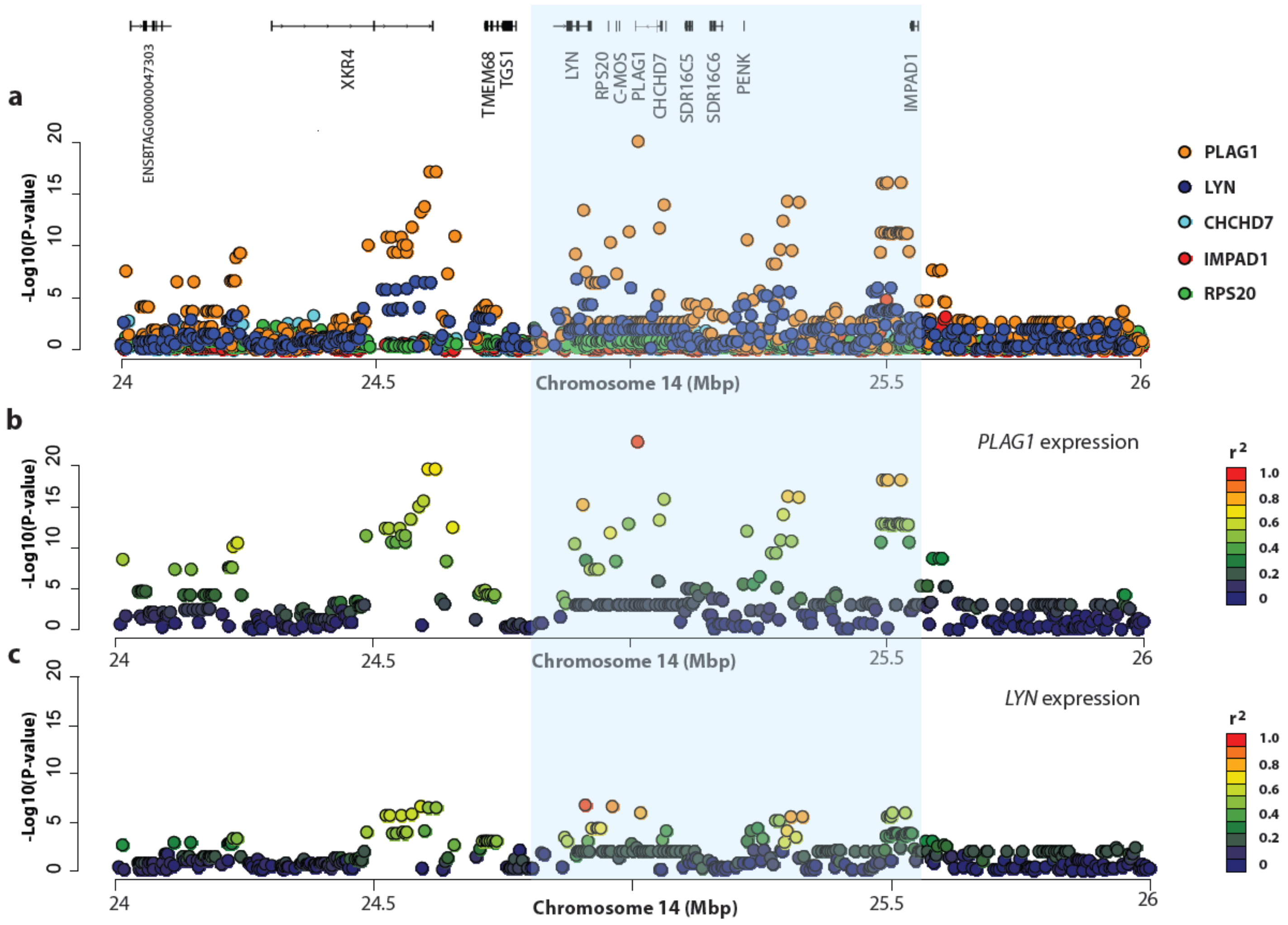


Figure 7.1 Expression QTL analysis at the chromosome 14 locus

(a) shows layered Manhattan plots for the five nominally expressed genes at the previously reported, ~780 kb body weight locus (blue-shaded area) in RNA-sequenced animals. The X-axis shows bp position on chromosome 14, the Y-axis shows −log10 P-values of marker association. The location and structure of 14 genes mapping to the broader, 2 Mbp interval are shown at the top of field. (b) indicates the marker association of the 432 SNPs with PLAG1 expression. (c) indicates marker association with LYN expression. The top gene expression-associated SNP is coloured red in (b) and c, with other variants coloured according to their linkage disequilibrium relationship with these SNPs.

Table 7.1 Chromosome 14 body weight locus effect on gene expression

For each of the five mammary-expressed genes in the 2 Mbp interval of interest, the top-associated SNP, its position and P-value of association is indicated. Significant effects are bolded; multiple testing significance threshold P=2.31×10^{-5}

Gene	Top SNP	Location on chr 14 (bp)	P-value
PLAG1	**rs109815800**	**25015640**	**1.33x10^{-23}**
LYN	**rs109116062**	**24909247**	**1.71x10^{-07}**
CHCHD7	rs42648880	24378496	4.33x10^{-04}
RPS20	rs134518689	24278284	0.008
IMPAD1	**rs110632518**	**25501417**	**1.69x10^{-05}**

Table 7.2 Association between rs109815800 and the expression of genes at the chromosome 14 body weight locus

Effect estimates are given with standard errors in units of VST read counts. For significant effects (bolded), the 'Pheno var explained' and 'Geno var explained' columns represent the percentage of phenotypic and genotypic variance accounted for by the rs109815800 SNP. P-values of genetic association are indicated in the right-most column. Significance threshold P=2.31×10^{-5}

Gene	Effect ± SE	Pheno var explained	Geno var explained	P-value
PLAG1	**−0.5293 (±0.0491)**	**32.59**	**45.59**	**1.33x10^{-23}**
LYN	**0.0787 (±0.0159)**	**9.61**	**45.22**	**1.15x10^{-06}**
CHCHD7	−0.0305 (± 0.0165)	N/A	N/A	0.065
RPS20	−0.0179 (± 0.0251)	N/A	N/A	0.476
IMPAD1	0.0059 (± 0.0193)	N/A	N/A	0.760

7.2.2 Analysis of body weight and milk composition effects

Having observed differential expression of *PLAG1* and *LYN* by rs109815800 genotype, we wondered whether these eQTL might have phenotypic consequences in the mammary gland. To answer this question, we used a population of 39,391 lactating cows for which both milk production and body weight data were available, using association models similar to those used for eQTL analysis. Given the major milk production effects attributed to another (albeit distant) chromosome 14 mutation, the *DGAT1* K232A variant (Grisart, Coppieters, & Farnir, 2002), these models also incorporated imputed *DGAT1* K232A genotypes as a fixed effect. Targeting the same interval of BovineHD SNPs used previously, analysis of body weight confirmed a very large effect at this locus, with the causative mutation tag-SNP rs109815800 the most significant variant ($P<2.2\times10^{-308}$, Table 7.3). This SNP accounted for 32.8% of the genetic variance and 15.8% of the phenotypic variance in body weight for this population, and notably, the body weight-increasing 'G' allele was the same allele associated with increased *PLAG1* gene expression.

Next, we conducted association mapping of milk volume, milk fat and protein yield, and milk fat and protein percentage traits. The frequency of the body weight-increasing 'G' allele was 0.46 in this population of mixed breed cattle. Significant effects were observed for all traits except milk protein percentage (Figure 7.2a; Table 7.3), where the body weight increasing rs109815800 'G' allele was associated with increased milk volume and milk protein and fat yield, and decreased milk fat percentage. Given the magnitude of association of rs109815800 genotype with body weight, we reasoned that some of these associations might be due to differences in animal (and thus mammary) size (Morris & Wtlron, 1976). To attempt to differentiate these effects from those that might be impacting secretory pathways and/or energy utilisation irrespective of size, we conducted an alternative analysis that fitted animal body weight as an additional covariate in the association models. Interestingly, rs109815800 showed significant associations for all milk traits using these models (Table 7.4), including a highly significant reduction in milk fat yield, and decreases in milk volume and protein yield in animals carrying the *PLAG1* high-expression 'G' allele. The sign of effect was reversed for milk volume in these body weight-adjusted models, suggesting that the increased size of animals carrying rs109815800 'G' alleles was indeed playing a role in the milk production effects observed at this locus.

Since the chromosome 14 body weight QTL minor allele is known to differ between Holstein-Friesian and Jersey breeds (Karim et al., 2011), and that this might be expected to lead to population stratification and confound analyses in a mixed breed population, we also conducted analysis of body weight and lactation traits in purebred animals. These analyses used purebred-segregated (16/16ths breed proportions) Holstein-Friesian (N=8086) and Jersey (N=4322) subpopulations, and confirmed significant effects for a subset of the traits (Tables 7.3 and 7.4). Although body weight was the only significantly associated phenotype in Jersey animals, statistical power was limited in this population due to a low MAF (0.02). With the exception of milk fat yield in the body weight-unadjusted models (Table 7.3), the direction of effects was otherwise identical across all populations and traits (including significant and non-significant associations alike). These observations suggested that association-confounding by breed was unlikely to be a major issue in these analyses, further supporting the role of *PLAG1* as a gene with pleiotropic impacts on these traits.

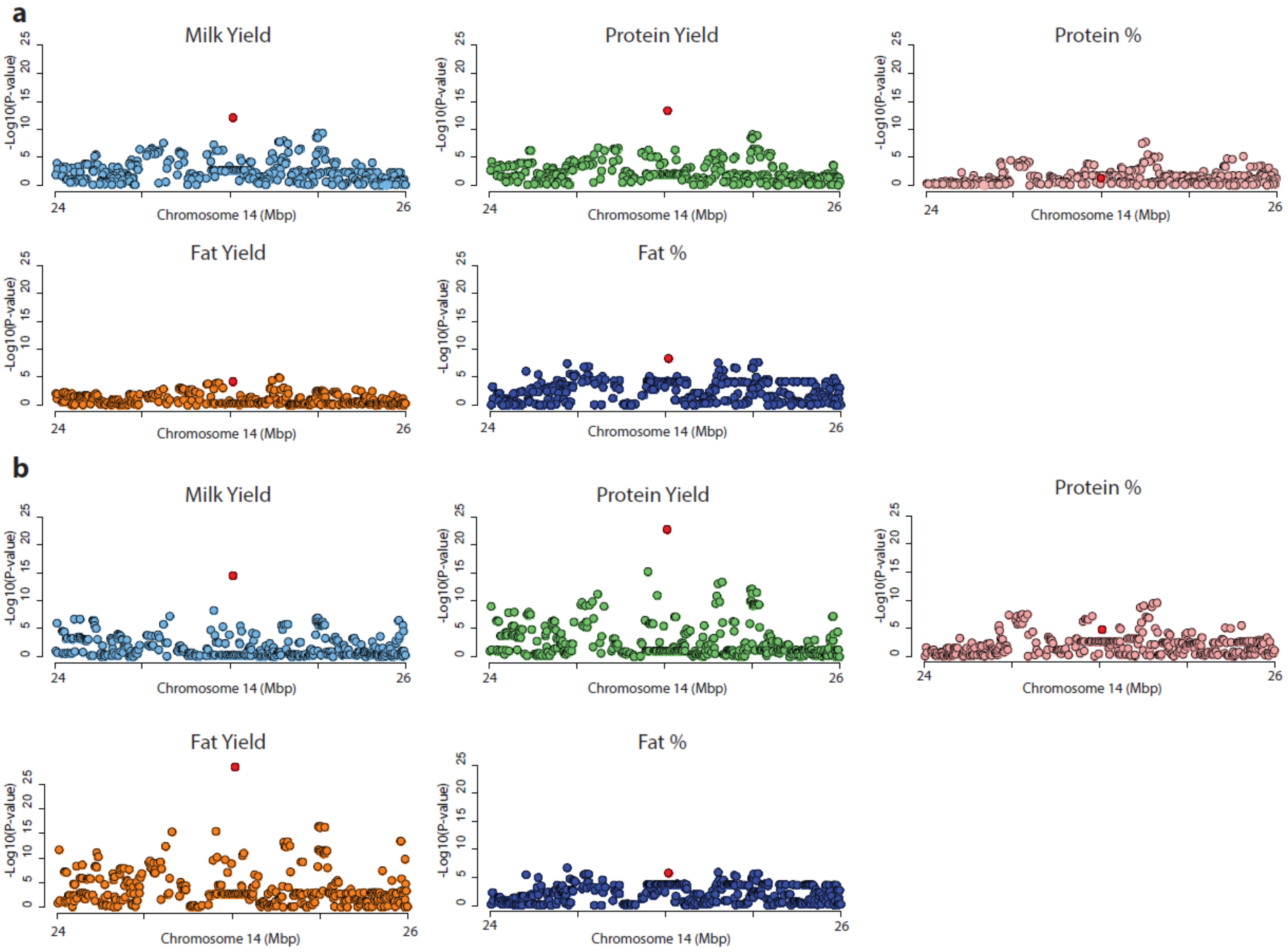


Figure 7.2 Milk composition QTLs at the chromosome 14 locus

(a) represents the Manhattan plots for the five milk composition and yield phenotypes, with the X-axis showing Mbp position on chromosome 14, and the Y-axis showing −log10 P-values of marker association. (b) represents similar plots where animal body weight has been fitted as an additional covariate in otherwise identical association models. The rs109815800 SNP is coloured red in both (a) and (b).

Table 7.3 Association between rs109815800 and milk composition traits and body weight in NZ dairy cows

Association results for the rs109815800 SNP with milk composition and body weight phenotypes are shown for models fitted across all animals and when segregated to Holstein-Friesian and Jersey breeds. Effects are expressed for each 'T' allele relative to homozygous 'G' animals, displayed using units of grams for yield traits, litres for milk volume, and kilograms for body weight. For significant effects (bolded), 'Pheno var explained' and 'Geno var explained' columns represent the percentage of phenotypic and genotypic variance accounted for by the rs109815800 SNP. Significance is based on a Bonferroni-adjusted threshold of P=1.51×10^{-3}

	All animals					Holstein-Friesian			Jersey		
Phenotype	**N**	**Effect ± SE**	**Pheno var explained**	**Geno var explained**	**P-value**	**N**	**Effect ± SE**	**P-value**	**N**	**Effect ± SE**	**P-value**
Milk volume	**39380**	**−0.0815 ± 0.0114**	**0.239**	0.879	**8.96x10^{-13}**	8082	− 0.0841 ± 0.027	1.98x10^{-03}	4321	− 0.094 ± 0.079	0.234
Fat %	**39391**	**0.0198 ± 0.0034**	**0.155**	0.309	**4.21x10^{-09}**	**8086**	**0.0401 ± 0.0067**	**2.48x10^{-09}**	4322	0.0134 ± 0.0296	0.650
Fat yield	**39374**	**−1.628 ± 0.404**	**0.076**	0.345	**5.52x10^{-05}**	8084	0.2218 ± 0.8828	0.802	4321	− 3.552 ± 3.132	0.257
Protein %	39391	0.0028 ± 0.0017	N/A	N/A	0.098	8086	0.0123 ± 0.0035	3.84x10^{-03}	4322	0.0186 ± 0.0143	0.194
Protein yield	**39376**	**−2.293 ± 0.304**	**0.267**	**1.239**	**4.92x10^{-14}**	8082	− 1.518 ± 0.709	0.032	4320	− 2.081 ± 2.238	0.352
Body weight	**39391**	**−17.25 ± 0.27**	**15.80**	**32.78**	**<2.23x10^{-308}**	**8086**	**−15.72 ± 0.633**	**4.01x10^{-131}**	**4321**	**−19.997 ± 1.939**	**1.20x10^{-24}**

Table 7.4 Association between rs109815800 and milk composition traits conditioned on body weight in NZ dairy cows
Association results for the rs109815800 SNP with milk composition phenotypes conditioned on body weight are shown for models fitted across all animals and when segregated to Holstein-Friesian and Jersey breeds. Effects are expressed for each 'T' allele relative to homozygous 'G' animals, displayed using units of grams for yield traits, and litres for milk volume. For significant effects (bolded), 'Pheno var explained' and 'Geno var explained' columns represent the percentage of phenotypic and genotypic variance accounted for by the rs109815800 SNP. Significance is based on a Bonferroni-adjusted threshold of P=1.51×10^{-3}

	All animals					**Holstein-Friesian**			**Jersey**		
Phenotype	**N**	**Effect ± SE**	**Pheno var explained**	**Geno var explained**	**P-value**	**N**	**Effect ± SE**	**P-value**	**N**	**Effect ± SE**	**P-value**
Milk volume	**39380**	**0.0918 ± 0.0116**	**0.322**	**1.226**	**3.17x10^{-15}**	8082	0.0808 ± 0.0274	3.17x10^{-03}	4321	0.086 ± 0.078	0.27
Fat %	**39391**	**0.017 ± 0.0035**	**0.115**	**0.23**	**1.49x10^{-06}**	**8086**	**0.0341 ± 0.007**	**1.04x10^{-06}**	4322	0.0127 ± 0.0299	0.671
Fat yield	**39374**	**4.615 ± 0.412**	**0.65**	**3.087**	**4.37x10^{-29}**	**8084**	**5.411 ± 0.891**	**1.30x10^{-09}**	4321	4.178 ± 3.08	0.175
Protein %	**39391**	**0.0078 ± 0.0018**	**0.091**	**0.165**	**1.78x10^{-05}**	**8086**	**0.0156 ± 0.0036**	**1.50x10^{-05}**	4322	0.0222 ± 0.0145	0.125
Protein yield	**39376**	**3.066 ± 0.307**	**0.516**	**2.614**	**2.11x10^{-23}**	**8082**	**3.311 ± 0.706**	**2.76x10^{-06}**	4320	3.586 ± 2.196	0.103

7.3 Discussion

We report a strong mammary eQTL for *PLAG1* which bears the same genetic signal underpinning the body weight and developmental QTLs reported for this locus. To our knowledge, these data represent the first functional confirmation of this expression-based effect. While contrary to previous analysis in foetal tissues showing cis eQTL for multiple genes (Karim et al., 2011), the current analysis suggests *PLAG1* alone as responsible for these effects. We additionally report new associations of *PLAG1* genotype with milk composition and yield phenotypes, adding lactation effects to the long list of physiological traits that are impacted by this locus.

Karim et al. (2011) first reported the presence of eQTLs underpinning the bovine stature locus on chromosome 14. They performed quantitative PCR using foetal brain, bone, muscle and liver samples representing 79 individuals and found significant associations with the expression of *RPS20*, *MOS*, *PLAG1*, *CHCHD7*, *SDR16C5*, *SDR16C6* and *PENK* genes. While it is possible that all genes in this interval are affected by a single control element in foetal tissues, it is also plausible that these associations were due to the genotype tagging multiple independent eQTL, given that association testing was restricted to analysis of a single variant. Critically, of the seven candidates above, only *PLAG1* was significantly differentially expressed by body weight QTL genotype in mammary tissue, with both eQTL and physiological trait QTL sharing the same top-associated SNP. Although we also observe an association with *LYN* expression, the rank order of associated SNPs suggests these QTLs may be driven by a different genetic element, and given that no eQTL was reported in foetal tissues, *LYN* can likely be discounted as a candidate for the stature and body weight effects. Taken together, these observations provide further evidence for *PLAG1*, likely under regulatory control of the rs209821678 and/or rs210030313 variants proposed be Karim et al. (2011), as the causative gene responsible for these effects. Since mammary tissue was used for eQTL mapping in the current study, this hypothesis assumes a shared regulatory architecture between this tissue and tissues of more direct relevance to growth and development processes. As such, an alternative hypothesis proposes the lactation effects being driven by *PLAG1*, and the previously reported stature and body weight QTLs being underpinned by one of the other candidate genes highlighted by Karim et al. (2011). Although technically conceivable, we contend the simplest and most plausible hypothesis is one that sees a single, major-effect pleiotropic *PLAG1* eQTL underpinning all physiological effects.

Given observation of the mammary eQTL effects, and the many other pleiotropic observations at this locus, we wondered whether differential expression of *PLAG1* might impact lactation traits directly. We analysed milk volume, milk fat percentage and yield, and milk protein percentage and yield, and found highly significant associations of rs109815800 with all traits except protein percentage. The body weight-increasing 'G' allele was associated with increased milk volume, fat, and protein yield, and decreased fat percentage. This opposing sign of effect between component yields and percentages could reflect these animals producing a higher volume of milk relative to the increases seen in the milk components, resulting in milk that is marginally more dilute. This phenomenon of milk component and yield effects being co-ordinately impacted is something that we (M. D. Littlejohn et al., 2016; M. D. Littlejohn, Tiplady, et al., 2014), and others (K.E. Kemper, Hayes, Daetwyler, & Goddard, 2015), have observed for other major QTL previously.

However, given the profound impact of this locus on animal stature and body weight, and the fact that larger cattle produce more milk (Holmes et al., 2002), we reasoned that volume effects might reflect differences in mammary size and capacity. As expected, adjusting for animal body weight in the association models revealed that this initial association was likely driven by differences in animal size, and notably, the sign of the SNP effect was reversed in this model. This observation was apparent for milk volume, protein, and fat yield traits, where the 'G' allele was associated with decreased yields, suggesting an efficiency advantage to the alternate allele. The apparent reduction in milk fat yield in animals carrying the allele normally associated with increased body weight was the most significant effect in these models and is of further note, since effects on reduced intramuscular fat and fat deposition have also been reported for this allele (Fortes, Kemper, et al., 2013). The mobilisation of body fat reserves to support the greater energy requirements of the lactating mammary gland is well described (Bauman & Currie, 1980; Friggens, Ingvartsen, & Emmans, 2004). We speculate that this reduction in milk fat may be due to differential energy utilisation between genotypes, whereby energy normally partitioned into subcutaneous fat or milk triglyceride synthesis shifts to a balance favouring increased lean tissue mass. Additional physiological indicators of energy balance and lipolysis could be examined in animals of contrasting *PLAG1* genotype to further test this hypothesis.

It is interesting to contemplate what mammary-specific pathways may be involved in the lactation effects proposed in our study. Two well-demonstrated targets of PLAG1 signalling include molecules of the IGF2 and WNT pathways (Voz et al., 2000; Y. Wang et al., 2013), with the former speculated as the underlying mechanism of the growth and body weight effects attributed to this QTL (Juma, Damdimopoulou, Grommen, Van de Ven, & De Groef, 2016; Karim et al., 2011). Transgenic mouse lines engineered to overexpress *plag1* in mammary tissue show differential expression of IGF2 and WNT signalling genes (Declercq et al., 2008), with mammary hyperplasia and development of adenomyoepitheliomas the primary phenotypes of these models. There is limited data to suggest IGF2 may increase milk synthesis in the lactating mammary gland (Prosser, Davis, Farr, Moore, & Gluckman, 1994), though the role of the hormone in mammary development and involution is clearly demonstrated (Cathrin Brisken et al., 2002; Moorehead, Fata, Johnson, & Khokha, 2001). Likewise, WNT signalling is proposed to play important roles in the development and differentiation of the mammary gland during pregnancy (Boras-Granic & Wysolmerski, 2008), and assuming the involvement of PLAG1 in these pathways is relevant outside of a tumorigenic context, the effects demonstrated might derive from morphological differences between animals of different QTL genotype. It is also possible that the milk composition and yield effects may reflect secondary impacts of the QTL deriving from effects in other tissues. Given that *PLAG1* is expressed during lactation and the eQTL is observed during this period, however, another appealing mechanism is one that acts through some as yet unidentified factors with direct, modulatory roles on milk synthesis and secretion.

7.4 Conclusions

In summary, we describe a strong mammary eQTL for *PLAG1* that bears the same genetic signal underpinning the previously described body weight and developmental effects at this locus. We additionally report new associations of *PLAG1* genotype with milk composition and yield phenotypes. These data provide the first functional validation of an eQTL-mediated mechanism underpinning these QTLs, and further expand the list of pleiotropic effects attributed to *PLAG1* in bovine species.

7.5 Methods

7.5.1 Ethics statement

All animal experiments were conducted in strict accordance with the rules and guidelines outlined in the NZ Animal Welfare Act 1999. For the mammary tissue biopsy experiment, samples were obtained in accordance with protocols approved by the Ruakura Animal Ethics Committee, Hamilton, NZ (approval AEC 12845). All other data were generated as part of routine commercial activities outside the scope of that requiring formal committee assessment and ethical approval (as defined by the above guidelines). No animals were sacrificed for this study.

7.5.2 Primary data

Primary datasets consisting of relevant genotypes, and milk production and gene expression phenotypes have been deposited into the Dryad digital data repository (doi: 10.5061/dryad.r8251), and NCBI Short Read Archive (SRP075939).

7.5.3 Animal populations, phenotypes and genotypes

Animals used for RNAseq analysis comprised 375 mostly Holstein-Friesian NZ dairy cows, representing a subset of 406 sequenced animals described in detail previously (M. D. Littlejohn et al., 2016; M. D. Littlejohn, Tiplady, et al., 2014).

High-depth RNAseq was undertaken using mammary gland biopsies from lactating cows as described in detail previously (M. D. Littlejohn et al., 2016). Briefly, 21 of the 375 samples were collected in 2004 and 2012, and were sequenced by NZGL (Dunedin, NZ) using the Illumina HiSeq 2000 instrument. The remaining samples were collected in 2013 and 2014 and were sequenced by the Australian Genome Research Facility (AGRF; Melbourne, Australia) using the Illumina HiSeq 2000 instrument.

RNA sequence reads were mapped to the UMD3.1 genome using Tophat2 (version 2.0.12; Kim et al., 2013) as previously described (M. D. Littlejohn et al., 2016). Cufflinks software (version 2.1.1; Trapnell et al., 2010) was used to quantify expressed transcripts, yielding fragments per kilobase of exon model per million mapped (FPKM) expression values. The genes in the ~780 kb region of interest on chromosome 14 were considered for downstream analysis if they had non-zero FPKM values in at least 75% of samples, and had a mean expression of 0.5 FPKM or greater. To derive gene expression phenotypes suitable for eQTL analysis, the read counts from the nine genes in this interval were also processed using the variance-stabilising transformation (VST) method in DESeq (version 1.18; Anders & Huber, 2010). This transformation addresses issues of heteroscedasticity inherent in RNA-seq data, and normalises the count data to a form suitable for linear model analysis.

The animal population used for GWAS comprised 39,391 dairy cows, consisting of 8,086 Holstein-Friesians, 4,322 Jerseys, and 26,983 Holstein-Friesian x Jersey cross breeds, where Holstein-Friesians and Jerseys were considered pure with a breed proportion of 16/16ths. This population represents part of a larger phenotypic and genotypic database of animals used for evaluation of sire performance, similar to populations described previously (M. D. Littlejohn et al., 2016; M. D. Littlejohn, Tiplady, et al., 2014). Animals were also segregated by breed (as defined above) to assess within-breed effects and potential confounding impacts of population stratification. Slight differences in the animal numbers quoted for each analysis is a reflection of the quality filtering performed on each trait.

Milk composition phenotypes were derived from first lactation herd test data. Concentrations of major milk components were measured using Fourier transform infrared spectroscopy as part of standard herd testing procedures as described in (M. D. Littlejohn et al., 2016; M. D. Littlejohn, Tiplady, et al., 2014). These concentrations were adjusted using linear models with age at calving and stage of lactation as fixed effects and contemporary group as an absorbed/sparse fixed effect. Residuals from these models were used for subsequent association analyses. Body weight measurements were from two sources, either representing a weight where the animal walked over a scale or a weight derived from visual scoring carried out by certified assessors in accordance with published guidelines (DairyNZ, 2014a). Body weight records were restricted to values measured on two year olds in their first lactation and were filtered to remove outliers. Individual estimates for each animal were derived by fitting a repeated measures model in ASReml-R and were used for subsequent association analyses. Fixed effects included in the model were method of measurement (scale weight/inspector weight), age at calving, and stage of lactation with contemporary group fitted as an absorbed/sparse fixed effect.

Genomic DNA extraction was conducted as previously described (M. D. Littlejohn et al., 2016; M. D. Littlejohn, Tiplady, et al., 2014). Briefly, DNA was extracted from either blood or ear-punch tissue or processed using Qiagen Biosprint kits (Qiagen) or a MagMAX system (Life Technologies) by GeneMark (Hamilton, NZ), and GeneSeek (Lincoln, NE, USA), respectively. Genotyping was conducted by GeneSeek (Lincoln, NE, USA), using the Illumina BovineHD BeadChip or BovineSNP50 BeadChip (Illumina) platforms. For samples genotyped on BovineSNP50 chips, these were imputed to the BovineHD platform using Beagle software (Beagle v3.3.2; Browning & Browning, 2009) prior to association analysis, using methods similar to those described previously (M. D. Littlejohn et al., 2016; M. D. Littlejohn, Tiplady, et al., 2014). Briefly, for the small subset of RNAseq animals that had not been physically genotyped on the BovineHD BeadChip (N = 27 cows), imputation was performed on a genome-wide basis for 659,811 SNP using a reference population of 3,460 animals (with 46,805 SNPs overlapping between platforms). For the population of 39,391 cows used for milk composition and body weight analysis, a reference population of 3389 animals was used to impute 675,321 SNPs (with 46,621 SNPs overlapping between platforms). The *DGAT1* K232A variant was imputed from whole genome sequence data using a reference population of 556 animals, in an approach similar to that described previously (M. D. Littlejohn et al., 2016). Linkage disequilibrium statistics for all chromosome 14 SNPs used for association analysis are shown in Appendix IV, calculated on the larger (N = 39,391) of the two animal populations.

7.5.4 Genetic association analysis

For the 2 Mbp interval of Illumina BovineHD SNPs encompassing the rs109815800 variant, associations with gene expression, body weight, and milk composition phenotypes were quantified using pedigree-based mixed models in ASReml-R (A R Gilmour et al., 2009; Arthur R. Gilmour et al., 1995). The RNAseq analysis used a total of 432 SNPs, the body weight and milk composition analyses used a subset of 421 SNPs with the difference reflecting the impact of slightly different imputation and quality-filtering criteria applied between populations. Each SNP was fitted in a separate sire-maternal grandsire single-trait model with the SNP treated as a quantitative variable based on the number of copies of the alternative allele and variance components estimated in a restricted maximum-likelihood (REML) framework. Covariates for birth year, the proportions of NZ Holstein-Friesian ancestry, US Holstein-Friesian ancestry, Jersey ancestry and breed heterozygosity effects were also included in the models. The additive genetic variance for each SNP was calculated using $\sigma^2_{SNP} = 2p(1-p)a^2$, where p is the frequency of the highest frequency allele and a is the estimated allele substitution effect. Polygenic genetic variances were evaluated as $\sigma^2_{anim} = 4\sigma^2_{sire}$ where $\sigma^2_{sire}$is the estimate of sire variance from the model. Total genetic variance was evaluated as $\sigma^2_g = \sigma^2_{SNP} + \sigma^2_{anim}$ and phenotypic variance was evaluated as $\sigma^2_p = \sigma^2_{SNP} + \sigma^2_{anim} + \sigma^2_e$ where σ^2_e is the residual variance. The proportion of phenotypic variance explained by each SNP for each phenotype was calculated as $\sigma^2_{SNP}/\sigma^2_p$ and the proportion of genetic variance explained by each SNP was calculated as $\sigma^2_{SNP}/\sigma^2_g$.

For eQTL analysis, these models used VST-normalised read counts from the mapped RNA-seq data, representing the five nominally expressed genes in the QTL interval. These models also included a fixed effect for biopsy year to address batch variation between the different sequencing submissions. For milk composition and body weight analysis, models used the body weight and milk composition phenotypes described above. For the association analyses that considered body weight in the analysis, these models were conducted in the same way, with the addition of body weight as another covariate. Models also included a fixed effect to account for potentially confounding impacts of the *DGAT1* K232A mutation. This variant is known to have profound impacts on milk composition (Grisart, Coppieters, & Farnir, 2002), and despite being > 23 Mbp away, long distance linkage disequilibrium might be anticipated to influence milk composition association results.

Associations were considered significant using an alpha value of 0.05 that incorporated Bonferroni corrections for multiple hypothesis testing across the different study populations. For eQTL analyses, a total of 2160 tests were conducted (five gene expression traits x 432 SNPs), yielding a nominal significance threshold of $P=2.31\times10^{-5}$. For milk composition and body weight analyses, a nominal significance value of $P=1.51\times10^{-3}$ was used (33 tests).

Chapter 8: General Discussion

8.1. Overview

The overarching objective of the research presented in this thesis was to conduct detailed genetic analysis of four of the largest milk production loci in New Zealand (NZ) dairy cattle, focusing on identifying the causative gene and/or variant responsible for their respective impacts on milk composition and production. Identifying the causative variant(s) for the genetic signals at these loci requires understanding mechanism of action, and in most cases the identity of the causative gene. This has traditionally been a challenging task since association signals at implicated loci may encompass several megabase (Mbp) of DNA sequence, and in the case where causative variants reside in non-coding regions of the genome, the diversity of functions and scarcity of annotation resources makes attributing functional consequences difficult.

The work presented in chapters 4 – 7 of this thesis incorporated the use of a high-depth lactating mammary RNA sequence (RNAseq) dataset to look for milk production-associated variants that impacted the expression of mammary genes, helping to identify the genes (and variants) responsible. To this end, the four major milk production effects were attributed to four *cis*-transcriptionally regulated genes; *MGST1*, *DGAT1*, *AGPAT6* and *PLAG1*. The transcriptional level of these genes mirrored the milk production QTL at each of the four loci. The co-segregation of gene expression and milk production QTLs provides strong evidence of these QTL being driven by regulatory-based effects, and further work was conducted as described in these chapters to explore these mechanisms. This work employed the use of annotation and functional testing approaches to further refine the identity of the causative variants.

Additionally, methods to efficiently gene edit mammary cells were established towards directly testing the effects of the candidate causative variants on target gene expression in an isogenic background. Outlined in Chapter 3, these CRISPR-Cas9-based methods provide a platform for future analysis of both regulatory and protein-coding variants, enabling testing of the causality of candidates that might otherwise not be resolvable through statistical and bioinformatics means.

The main outcomes from this thesis are summarised and discussed below including implications and future directions.

8.2. Major research outcomes

8.2.1. Detailed investigation of the *MGST1* locus highlights a gene with major yet functionally unresolved effects on milk composition

In chapter 4, a QTL on bovine chromosome 5 with a large effect on milk fat percentage was investigated as part of a broader study led by Littlejohn et al., (2016). This locus was found to be responsible for 2.78% and 4.57% of the phenotypic and genotypic variance, respectively, in milk fat percentage in a large outbred NZ dairy cattle population (M. D. Littlejohn et al., 2016). This locus also had significant pleiotropic effects on milk composition, accounting for 0.10–2.78% and 0.23-4.57% of the phenotypic and genotypic variance, respectively, for the other lactation traits measured in this population (Littlejohn et al., 2016; summarised in Table 8.1). We attributed this QTL to *MGST1* due to the presence a strong *MGST1* expression QTL (eQTL) in the lactating mammary gland, which notably, bears the same genetic signal underpinning the milk fat percentage QTL. The top milk fat percentage QTL tag-SNP accounted for 60.73% and 88.8% of the phenotypic and genotypic variance, respectively, in mammary *MGST1* expression and importantly, the milk fat increasing allele was also associated with increased *MGST1* gene expression (summarised in Table 8.2). These data represent the first functional evidence for *MGST1* as the causative gene for this QTL. This contradicts previously published work that proposed these milk phenotypic effects were due to the gene *EPS8* (Wang et al., 2012).

The next step in characterising the *MGST1* locus was to examine candidate variants responsible for the QTLs, aiming to provide a molecular mechanism for these effects. Local reference sequence refinement was performed to identify any additional variants not captured in the genome assembly, and precise characterisation of an 8.2 kilobase (kb) deletion was conducted to assess the relative merits of these variants as functional polymorphisms for the QTLs. Despite the outstanding candidacy of the 8.2 kb deletion in particular, subsequent association analyses suggested that these polymorphisms were unlikely to be the source of the genetic signal at this locus (M. D. Littlejohn et al., 2016). Further association analyses were conducted to investigate the impact of *MGST1* variants on individual milk fatty acid profiles, and potential *trans*-eQTL effects as a consequence of differential *MGST1* expression. However, these approaches did not provide any additional insight into the cellular mechanisms or pathways underpinning the impact of *MGST1* on milk composition and production. Given that there are no obvious pre-existing mechanistic links between *MGST1* and lactation, these results do little to explain how the effect on milk composition is mediated. However, the results demonstrate the usefulness of a 'hypothesis-free' approach through genetic mapping. The GWAS and eQTL analysis combined nearly unequivocally demonstrate the involvement of *MGST1*, in the absence of an obvious cellular basis for these effects. Ultimately, further work is required to definitively prove the pathways and causal variant(s) responsible for the effects at this locus (discussed further in 8.3.2).

8.2.2. Characterisation of mammary *DGAT1* expression reveals a novel expression-based effect of K232A

Chapter 5 describes the results of a detailed investigation of transcriptional regulation at the *DGAT1* locus in the lactating mammary gland. This locus explains approximately 40% of the total genetic variance in milk fat content in the milk of Holstein cattle (Grisart, Coppieters, Farnir, et al., 2002), and is one of the most widely reported QTLs in analyses of bovine lactation traits. The effects have been near unanimously attributed to the *DGAT1* K232A polymorphism, though curiously, we identified a *cis*-eQTL for *DGAT1.* This eQTL appeared to be underpinned by the same genetic signal as the milk fat percentage QTL, where the K232A polymorphism was one of the most highly associated variants in our analysis. Notably, the milk fat increasing allele was the same allele associated with increased *DGAT1* expression, for the first time suggesting a potential expression-based mechanism to the QTL.

Further work revealed that the mammary *DGAT1* eQTL appears to be the result of an exon splice enhancer (ESE) that overlaps the K232A polymorphism. The dinucleotide substitution responsible for K232A results in the disruption of a consensus ESE motif which influences the splicing efficiency at numerous junctions of the gene. This effect was confirmed in cell-based experiments, whereby the amounts of spliced and unspliced RNA transcripts were quantified for four intron-exon *DGAT1* junctions. In agreement with the results of RNAseq analysis, the ratio of spliced to unspliced transcripts differed between the K and A alleles at a proportion of *DGAT1* splice junctions. Notably, the *DGAT1* expression-increasing K allele demonstrated a greater splicing efficiency at the affected intron junctions, suggesting increased yield of mRNA as a consequence of increased pre-mRNA to mRNA conversion. Taken together, these observations provide functional evidence of an expression-based mechanism for the *DGAT1* K232A polymorphism, which is likely to be, at least in part, responsible for lactation effects at the locus.

8.2.3. Genetic and functional characterisation of the *AGPAT6* locus and investigation of a candidate causative 5'UTR VNTR variant

Chapter 6 describes the investigation of a QTL on bovine chromosome 27 with a large effect on milk fat percentage. This locus was responsible for 1.26% and 2.22% of the phenotypic and genetic variance, respectively, in milk fat percentage in a large outbred NZ dairy cattle population. Similarly, this locus also had significant pleiotropic effects on milk composition, accounting for 0.08-1.26% and 0.002-2.22% of the phenotypic and genotypic variance, respectively, of the other milk composition production traits (summarised Table 8.1). As part of a broader study led by Littlejohn et al., (2014) we provided the first functional evidence supporting *AGPAT6* underpinning this QTL, that, like the *MGST1* locus, co-segregated with a strong expression QTL (eQTL) in the lactating mammary gland.

Based on observation of this expression-based effect, a variable number tandem repeat (VNTR) in the 5' untranslated region (UTR) of *AGPAT6* represented an appealing candidate causative variant. The development and testing of a custom genotyping assay to directly interrogate this variant in the FJXB pedigree is described. Subsequent association analysis incorporating these genotypes revealed large pleiotropic effects for a variety of milk composition and production phenotypes, and notably, the *AGPAT6* VNTR was the most highly associated variant in this analysis (presented in Littlejohn et al., 2014).

To investigate the functionality of the *AGPAT6* VNTR directly, CRISPR-Cas9 mediated genome editing was carried out to attempt to generate a series of bovine mammary cell lines containing alternate genotypes of this, and other, candidate variants. The screening of hundreds of clonal cell lines unfortunately revealed that targeting of these alleles had been unsuccessful. This negative result remains to be understood. However prior to this experiment, the method optimisation components of this work, which was very successful, will serve as exemplification of how the methods could be used in future experiments. This approach should ultimately provide an answer as to what non-coding variants underpin modulation of milk composition-implicated genes such as *AGPAT6* and *MGST1*.

8.2.4. Expression-based effects underpin the impact of *PLAG1* on bodyweight and milk composition

In Chapter 7, the major growth and body weight QTL located on chromosome 14 was investigated. This QTL has been attributed to two functional variants in the bidirectional promoter of *PLAG1* and *CHCHD7*, although the precise involvement of these and several other genes at the broader locus remains unresolved (Karim et al., 2011). A tag-SNP representing this locus was responsible for 15.80% and 32.78% of the phenotypic and genetic variance, respectively, in body weight in NZ dairy cattle (summarised in Table 8.1). We identified a *cis*-eQTL for *PLAG1* in the lactating mammary gland which, notably, showed similar variant association statistics to the body weight QTL at this locus. The top body weight QTL tag-SNP was the most highly associated variant in this analysis, and accounted for 32.59% and 46.59% of the phenotypic and genotypic variance, respectively, in mammary *PLAG1* expression (Table 8.2). These data represent the first functional conformation of the expression-based effect at this locus, and implicate *PLAG1* alone as responsible for these effects, contrary to the previous study that demonstrated *cis*-eQTLs for multiple genes in foetal tissue (Karim et al., 2011).

Given the observation of the mammary eQTL effects, we wondered whether differential expression of *PLAG1* might impact lactation traits directly. Chapter 7 also reports new associations of *PLAG1* genotype with pleiotropic effects on milk composition, whereby this locus accounts for 0.08–0.27% and 0.31-1.24% of the phenotypic and genotypic variance, respectively, in milk composition and production traits (Table 8.1). Based on the correlation between body size and milk production, we adjusted for body weight in our models revealing that, while this association is predominately driven by differences in animal size, there appears to be an effect of this locus on milk composition independent of its influence on body weight (Table 8.1).Taken together, this work is the first to provide functional evidence supporting the causal status of *PLAG1* underlying these effects, and adds lactation effects to the list of physiological traits impacted by this locus.

Table 8.1 Summary of the phenotypic and genotypic variance in bovine lactation traits explained by the four genes investigated in this thesis

Gene	**MGST1^**		**AGPAT6**		**DGAT1***		**PLAG1**		***PLAG1* cond. on bodyweight**	
	(N=~64,100)		(N=~37,200)		(N=~39,400)		(N=~39,400)		(N=~39,400)	
	Pheno var	**Geno var**	**Pheno var**	**Geno var**	**Pheno var**	**Geno var**	**Pheno var**	**Geno var**	**Pheno var**	**Geno var**
Fat %	**2.78**	**4.57**	**1.26**	**2.22**	**27.95**	**43.56**	**0.24**	**0.88**	**0.115**	**0.23**
Fat yield	**0.22**	**0.89**	**0.19**	**0.003**	**2.89**	**11.82**	**0.16**	**0.31**	**0.65**	**3.087**
Protein %	**1.10**	**1.93**	**0.19**	**0.002**	**8.33**	**14.17**	**0.08**	**0.35**	**0.091**	**0.165**
Protein yield	**0.26**	**1.03**	0.08	0.14	**2.24**	**9.63**	N/A	N/A	**0.516**	**2.614**
Lactose %	**0.10**	**0.23**	**0.89**	**1.91**	ND	ND	ND	ND	ND	ND
Lactose yield	**0.64**	**2.25**	**0.18**	**0.69**	ND	ND	ND	ND	ND	ND
Milk volume	**0.83**	**2.70**	**0.08**	**0.27**	**6.59**	**20.57**	**0.27**	**1.24**	**0.322**	**1.226**
Body weight	ND	ND	ND	ND	**ND**	**ND**	**15.80**	**32.78**	**N/A**	**N/A**

ND – not determined, N/A – not applicable

*conducted as part of the *PLAG1* analysis and not presented in Chapter 5

^conducted in Littlejohn et al., (2016)

Table 8.2 Summary of the phenotypic and genotypic variance in cis-gene expression for four genes with large effect on bovine milk composition

Gene	***MGST1***		***DGAT1***		**AGPAT6***		***PLAG1***	
	Pheno var	**Geno var**	**Pheno var**	**Geno var**	**Pheno var**	**Geno var**	**Pheno var**	**Geno var**
***Cis*-gene expression**	**60.73**	**88.8**	**31.23**	**99.99**	**18.02**	**60.14**	**32.59**	**46.59**

*conducted in Littlejohn et al., (2014)

8.2.5. Major impact variants underpin bovine milk lactation traits in NZ dairy animals

In this thesis, I have focused on analysis of some of the largest QTL in the NZ dairy population (Table 8.1). These QTL also represent major-effect loci in other populations (Daetwyler et al., 2014; Fontanesi et al., 2014; Grisart, Coppieters, & Farnir, 2002; Iso-Touru, Sahana, Guldbrandtsen, Lund, & Vilkki, 2016; Karim et al., 2011; Komisarek et al., 2011; M. Littlejohn et al., 2012; Raven et al., 2015; X. Wang et al., 2012), and it is interesting to consider the genetic history of these QTL in NZ and other bovine populations around the world. A prevailing theory as to the very large effect sizes of the described QTLs relates to the strong selection pressure placed on bovine milk composition and production. Since the advent of agriculture, dairy cattle have undergone significant phenotypic changes and genetic adaptation to various farming conditions, and alleles with moderate to large effects have been enriched due to this strong directional selection (Andersson & Georges, 2004). Indeed, *PLAG1* accounts for almost 16% of the phenotypic variance in body weight, and the opposing alleles responsible for this effect have reached near fixation in Holstein-Friesians (characteristically large animals) and Jerseys (characteristically small animals). As a comparison, the effect sizes of many of the loci influencing human height typically account for less than half a percent of the phenotypic variance, displaying a much more infinitesimal architecture (Gudbjartsson et al., 2008; Lango Allen et al., 2010; Visscher et al., 2010).

The associations between these four loci and milk composition and production phenotypes provides evidence of the pleiotropic effects of major effect genes, whereby they co-ordinately impact multiple phenotypes (Table 8.1). This phenomenon of milk composition and production effects being concomitantly influenced at major QTL has been demonstrated previously (Andersson & Georges 2004; Littlejohn et al., 2014; Kemper, Hayes, Daetwytler & Goddard, 2015). While further work is required to demonstrate the effects of QTL genotype on mammary physiology and function, this phenomenon is likely the result of the multi-faceted regulation of mammary and lactation biology. Indeed, this work enables us to speculate about the mammary-specific pathways impacted by these variants and generate hypothesises for further functional characterisation of mammary physiology. This is of particular note for *MGST1*, whereby the results of this work have, for the first time, linked this gene to a role in lactation based on its major impact on multiple milk phenotypes. Similarly, the demonstration of the reduction in milk fat yield (when accounting for body weight) in animals carrying the *PLAG1* body weight increasing allele enables us to speculate regarding the differential energy utilisation between animals of different *PLAG1* genotypes. It is likely that other phenotypes not measured or analysed through the course of this study are also impacted by the investigated variants, and it is also possible that not all physiological impacts would be desirable. In the same way that aggressive artificial selection has increased the frequencies of major effect alleles, negative, pleiotropic consequences that might otherwise have been subject to purifying selection in wild populations, may have been overridden. Indeed, multiple examples of balancing selection in domestic species have recently emerged (Charlier et al., 2016; Rupp et al., 2015; Tamma et al., 2012), so future work could examine further-reaching consequences of the *MGST1*, *DGAT1*, *AGPAT6* and *PLAG1* variants.

8.2.6. Regulatory variants impact milk lactation traits in NZ dairy cattle

Transcriptional profiling of the bovine lactating mammary gland identified very significant *cis*-eQTLs for *MGST1*, *AGPAT6*, *DGAT1* and *PLAG1* (summarised in Table 8.2). Importantly, these eQTL appear to co-segregate, and therefore likely underlie the milk composition and production QTLs at each locus. The demonstration of shared genetic signal between co-locating QTLs is a powerful method to identify the causative gene for these traits, and in some respects, makes this class of QTLs more tractable than those underpinned by protein-function based mechanisms, since large transcriptome-wide datasets can simultaneously investigate multiple loci. The limitation, however, is in identification of causative *variants* for such eQTLs, since unlike analyses of protein coding variants, the consequences of non-coding variants are much more difficult to predict. This is of note for the *AGPAT6* and *MGST1* loci, as despite identifying *cis*-eQTLs for these genes, we were unable to delineate the causal variant from the myriad of other associated variants. An unusual example in this context, however, exists in our discovery of an expression-based effect for the *DGAT1* K232A mutation. Remarkably, by virtue of a functional study conducted in 2004 that attributed enzymatic differences to the two protein isoforms (Grisart et al., 2004), the K232A variant has long been assumed to mediate its action solely through these protein-based effects. In this thesis, I have demonstrated that the MNP encoding this amino acid substitution also promotes splicing and mRNA expression, presenting an unusual case whereby a single variant likely modulates phenotype through multiple, though genetically inseparable, effects.

To create a system to identify and/or more fully characterise causative variants, I investigated the use of CRISPR-Cas9 genome editing in mammary cells. Despite not being able to provide functional evidence supporting one or more of the *AGPAT6* variants, I demonstrated highly efficient CRISPR-Cas9 genome editing and the high-throughput screening of CRISPR-Cas9 edited cell lines. Similar protocols have been successful in identifying functional regulatory elements in the human genome (Klann et al., 2017), and recapitulating the effects of oncogenic mutations *in vitro* and *in vivo* (reviewed in Guernet & Grumolato, 2017). Ultimately, further work using the tools established here will allow for the rapid modelling of the impacts of candidate causal variants identified through GWAS for bovine milk composition and production.

8.3. Implications and future directions

8.3.1. Regulatory variants can have large and complex effects on quantitative traits

The identification of expression-based effects at four of the largest effect milk production loci demonstrates that regulatory variants can have major effects on quantitative traits. These data suggest we need to move away from the assumption that large genetic effects can only be the result of high-penetrance coding variants. While on a molecular level, it can be easily rationalised that changes in the amount of a gene product can have similar effects to changing the structure of that product, there are limited examples of large expression-based effects influencing both quantitative and Mendelian traits (Makrythanasis & Antonarakis, 2013). This is likely a reflection of the experimental approaches taken, which predominately interrogate exome DNA sequence to find genetic variants that associate with the trait. As such, we speculate that this ascertainment bias is likely to diminish as more studies leverage whole genome sequencing technologies to establish the full catalogues of genetic variation and their relationship to phenotype.

Transcriptional regulation is controlled at many levels, including chromatin formation, histone modification, transcription initiation, RNA polyadenation, pre-mRNA splicing, mRNA stability, and translation initiation (de Vooght, van Wijk, & van Solinge, 2009). As such, the variants implicated in this thesis are likely to influence one (or more) of these mechanisms involved in transcriptional regulation to contribute to changes in *cis*-gene expression. However, given the lack of annotation of the bovine genome it is hard to ascertain the regulatory mechanisms implicated by each of these sequence variants without directly interrogating them experimentally. While I was able to functionally characterise the influence of *DGAT1* K232A on *DGAT1* pre-mRNA splicing, we can only speculate on the possible molecular mechanisms underpinning the gene expression effects of the remaining variants at the *MGST1*, *PLAG1* and *AGPAT6* loci. To this end, it is plausible the variants implicated at the *MGST1* locus, which reside within the 4 kb upstream of the TSS of *MGST1,* co-locate to an upstream promoter element, and that one or more of these variants alter the binding capacity of the *cis*-acting DNA sequence motifs for protein factors that usually interact with them. Similarly, the candidate causal variant at the *PLAG1* locus sits in the core promoter of the gene and it is predicted that this polymorphism influences the activation of the promoter, and/or alters the interaction between transcription factors and chromatin modifying enzymes with the promoter to drive the expression of *PLAG1*. In contrast, polymorphisms within 5′ UTR often alter the post-transcriptional regulation of gene expression through the modulation of the transport of mRNAs out of the nucleus, translation efficiency, subcellular localisation and transcript stability (Migone et al., 2002). As such, the VNTR in the 5′ UTR of *AGPAT6* represents a strong candidate causal variant as it potentially plays a role in the initiation of translation or the secondary structure of the mRNA transcript to alter *AGPAT6* expression. It is also possible that this variant also changes the affinity of the DNA sequence to bind transcription factors to alter the expression of the gene. Indeed, while these variants all influence *cis*-gene expression in the bovine mammary gland, the mechanism by which they alter the transcriptional regulation of these genes is specific to each polymorphism, highlighting the complexity of functional relationship between genotype and phenotype for regulatory variants.

8.3.2. Understanding and characterising causative variants requires new data and methods

Various approaches have been developed to identify genetic variants that influence quantitative traits; however most of these focus on predicting the effect of variants in coding regions. Most of the implicated genetic variants in GWAS reside in non-coding sequence, presenting a major challenge since the functional prediction of variant effects is difficult due to lack of annotation resources. Further, statistical delineation of association signals may be impractical or impossible, particularly in cattle populations where stretches of linkage disequilibrium extend over long physical distances (Farnir et al., 2000). To this end, we need to develop specific and scalable tools to identify causative variants for regulatory-based effects. There are two plausible avenues to achieve this: indirectly by generating genome-wide functional annotations in the lactating mammary gland; or through the direct functional testing of candidate causative variants.

8.3.2.1. Generating functional annotations to filter regulatory variants

Non-coding variants underpinning GWAS signals co-locate to eQTL more often than expected by chance (Maurano et al., 2012). However, identifying the exact polymorphism responsible for the signal, and the molecular pathways impacted by these variants, is difficult as there are a variety of processes involved in the regulation of gene expression (Encode Consortium, 2012). There are a number of elements essential for gene expression that are potentially perturbed by genetic variants, including transcription factors binding at promoters and enhancers (Sheffield et al., 2013; Valouev et al., 2008), and chromatin interactions and structure (Song et al., 2011; Thurman et al., 2012).

Advancements in genomic technologies have opened new avenues for the detailed examination of how regulatory variants can influence the individual steps of gene expression (Ulirsch et al., 2016). Numerous large-scale and multi-disciplinary projects, like the Encyclopaedia of DNA elements (ENCODE) (Encode Consortium, 2012), and the Roadmap Epigenomics Project (Kundaje et al., 2015), have been established to understand and catalogue the functional regulatory elements in the human genome. Pioneered by the ENCODE project, several powerful approaches now exist to annotate and map cell- and tissue-specific regulatory elements and features. These include identification of open chromatin using DNaseI-hypersensitivity, histone modifications and transcription factor binding sites using chromatin immunoprecipitation, DNA interactions using chromatin conformation capture, and differential DNA methylation using bisulphate sequencing. Subsequently, studies have demonstrated a substantial enrichment of GWAS variants in ENCODE-defined regions containing regulatory elements (Handel, Gallone, Cader, & Ponting, 2017), with one study demonstrating that 67.7% of brain disease associated eQTLs co-located to within DNase-I hypersensitivity sites (Handel et al., 2017).

The ENCODE methods can be used to filter candidate causal variants based on whether they co-locate to functional regulatory elements, and this also generates hypotheses regarding the mechanism of the regulatory variant responsible. Consequently, this reduces the number of candidate variants for manual interrogation, and enables the implementation of the most appropriate functional assay for the direct testing of these polymorphisms. Based on the success of eQTL mapping to identify causative variants underlying milk composition and production QTL, extending association mapping to additional molecular phenotypes such as methylation, transcription factor binding, and chromatin conformation traits will be highly useful.

8.3.2.2. High-throughput functional screens of candidate causal variants

The predominant approach to provide direct functional evidence that a genetic variation has an affect is to assay it in an *in vitro* system or model organism. This approach does not scale well to the vast numbers of non-coding regulatory variants derived from GWAS. However, the recent advent of massively parallel reporter assays (MPRAs) provides a powerful tool for comprehensively assessing the effects of all possible regulatory variants within associated loci (Tewhey et al., 2016; Vockley et al., 2015). In these assays, a construct variant library or allelic series is synthesised encompassing all the variants of interest and integrated into reported gene plasmids, which are subsequently delivered into *in vitro* or *in vivo* system. Functional assays are conducted and variants are stratified based on impact on gene expression e.g. if they increase or decrease transcription relative to the wild-type sequence. A functional score for each genetic variant is derived using high-throughput sequencing and the barcodes placed in the reporter gene. Using this approach Tewhey et al. (2016) quantified the effects of common genetic variants within 3,642 eQTLs and, notably, were able to validate the effects of a subset of the expression-modulating variants using CRISPR-Cas 9 genome editing.

While the use of MPRAs to investigate the effects of trait associated regulatory variants provides unprecedented opportunity to screen large numbers of polymorphisms, there are some drawbacks associated with this highly *in vitro* approach given its use of exogenous DNA expression plasmids. The differentiation of the effects of sequence variants using outside of their genomic context where additional genomic complexity (e.g. modifier elements) is lacking limits the physiological relevance as they do not recapitulate the regulatory element interactions at the native locus (Klann et al., 2017; J. B. Wright & Sanjana, 2016). This approach also assumes that the relevant transcription factors are expressed in the cell line and are able to interact with the short oligonucleotides within the plasmid. Further complicating the use of MPRAs to model regulatory variants, which typically only have modest effects, is that large numbers of individual lines are needed to provide sufficient statistical power to detect significant impacts and any implicated variants will need to be further validated as this assay provides no information on the particular mechanism underlying the effects of these polymorphisms. In contrast, targeted CRISPR-Cas9 mediated genome editing can be used to introduce individual alleles outside of their native haplotypic context, yet within the complex genomic environment (T. Wang et al., 2014). Generating an isogeneic series of cell lines representing the different genotypes of candidate causative variants therefore provides another, and arguably more physiologically relevant, method for identifying regulatory variants that modulate gene expression or cell function. CRISPR-based methods of variant testing are lower throughput than MPRA, with the relatively low efficiency of the HDR-mediated editing pathway one drawback in particular. However, the generation of CRISPR gRNA libraries has been successfully used for high-throughput loss-of-function screens of genes (Shalem, Sanjana, Hartenian, & Zhang, 2014; T. Wang et al., 2014) and regulatory elements (Canver, Bauer, & Orkin, 2017; Diao et al., 2017; Sanjana, 2016), and will likely represent a powerful and lucrative approach to testing variant function as the methods continue to improve.

Rapid advancements in high-throughput screens of genetic variants have opened new avenues for both hypothesis driven and unbiased interrogation of genetic variants associated with gene expression (and potentially other molecular phenotypes). These methods overcome some of the difficulties of identifying regulatory causal variants and have massive potential to increase our understanding of the biological roles of these polymorphisms in milk production traits. As our knowledge of the complexities of gene expression increase, the more diverse mechanisms of action of regulatory variants will be recognised (Albert & Kruglyak, 2015).

8.3.3. Application of these genetic variants in animal selection and generation

The work presented in this thesis investigated molecular and genetic aspects of major-effect loci influencing bovine milk composition. This information could be used to attempt to increase the accuracy of genomic predictions for genomic selection (VanRaden, Tooker, O'Connell, Cole, & Bickhart, 2017). Currently genomic selection does not use these genetic variants directly, but researchers at Livestock Improvement Corporation are experimenting with methods to genotype directly or impute the variants highlighted in this thesis and examine the impact of these on genomic prediction results.

Another potential application for the variants described in this thesis is their targeted introduction in one-cell embryos via CRISPR-Cas9 mediated genome editing. This could, in theory, enable the generation of dairy cows with dramatically different milk composition profiles. Assuming the effects of the variants combine in a purely additive fashion, causative alleles in *MGST1, DGAT1, AGPAT6,* and *PLAG1* together account for 32.23% and 51.23% of the phenotypic and genotypic variance, respectively, in milk fat percentage in NZ dairy cattle (Table 8.1). Added together, stacking of high and low milk fat percentage alleles could be expected to create milks differing by up to 1.2% milk fat, a similar contrast to commercial products marketed as 'standard' (3% milk fat), and 'lite' (1.5%) liquid milks. The generation of animals with such divergent milk composition profiles has many possibilities, though one caveat here is that the causative mutations would need to be known to enable stacking of alleles. This creates an economic incentive to identify the causative mutations for QTLs, and should warrant expanded investment in the functional annotation and *in vitro* testing methods discussed above, as well as contributing academic value through identification of causal genes important to regulation of milk composition and lactation.

8.4. Concluding remarks

The research described in this thesis has revealed the causative gene and in some cases the variants responsible for major impacts on bovine lactation traits. The characterisation of gene expression-based effects at each of these loci provides an enhanced understanding of the chain of causality from genetic variants, to genes, and ultimately phenotype.

This work also highlights the complexity of interpreting regulatory effects, and we envisage that further studies to identify causative regulatory variants will need to incorporate new genomic and genetic technologies, such as the CRISPR-Cas9 genome editing described in this research. Ultimately, improved methods to annotate, filter, or directly test the causality of regulatory variants will help overcome the substantial 'burden of proof' required to assign causality to this class of genetic variants, and should also help to understand and better utilise the wealth of selectable variation that exists in the bovine genome.

List of References

Adelson, D. L., Raison, J. M., & Edgar, R. C. (2009). Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12855–12860. http://doi.org/DOI 10.1073/pnas.0901282106

Aida, T., Chiyo, K., Usami, T., Ishikubo, H., Imahashi, R., Wada, Y., … Tanaka, K. (2015). Cloning-free CRISPR/Cas system facilitates functional cassette knock-in in mice. *Genome Biology*, *16*(1), 87. http://doi.org/10.1186/s13059-015-0653-x

Akers, M. R. (2002). *Lactation and the Mammary Gland* (1st Editio). Blackwell Publishing Company.

Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212. http://doi.org/10.1038/nrg3891

Alim, M. A., Wang, P., Wu, X. P., Li, C., Cui, X. G., Zhang, S. L., … Sun, D. X. (2014). Effect of FASN gene on milk yield and milk composition in the Chinese Holstein dairy population. *Animal Genetics*, *45*(1), 111–113. http://doi.org/10.1111/age.12089

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. http://doi.org/10.1186/gb-2010-11-10-r106

Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., & Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*, *8*(9), 1765–86. http://doi.org/10.1038/nprot.2013.099

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq A Python framework to work with high-throughput sequencing data. *bioRxiv*, *31*(2), 166–169. http://doi.org/10.1101/002824

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, *22*(10), 2008–17. http://doi.org/10.1101/gr.133744.111

Anderson, S. M., Rudolph, M. C., McManaman, J. L., & Neville, M. C. (2007). Key stages in mammary gland development. Secretory activation in the mammary gland: it's not just about milk protein synthesis! *Breast Cancer Research : BCR*, *9*(1), 204. http://doi.org/10.1186/bcr1653

Andersson, L., & Georges, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. *Nature Reviews. Genetics*, *5*(3), 202–12. http://doi.org/10.1038/nrg1294

Arvey, A., Agius, P., Noble, W. S., & Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Research*, *22*(9), 1723–34. http://doi.org/10.1101/gr.127712.111

Avery, O. T., MacLeod, C. M., & McCarthy, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental*

Medicine, *2*(6).

Banos, G., Woolliams, J. A., Woodward, B. W., Forbes, A. B., & Coffey, M. P. (2008). Impact of Single Nucleotide Polymorphisms in Leptin, Leptin Receptor, Growth Hormone Receptor, and Diacylglycerol Acyltransferase (DGAT1) Gene Loci on Milk Production, Feed, and Body Energy Traits of UK Dairy Cows. *Journal of Dairy Science*, *91*(8), 3190–3200. http://doi.org/10.3168/jds.2007-0930

Bauman, D. E., & Currie, W. B. (1980). Partitioning of nutrients during pregnancy and lactation: a review of mechanisms involving homeostasis and homeorhesis. *Journal of Dairy Science*, *63*(9), 1514–1529. http://doi.org/10.3168/jds.S0022-0302(80)83111-0

Beigneux, A. P., Vergnes, L., Qiao, X., Quatela, S., Davis, R., Watkins, S. M., … Young, S. G. (2006). Agpat6--a novel lipid biosynthetic gene required for triacylglycerol production in mammary epithelium. *Journal of Lipid Research*, *47*(4), 734–44. http://doi.org/10.1194/jlr.M500556-JLR200

Bennewitz, J., Reinsch, N., Paul, S., Looft, C., Kaupe, B., Weimann, C., … Kalm, E. (2004). The DGAT1 K232A Mutation Is Not Solely Responsible for the Milk Production Quantitative Trait Locus on the Bovine Chromosome 14. *Journal of Dairy Science*, *87*(2), 431–442. http://doi.org/10.3168/jds.S0022-0302(04)73182-3

Berry, S., Coppieters, W., Davis, S., Burrett, A., Thomas, N., Palmer, D., … Snell, R. (2013). A triad of highly divergent polymeric immunoglobulin receptor (PIGR) haplotypes with major effect on IgA concentration in bovine milk. *PloS One*, *8*(3), e57219. http://doi.org/10.1371/journal.pone.0057219

Berry, S. D., Davis, S. R., Beattie, E. M., Thomas, N. L., Burrett, a. K., Ward, H. E., … Snell, R. G. (2009). Mutation in bovine β-carotene oxygenase 2 affects milk color. *Genetics*, *182*(3), 923–926. http://doi.org/10.1534/genetics.109.101741

Bétermier, M., Bertrand, P., & Lopez, B. S. (2014). Is Non-Homologous End-Joining Really an Inherently Error-Prone Process ? *PLoS Genetics*, *10*(1), e1004086. http://doi.org/10.1371/journal.pgen.1004086

Beuzen, N. D., Stear, M. J., & Chang, K. C. (2000). Molecular markers and their use in animal breeding. *The Veterinary Journal*, *160*(1), 42–52. http://doi.org/10.1053/tvjl.2000.0468

Bhaya, D., Davison, M., & Barrangou, R. (2011). CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation. *Annual Review of Genetics*, *45*(1), 273–297. http://doi.org/10.1146/annurev-genet-110410-132430

Bionaz, M., & Loor, J. J. (2008). Gene networks driving bovine milk fat synthesis during the lactation cycle. *BMC Genomics*, *9*(366). http://doi.org/10.1186/1471-2164-9-366

Blott, S., Kim, J., Moisio, S., Schmidt-ku, A., Cornet, A., Berzi, P., … Georges, M. (2003). Molecular Dissection of a Quantitative Trait Locus : A Phenylalanine-to-Tyrosine Substitution in the Transmembrane Domain of the Bovine Growth Hormone Receptor Is Associated With a Major Effect on Milk Yield and Composition. *Genetics*, *163*, 253–266.

Bolormaa, S., Hayes, B. J., Savin, K., Hawken, R., Barendse, W., Arthur, P. F., … Goddard, M. E. (2011). Genome-wide association studies for feedlot and growth traits in cattle. *Journal of Animal Science*, *89*(6), 1684–1697. http://doi.org/10.2527/jas.2010-3079

Boras-Granic, K., & Wysolmerski, J. J. (2008). Wnt signaling in breast organogenesis. *Organogenesis*, *4*(2), 116–122. http://doi.org/10.4161/org.4.2.5858

Böttcher, R., Hollmann, M., Merk, K., Nitschko, V., Obermaier, C., Philippou-Massier, J., … Förstemann, K. (2014). Efficient chromosomal gene modification with CRISPR/cas9 and PCR-based homologous recombination donors in cultured Drosophila cells. *Nucleic Acids Research*, *42*(11). http://doi.org/10.1093/nar/gku289

Boutz, P. L., Bhutkar, A., & Sharp, P. A. (2015). Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes and Development*, *29*(1), 63–80. http://doi.org/10.1101/gad.247361.114

Bouwman, A. C., Visker, M. H., van Arendonk, J. A., & Bovenhuis, H. (2012). Genomic regions associated with bovine milk fatty acids in both summer and winter milk samples. *BMC Genetics*, *13*(1), 93. http://doi.org/10.1186/1471-2156-13-93

Bovenhuis, H., Van Arendonk, J. a, & Korver, S. (1992). Associations between milk protein polymorphisms and milk production traits. *Journal of Dairy Science*, *75*(9), 2549–2559. http://doi.org/10.3168/jds.S0022-0302(92)78017-5

Bovenhuis, H., Visker, M. H. P. W., & Lundén, a. (2013). Selection for milk fat and milk protein composition. *Advances in Animal Biosciences*, *4*(3), 612–617. http://doi.org/10.1017/S2040470013000174

Boyko, A. R., Quignon, P., Li, L., Schoenebeck, J. J., Degenhardt, J. D., Lohmueller, K. E., … Ostrander, E. A. (2010). A Simple Genetic Architecture Underlies Morphological Variation in Dogs. *PLoS Biology*, *8*(8), e1000451. http://doi.org/10.1371/journal.pbio.1000451

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., … Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nature Biotechnology*, *26*(10), 1146–1153. http://doi.org/10.1038/nbt.1495

Brennan, A. J., Sharp, J. A., Lefevre, C., Topcic, D., Auguste, A., Digby, M., & Nicholas, K. R. (2007). The Tammar Wallaby and Fur Seal: Models to Examine Local Control of Lactation. *Journal of Dairy Science*, *90*, E66–E75. http://doi.org/10.3168/jds.2006-483

Brisken, C., Ayyannan, A., Nguyen, C., Heineman, A., Reinhardt, F., Jan, T., … Weinberg, R. A. (2002). IGF-2 is a mediator of prolactin-induced morphogenesis in the breast. *Developmental Cell*, *3*(6), 877–887. http://doi.org/10.1016/S1534-5807(02)00365-9

Brisken, C., Kaur, S., Chavarria, T. E., Binart, N., Sutherland, R. L., Weinberg, R. a, … Ormandy, C. J. (1999). Prolactin controls mammary gland development via direct and indirect mechanisms. *Developmental Biology*, *210*(1), 96–106. http://doi.org/10.1006/dbio.1999.9271

Brotherstone, S. (1994). Genetic and phenotypic correlations between linear type traits and production traits in Holstein-Friesian dairy cattle. *Animal Science*, *59*, 183–187. http://doi.org/doi:10.1017/S0003356100007662

Browning, B. L., & Browning, S. R. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics*, *84*(2), 210–223. http://doi.org/10.1016/j.ajhg.2009.01.005

Brunham, L. R., & Hayden, M. R. (2013). Hunting human disease genes: lessons from the past, challenges for the future. *Human Genetics*, *132*(6), 603–17. http://doi.org/10.1007/s00439-013-1286-3

Brym, P., Kamiński, S., & Wójcik, E. (2005). Nucleotide sequence polymorphism within exon 4 of the bovine prolactin gene and its associations with milk performance traits. *Journal of Applied Genetics*, *46*(2), 179–185. Retrieved from http://europepmc.org/abstract/MED/15876685

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–8. http://doi.org/10.1038/nmeth.2688

Buitenhuis, B., Janss, L. L., Poulsen, N. a, Larsen, L. B., Larsen, M. K., & Sørensen, P. (2014). Genome-wide association and biological pathway analysis for milk-fat composition in Danish Holstein and Danish Jersey cattle. *BMC Genomics*, *15*(1). http://doi.org/10.1186/1471-2164-15-1112

Canver, M. C., Bauer, D. E., & Orkin, S. H. (2017). Functional interrogation of non-coding DNA through CRISPR genome editing. *Methods*. http://doi.org/10.1016/j.ymeth.2017.03.008

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. http://doi.org/10.1186/s13742-015-0047-8

Charlier, C., Li, W., Harland, C., Littlejohn, M., Coppieters, W., Creagh, F., … Georges, M. (2016). NGS-based reverse genetic screen for common embryonic lethal mutations compromising fertility in livestock. *Genome Research*, *26*(10), 1333–1341. http://doi.org/10.1101/gr.207076.116

Chu, V. T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., & Kühn, R. (2015). letters Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells, *33*(5). http://doi.org/10.1038/nbt.3198

Cobanoglu, O., Zaitoun, I., Chang, Y. M., Shook, G. E., & Khatib, H. (2006). Effects of the Signal Transducer and Activator of Transcription 1 (STAT1) Gene on Milk Production Traits in Holstein Dairy Cattle. *Journal of Dairy Science*, *89*(11), 4433–4437. http://doi.org/http://doi.org/10.3168/jds.S0022-0302(06)72491-2

Cochran, S. D., Cole, J. B., Null, D. J., & Hansen, P. J. (2013). Single nucleotide polymorphisms in candidate genes associated with fertilizing ability of sperm and subsequent embryonic development in cattle. *Biology of Reproduction*, *89*(3), 69. http://doi.org/10.1095/biolreprod.113.111260

Cohen-Zinder, M., Seroussi, E., Larkin, D. M., Loor, J. J., Everts-Van Der Wind, A., Lee, J. H., … Ron, M. (2005). Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. *Genome Research*, *15*(7), 936–944. http://doi.org/10.1101/gr.3806705

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., … Jacobsen, S. E. (2008). Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature*, *452*(7184), 215–219. Retrieved from http://dx.doi.org/10.1038/nature06745

Cole, J. B., Wiggans, G. R., Ma, L., Sonstegard, T. S., Lawlor, T. J., Crooker, B. A., … Da, Y. (2011). Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. *BMC Genomics*, *12*(1), 408. http://doi.org/10.1186/1471-2164-12-408

Cyranoski, D. (2016). CRISPR gene editing tested in a person. *Nature*, *539*(7630), 479. http://doi.org/10.1038/nature.2016.20988

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., Binsbergen, R. Van, Brøndum, R. F., … Hayes, B. J. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature Genetics*, *46*, 858–865. http://doi.org/10.1038/ng.3034

DairyNZ. (2014a). Evaluation system for traits other than production (TOP) for dairy cattle in New Zealand. Retrieved from http://www.dairynz.co.nz/animal/animal-evaluation/interpreting-the-info/breeding-values/

DairyNZ. (2014b). New Zealand Dairy Statistics 2013-2014. Retrieved from https://dairy.ahdb.org.uk/resources-library/market-information/dairy-statistics/dairy-statistics-an-insiders-guide-2016/#.WJ2cFBicaCR

de Vooght, K. M. K., van Wijk, R., & van Solinge, W. W. (2009). Management of Gene Promoter Mutations in Molecular Diagnostics. *Clinical Chemistry*, *55*(4), 698 LP-708. Retrieved from http://clinchem.aaccjnls.org/content/55/4/698.abstract

De Vos, L., Declercq, J., Rosas, G. G., Van Damme, B., Roebroek, A., Vermorken, F., … Creemers, J. (2008). MMTV-cre-mediated fur inactivation concomitant with PLAG1 proto-oncogene activation delays salivary gland tumorigenesis in mice. *International Journal of Oncology*, *32*(5), 1073–83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18425334

Deb, R., Singh, U., Kumar, S., Singh, R., Sengar, G., & Sharma, A. (2014). Genetic polymorphism and association of kappa-casein gene with milk production traits among Frieswal (HF x Sahiwal) cross breed of Indian origin. *Iranian Journal of Veterinary Research*, *15*(4), 406–408.

DeChiara, T. M., Efstratiadis, A., & Robertson, E. J. (1990). A growth-deficiency phenotype in heterozygous mice carrying an insulin-like growth factor II gene disrupted by targeting. *Nature*, *3*(345), 78–80.

Declercq, J., Skaland, I., Van Dyck, F., Janssen, E. A. M., Baak, J. P., Drijkoningen, M., & Van De Ven, W. J. M. (2008). Adenomyoepitheliomatous lesions of the mammary glands in transgenic mice with targeted PLAG1 overexpression. *International Journal of Cancer*, *123*(7), 1593–1600. http://doi.org/10.1002/ijc.23586

Dekkers, J. C. M., & Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews. Genetics*, *3*(1), 22–32. http://doi.org/10.1038/nrg701

Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., … Ren, B. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods*, (March), 1–11. http://doi.org/10.1038/nmeth.4264

Doerge, R. (2002). Multifactorial Geneticsmapping and Analysis of Quantitative Trait Loci in Experimental Populations. *Nature Reviews Genetics*, *3*(1), 43–52. http://doi.org/10.1038/nrg703

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213), 1258096–1258096. http://doi.org/10.1126/science.1258096

Dove, P. (2002). Genetic polymorphisms in milk protein genes and their impact on milk composition. In *Biology of the Mammary Gland* (pp. 225–230). Springer.

Dvinge, H., & Bradley, R. K. (2015). Widespread intron retention diversifies most cancer transcriptomes. *Genome Medicine*, *7*(1), 45. http://doi.org/10.1186/s13073-015-0168-9

Edwards, S. L., Beesley, J., French, J. D., & Dunning, M. (2013). Beyond GWASs: Illuminating the dark road from association to function. *American Journal of Human Genetics*, *93*(5), 779–797. http://doi.org/10.1016/j.ajhg.2013.10.012

Encode Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. http://doi.org/10.1038/nature11247

Fairbrother, W. G., Yeo, G. W., Yeh, R., Goldstein, P., Mawson, M., Sharp, P. A., & Burge, C. B. (2004). RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Research*, *32*, 187–190. http://doi.org/10.1093/nar/gkh393

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics* (4th Editio). Pearson.

Farnir, F., Coppieters, W., Arranz, J., Coppieters, W., Berzi, P., Cambisano, N., … Georges, M. (2000). Extensive Genome-wide Linkage Disequilibrium in Cattle Extensive Genome-wide Linkage Disequilibrium in Cattle, 220–227. http://doi.org/10.1101/gr.10.2.220

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of

Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, *52*(2), 399–433. http://doi.org/DOI: 10.1017/S0080456800012163

Fitarelli-Kiehl, M., Macedo, G. S., Schlatter, R. P., Koehler-Santos, P., Da Silveira Matte, U., Ashton-Prolla, P., & Giacomazzi, J. (2016). Comparison of multiple genotyping methods for the identification of the cancer predisposing founder mutation p.R337H in TP53. *Genetics and Molecular Biology*, *39*(2), 203–209. http://doi.org/10.1590/1678-4685-GMB-2014-0351

Flint, J., & Mackay, T. F. C. (2009). Genetic architecture of quantitative traits in mice , flies , and humans, 723–733. http://doi.org/10.1101/gr.086660.108.19

Fontanesi, L., Calo, D. G., Galimberti, G., Negrini, R., Marino, R., Nardone, A., … Russo, V. (2014). A candidate gene association study for nine economically important traits in Italian Holstein cattle. *Animal Genetics*, *45*(4), 576–580. http://doi.org/10.1111/age.12164

Fortes, M. R. S., Kemper, K., Sasazaki, S., Reverter, a., Pryce, J. E., Barendse, W., … Lehnert, S. a. (2013). Evidence for pleiotropism and recent selection in the PLAG1 region in Australian Beef cattle. *Animal Genetics*, *44*(6), 636–647. http://doi.org/10.1111/age.12075

Fortes, M. R. S., Reverter, a., Kelly, M., Mcculloch, R., & Lehnert, S. a. (2013). Genome-wide association study for inhibin, luteinizing hormone, insulin-like growth factor 1, testicular size and semen traits in bovine species. *Andrology*, *1*(4), 644–650. http://doi.org/10.1111/j.2047-2927.2013.00101.x

Friggens, N. C., Ingvartsen, K. L., & Emmans, G. C. (2004). Prediction of body lipid change in pregnancy and lactation. *Journal of Dairy Science*, *87*(4), 988–1000. http://doi.org/10.3168/jds.S0022-0302(04)73244-0

Fürbass, R., Winter, A., Fries, R., & Kühn, C. (2006). Alleles of the bovine DGAT1 variable number of tandem repeat associated with a milk fat QTL at chromosome 14 can stimulate gene expression. *Physiological Genomics*, *25*, 116–120. http://doi.org/10.1152/physiolgenomics.00145.2005

Gaffney, D. J., Veyrieras, J., Degner, J. F., Pique-regi, R., Pai, A. A., Crawford, G. E., … Pritchard, J. K. (2012). Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology*, *13*(1), R7. http://doi.org/10.1186/gb-2012-13-1-r7

Ge, Y., & Porse, B. T. (2014). The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression. *BioEssays*, *36*(3), 236–243. http://doi.org/10.1002/bies.201300156

Georges, M. (2007). Mapping, fine mapping, and molecular dissection of quantitative trait Loci in domestic animals. *Annual Review of Genomics and Human Genetics*, *8*, 131–62. http://doi.org/10.1146/annurev.genom.8.080706.092408

Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A. T., … Hoeschele, I. (1995). Mapping Quantitative Trait Loci Controlling Milk Production in Dairy Cattle by Exploiting Progeny Testing. *Genetics Society of America*, (139), 907–920.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., & Thompson, R. (2009). ASReml user guide release 3.0. *VSN International Ltd*.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. *Biometrics*, *51*(4), 1440–1450. http://doi.org/10.2307/2533274

Glazier, A. M. (2002). Finding Genes That Underlie Complex Traits. *Science*, *298*(5602), 2345–2349. http://doi.org/10.1126/science.1076641

Goddard, M. E., & Hayes, B. J. (2009). Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nature Reviews. Genetics*, *10*(6), 381–91. http://doi.org/10.1038/nrg2575

Grala, T. M., Phyn, C. V. C., Kay, J. K., Rius, A. G., Littlejohn, M. D., Snell, R. G., & Roche, J. R. (2011). Temporary alterations to milking frequency, immediately post-calving, modified the expression of genes regulating milk synthesis and apoptosis in the bovine mammary gland. *Proc. N. Z. Soc. Anim. Prod.*, *71*, 3–8.

Grisart, B., Coppieters, W., & Farnir, F. (2002). Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Research*, 222–231. http://doi.org/10.1101/gr.224202.1

Grisart, B., Coppieters, W., Farnir, F., Karim, L., Ford, C., Berzi, P., … Snell, R. (2002). Positional Candidate Cloning of a QTL in Dairy Cattle : Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. *Genome Research*, *12*, 222–231. http://doi.org/10.1101/gr.224202.1

Grisart, B., Coppieters, W., Riquet, J., Berzi, P., Cambisano, N., Karim, L., … Georges, M. (2002). Simultaneous Mining of Linkage and Linkage Disequilibrium to Fine Map Quantitative Trait Loci in Outbred Half-Sib Pedigrees : Revisiting the Location of a Quantitative Trait Locus With Major Effect on Milk Production on Bovine Chromosome 14, *287*, 275–287.

Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J.-J., Kvasz, A., … Georges, M. (2004). Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proceedings of the National Academy of Sciences*, *101*(8). Retrieved from http://www.pnas.org/content/101/8/2398.full.pdf+html&gt

Grobarczyk, B., Malgrange, B., Grobarczyk, B., Franco, B., Hanon, K., & Malgrange, B. (2015). Generation of Isogenic Human iPS Cell Line Precisely Corrected by Genome Editing Using the CRISPR / Cas9 System Generation of Isogenic Human iPS Cell Line Precisely Corrected by Genome Editing Using the CRISPR / Cas9 System. *Stem Cell Reviews and Reports*, *11*(5), 774–787. http://doi.org/10.1007/s12015-015-9600-1

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V, Zusmanovich, P., … Stefansson, K. (2008). Many sequence variants affecting diversity of adult human height. *Nature Genetics*, *40*(5), 609–615. http://doi.org/10.1038/ng.122

Guernet, A., & Grumolato, L. (2017). CRISPR/Cas9 editing of the genome for cancer modeling. *Methods*. http://doi.org/10.1016/j.ymeth.2017.03.007

Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Georgiev, S., Daly, M. J., … Chase, C. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics*, *48*(1), 22–29. http://doi.org/10.1038/ng.3461.Abundant

Handel, A. E., Gallone, G., Cader, M. Z., & Ponting, C. P. (2017). Most brain disease-associated and eQTL haplotypes are not located within transcription factor DNase-seq footprints in brain. *Human Molecular Genetics*, *26*(1), ddw369. http://doi.org/10.1093/hmg/ddw369

Hayden, E. C. (2014). The $1,000 genome. *Nature*, *507*, 295. http://doi.org/10.1038/nature.2014.14530

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2008). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, *1646*(92), 433–443.

Hayes, B. J., Pryce, J., Chamberlain, A. J., Bowman, P. J., & Goddard, M. E. (2010). Genetic architecture of complex traits and accuracy of genomic Prediction: Coat colour, Milk-fat percentage, and type in holstein cattle as contrasting model traits. *PLoS Genetics*, *6*(9). http://doi.org/10.1371/journal.pgen.1001139

Hayes, P. L., Lytle, B. L., Volkman, B. F., & Peterson, F. C. (2008). The solution structure of ZNF593 from Homo sapiens reveals a zinc finger in a predominately unstructured protein, (Iuchi 2001), 571–576. http://doi.org/10.1110/ps.073290408.including

Hennighausen, L., & Robinson, G. W. (2005). Information networks in the mammary gland. *Nature Reviews. Molecular Cell Biology*, *6*(9), 715–725. http://doi.org/10.1038/nrm1714

Hensen, K., Braem, C., Declercq, J., Van Dyck, F., Dewerchin, M., Fiette, L., … Van De Ven, W. J. M. (2004). Targeted disruption of the murine Plag1 proto-oncogene causes growth retardation and reduced fertility. *Development Growth and Differentiation*, *46*(5), 459–470. http://doi.org/10.1111/j.1440-169x.2004.00762.x

Holmes, C. W., Brookes, I. M., Garrick, D. J., Mackenzie, D. D. ., Parkinson, T., & Wilson, G. F. (2002). *Milk Production from Pasture*. Palmerston North, New Zealand: Massey University.

Hoshiba, H., Setoguchi, K., Watanabe, T., Kinoshita, A., Mizoshita, K., Sugimoto, Y., & Takasuga, A. (2013). Comparison of the effects explained by variations in the bovine PLAG1 and NCAPG genes on daily body weight gain, linear skeletal measurements and carcass traits in Japanese Black steers from a progeny testing program. *Animal Science Journal*, *84*(7), 529–534. http://doi.org/10.1111/asj.12033

Hsu, P. D., Lander, E. S., & Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, *157*(6), 1262–1278. http://doi.org/10.1016/j.cell.2014.05.010

Huynh, H. T., Robitaille, G., & Turner, J. D. (1991). Establishment of bovine mammary epithelial cells (MAC-T): an in vitro model for bovine lactation. *Experimental Cell Research*, *197*(2), 191–199. http://doi.org/10.1016/0014-4827(91)90422-Q

Ilie, D. E., Magdin, A., Sălăjeanu, A., Neamţ, R., & Vintilă, I. (2009). Influence of CSN3 Marker on Milk Composition in Romanian Brown and Romanian Simmental. *Facultatea de Zootehnie Si Biotehnologii, Timisoara*, *42*(1), 54–57.

Isaac, R. S., Jiang, F., Doudna, J. A., Lim, W. A., Narlikar, G. J., & Almeida, R. (2016). Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *eLife*, *5*, 1–14. http://doi.org/10.7554/eLife.13450

Iso-Touru, T., Sahana, G., Guldbrandtsen, B., Lund, M. S., & Vilkki, J. (2016). Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. *BMC Genetics*, *17*(1), 55. http://doi.org/10.1186/s12863-016-0363-8

Jensen, R. G. (2002). The Composition of Bovine Milk Lipids: January 1995 to December 2000. *Journal of Dairy Science*, *85*(2), 295–350. http://doi.org/10.3168/jds.S0022-0302(02)74079-4

Jiang, L., Liu, J., Sun, D., Ma, P., Ding, X., Yu, Y., & Zhang, Q. (2010). Genome Wide Association Studies for Milk Production Traits in Chinese Holstein Population, *5*(10). http://doi.org/10.1371/journal.pone.0013661

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, *337*(6096), 816 LP-821. Retrieved from http://science.sciencemag.org/content/337/6096/816.abstract

Joung, J. K., & Sander, J. D. (2013). TALENs: a widely applicable technology for targeted genome editing. *Nat Rev Mol Cell Biol*, *14*(1), 49–55. http://doi.org/10.1038/nrm3486

Juma, A. R., Damdimopoulou, P. E., Grommen, S. V. H., Van de Ven, W. J. M., & De Groef, B. (2016). Emerging role of PLAG1 as a regulator of growth and reproduction. *Journal of Endocrinology*, *228*(2), R45–R56. http://doi.org/10.1530/JOE-15-0449

Karim, L., Takeda, H., Lin, L., Druet, T., Arias, J. a C., Baurain, D., … Coppieters, W. (2011). Variants modulating the expression of a chromosome domain encompassing PLAG1 influence bovine stature. *Nature Genetics*, *43*(5), 405–413. http://doi.org/10.1038/ng.814

Kaupe, B., Winter, A., Fries, R., & Erhardt, G. (2004). DGAT1 polymorphism in Bos indicus and Bos taurus cattle breeds. *J.Dairy Res.*, *71*(2), 182–187. http://doi.org/10.1017/S0022029904000032

Kellner, M., Rohrmoser, M., Forné, I., Voss, K., Burger, K., Mühl, B., … Eick, D. (2015). DEAD-box helicase DDX27 regulates 3 0 end formation of ribosomal 47S RNA and stably associates with the PeBoW-complex. *Experimental Cell Research*, *334*(1), 146–159. http://doi.org/10.1016/j.yexcr.2015.03.017

Kemper, K. E., Hayes, B. J., Daetwyler, H. D., & Goddard, M. E. (2015). How old are quantitative trait loci and how widely do they segregate? *Journal of Animal Breeding and Genetics*, *132*(2), 121–134. http://doi.org/10.1111/jbg.12152

Kemper, K. E., Reich, C. M., Bowman, P. J., Vander Jagt, C. J., Chamberlain, A. J., Mason, B. A., … Goddard, M. E. (2015). Improved precision of QTL mapping using a nonlinear Bayesian method in a multi-breed population leads to greater accuracy of across-breed genomic predictions. *Genetics Selection Evolution*, *47*(29). http://doi.org/10.1186/s12711-014-0074-4

Kemper, K. E., Visscher, P. M., & Goddard, M. E. (2012). Genetic architecture of body size in mammals. *Genome Biology*, *13*(4), 244. http://doi.org/10.1186/gb4016

Khatib, H., Leonard, S. D., Schutzkus, V., Luo, W., & Chang, Y. M. (2017). Association of the <em>OLR1</em> Gene with Milk Composition in Holstein Dairy Cattle. *Journal of Dairy Science*, *89*(5), 1753–1760. http://doi.org/10.3168/jds.S0022-0302(06)72243-3

Khatkar, M. S., Thomson, P. C., Tammen, I., & Raadsma, H. W. (2004a). Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genetics, Selection, Evolution : GSE*, *36*(2), 163–190. http://doi.org/10.1186/1297-9686-36-2-163

Khatkar, M. S., Thomson, P. C., Tammen, I., & Raadsma, H. W. (2004b). Quantitative trait loci mapping in dairy cattle: review and meta-analysis. *Genetics, Selection, Evolution : GSE*, *36*(2), 163–90. http://doi.org/10.1186/1297-9686-36-2-163

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. http://doi.org/10.1186/gb-2013-14-4-r36

Kim, H., Um, E., Cho, S.-R., Jung, C., Kim, H., & Kim, J.-S. (2011). Surrogate reporters for enrichment of cells with nuclease-induced mutations. *Nature Methods*, *8*(11), 941–943. http://doi.org/10.1038/nmeth.1733

Klann, T. S., Black, J. B., Chellappan, M., Safi, A., Song, L., Hilton, I. B., … Gersbach, C. A. (2017). CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature Biotechnology*, (September 2016). http://doi.org/10.1038/nbt.3853

Komisarek, J., Michalak, A., & Walendowska, A. (2011). The effects of polymorphisms in DGAT1, GH and GHR genes on reproduction and production traits in jersey cows. *Animal Science Papers and Reports*, *29*(1), 29–36.

Kühn, C., Thaller, G., Winter, A., Bininda-Emonds, O. R. P., Kaupe, B., Erhardt, G., … Fries, R. (2004). Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle. *Genetics*, *167*(4), 1873–81. http://doi.org/10.1534/genetics.103.022749

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. http://doi.org/10.1038/nature14248

Kuss, A. W., Gogol, J., & Geldermann, H. (2003). Associations of a Polymorphic AP-2 Binding Site in the 5′-Flanking Region of the Bovine β-Lactoglobulin Gene with Milk Proteins. *Journal of Dairy Science*, *86*(6), 2213–2218. http://doi.org/10.3168/jds.S0022-0302(03)73811-9

Lander, E. S., & Botstein, D. (1989). Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics Society of America*, (121), 185–199.

Lander, E. S., & Schork, N. J. (1994). Genetic Dissection of Complex Traits. *Science*, *265*(5181), 2037–2048. http://doi.org/10.1016/S0065-2660(07)00409-9

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., … Hirschhorn, J. N. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, *467*(7317), 832–838. http://doi.org/10.1038/nature09410

Lappalainen, T. (2015). Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res.*, *25*, 1427–1431. http://doi.org/10.1101/gr.190983.115.

Lappalainen, T., Sammeth, M., Friedländer, M. R., ′t Hoen, P. a C., Monlong, J., Rivas, M. a, … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–11. http://doi.org/10.1038/nature12531

Lehnert, K., Ward, H., Berry, S. D., Ankersmit-Udy, A., Burrett, A., Beattie, E. M., … Snell, R. G. (2015). Phenotypic population screen identifies a new mutation in bovine DGAT1 responsible for unsaturated milk fat. *Scientific Reports*, *5*, 8484. http://doi.org/10.1038/srep08484

Lemay, D. G., Lynn, D. J., Martin, W. F., Neville, M. C., Casey, T. M., Rincon, G., … Rijnkels, M. (2009). The bovine lactation genome: insights into the evolution of mammalian milk. *Genome Biology*, *10*(4), R43. http://doi.org/10.1186/gb-2009-10-4-r43

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., … Hirschhorn, J. N. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genetics*, *40*(5), 584–591. http://doi.org/10.1038/ng.125

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Preprint arXiv*, *0*(0), 3. http://doi.org/arXiv:1303.3997 [q-bio.GN]

Li, M., Sun, X., Hua, L., Huang, Y., Wang, J., Cao, X., … Chen, H. (2013). Molecular characterization, alternative splicing and expression analysis of bovine DBC1. *Gene*, *527*(2), 689–693. http://doi.org/10.1016/j.gene.2013.05.065

Liang, X., Potter, J., Kumar, S., Zou, Y., Quintanilla, R., Sridharan, M., … Chesnut, J. D. (2015). Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *Journal of Biotechnology*, *208*, 44–53. http://doi.org/10.1016/j.jbiotec.2015.04.024

Lin, S., Staahl, B., Alla, R. K., & Doudna, J. a. (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife*, *3*, 1–13. http://doi.org/10.7554/eLife.04766

Litman, T., Brangi, M., Hudson, E., Fetsch, P., Abati, A., Ross, D. D., … Bates, S. E. (2000). The multidrug-resistant phenotype associated with overexpression of the new ABC half-transporter, MXR (ABCG2). *Journal of Cell Science*, *113 (Pt 1*(April 2017), 2011–21. http://doi.org/10.1055/s-0028-1096972

Littlejohn, M. D., Henty, K. M., Tiplady, K., Johnson, T., Harland, C., Lopdell, T., … Davis, S. R. (2014). Functionally reciprocal mutations of the prolactin signalling pathway define hairy and slick cattle. *Nature Communications*, *5*, 1–8. http://doi.org/10.1038/ncomms6861

Littlejohn, M. D., Tiplady, K., Fink, T. A., Lehnert, K., Lopdell, T., Johnson, T., … Spelman, R. J. (2016). Sequence-based association analysis reveals an MGST1 eQTL with pleiotropic effects on bovine milk composition. *Scientific Reports*, *6*(25376), 1–14. http://doi.org/10.5061/dryad.457br

Littlejohn, M. D., Tiplady, K., Lopdell, T., Law, T. a, Scott, A., Harland, C., … Snell, R. G. (2014). Expression variants of the lipogenic AGPAT6 gene affect diverse milk composition phenotypes in Bos taurus. *PloS One*, *9*(1), e85757. http://doi.org/10.1371/journal.pone.0085757

Littlejohn, M., Grala, T., Sanders, K., Walker, C., Waghorn, G., MacDonald, K., … Snell, R. (2012). Genetic variation in PLAG1 associates with early life body weight and peripubertal weight and growth in Bos taurus. *Animal Genetics*, *43*(5), 591–594. http://doi.org/10.1111/j.1365-2052.2011.02293.x

MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., Abecasis, G. R., … Gunter, C. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, *508*(7497), 469–476. Retrieved from http://dx.doi.org/10.1038/nature13127

Machlin, L. J. (1973). Effect of growth hormone on milk production and feed utilization in dairy cows. *Journal of Dairy Science*, *56*(5), 575–580. http://doi.org/10.3168/jds.S0022-0302(73)85221-X

Mackay, T. F. C. (2001). The Genetic Architecture of Quantitative Traits. *Annual Review of Genetics*, *35*, 303–339.

Mackay, T. F. C., Stone, E. a, & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, *10*(8), 565–577. http://doi.org/10.1038/nrg2612

Makrythanasis, P., & Antonarakis, S. (2013). Pathogenic variants in non-protein-coding sequences. *Clinical Genetics*, *84*(5), 422–428. http://doi.org/10.1111/cge.12272

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. a. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, *337*(6099), 1190–1195.

http://doi.org/10.1126/science.1222794

Mayorek, N., Grinstein, I., & Bar-tana, J. (1989). Triacylglycerol synthesis in cultured rat hepatocytes. *European Journal of Biochemistry*, *182*, 395–400.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., & Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, *9*(5), 356–369. Retrieved from http://dx.doi.org/10.1038/nrg2344

McGowan, K. a, Li, J. Z., Park, C. Y., Beaudry, V., Tabor, H. K., Sabnis, A. J., … Barsh, G. S. (2008). Ribosomal mutations cause p53-mediated dark skin and pleiotropic effects. *Nature Genetics*, *40*(8), 963–70. http://doi.org/10.1038/ng.188

McManaman, J. L., & Neville, M. C. (2003). Mammary physiology and milk secretion. *Advanced Drug Delivery Reviews*, *55*(5), 629–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12706546

Mendel, G. (1865). Experiments in Plant Hybridization. *Journal of the Royal Horticultural Society*, *IV*(1865), 3–47. Retrieved from http://www.esp.org/foundations/genetics/classical/gm-65.pdf

Mercer, T. R., Edwards, S. L., Clark, M. B., Neph, S. J., Wang, H., Stergachis, A. B., … Stamatoyannopoulos, J. a. (2013). DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nature Genetics*, *45*(8), 852–9. http://doi.org/10.1038/ng.2677

Mersch, B., Gepperth, A., Suhai, S., & Hotz-Wagenblatt, A. (2008). Automatic detection of exonic splicing enhancers (ESEs) using SVMs. *BMC Bioinformatics*, *9*, 369. http://doi.org/10.1186/1471-2105-9-369

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829. http://doi.org/11290733

Moioli, B., Contarini, G., Avalli, A., Catillo, G., Orrù, L., De Matteis, G., … Napolitano, F. (2007). Short Communication: Effect of Stearoyl-Coenzyme A Desaturase Polymorphism on Fatty Acid Composition of Milk. *Journal of Dairy Science*, *90*(7), 3553–3558. http://doi.org/http://doi.org/10.3168/jds.2006-855

Molee, A., Poompramun, C., & Mernkrathoke, P. (2015). Effect of casein genes - beta-LGB, DGAT1, GH, and LHR - on milk production and milk composition traits in crossbred Holsteins. *Genetics and Molecular Research*, *14*(1), 2561–2571. http://doi.org/10.4238/2015.March.30.15

Moorehead, R. a, Fata, J. E., Johnson, M. B., & Khokha, R. (2001). Inhibition of mammary epithelial apoptosis and sustained phosphorylation of Akt/PKB in MMTV-IGF-II transgenic mice. *Cell Death and Differentiation*, *8*(1), 16–29. http://doi.org/10.1038/sj.cdd.4400762

Morris, C. A., & Wtlron, A. (1976). Infleunce of Body Size on the Biological Efficiency of Cows: A Review. *Canadian Journal of Animal Science*, *56*(4), 613–647.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18516045

Nafikov, R. a, Schoonmaker, J. P., Korn, K. T., Noack, K., Garrick, D. J., Koehler, K. J., … Beitz, D. C. (2013). Sterol regulatory element binding transcription factor 1 (SREBF1) polymorphism and milk fatty acid composition. *Journal of Dairy Science*, *96*(4), 2605–16. http://doi.org/10.3168/jds.2012-6075

Neville, M. C. (1990). The physiological basis of milk secretion. *Annals of the New York Academy of Sciences*, *586*, 1–11. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2192630

Neville, M. C. (2001). Anatomy and physiology of lactation. *Pediatric Clinics of North America*, *48*(1), 13–34. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11236721

Neville, M. C., McFadden, T. B., & Forsyth, I. (2002). Hormonal regulation of mammary differentiation and milk secretion. *Journal of Mammary Gland Biology and Neoplasia*, *7*(1), 49–66. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12160086

Neville, M. C., Webb, P., Ramanathan, P., Mannino, M. P., Pecorini, C., Monks, J., … MacLean, P. (2013). The insulin receptor plays an important role in secretory differentiation in the mammary gland. *American Journal of Physiology. Endocrinology and Metabolism*, *305*(9), E1103-14. http://doi.org/10.1152/ajpendo.00337.2013

Ni, T., Yang, W., Han, M., Zhang, Y., Shen, T., Nie, H., … Zhu, J. (2016). Global intron retention mediated gene regulation during CD4+ T cell activation. *Nucleic Acids Research*, *44*(14), 1–13. http://doi.org/10.1093/nar/gkw591

Nishimura, S., Watanabe, T., Mizoshita, K., Tatsuda, K., Fujita, T., Watanabe, N., … Takasuga, A. (2012). Genome-wide association study identified three major QTL for carcass weight including the PLAG1-CHCHD7 QTN for stature in Japanese Black cattle. *BMC Genetics*, *13*(1), 40. http://doi.org/10.1186/1471-2156-13-40

Niu, Y., Shen, B., Cui, Y., Chen, Y., Wang, J., Wang, L., … Sha, J. (2014). Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell*, *156*(4), 836–843. http://doi.org/10.1016/j.cell.2014.01.027

Oftedal, O. T. (2002). The mammary gland and its origin during synapsid evolution. *Journal of Mammary Gland Biology and Neoplasia*, *7*(3), 225–52. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12751889

Oftedal, O. T. (2009). *Advanced Dairy Chemistry*. (P. McSweeney & P. F. Fox, Eds.) (4th Editio). New York, NY: Springer New York. http://doi.org/10.1007/978-0-387-84865-5

Oftedal, O. T., & Iverson, S. J. (1995). Comparative Analysis of Nonhuman Milks: A. Phylogenetic Variation in the Gross Composition of Milks A2. In *Handbook of Milk*

Composition (pp. 749–789). San Diego: Academic Press. http://doi.org/http://doi.org/10.1016/B978-012384430-9/50035-4

Olsen, H. G., Nilsen, H., Hayes, B., Berg, P. R., Svendsen, M., Lien, S., & Meuwissen, T. (2007). Genetic support for a quantitative trait nucleotide in the ABCG2 gene affecting milk composition of dairy cattle. *BMC Genetics*, *8*, 32. http://doi.org/10.1186/1471-2156-8-32

Paix, A., Folkmann, A., Rasoloson, D., & Seydoux, G. (2015). High Efficiency, Homology-Directed Genome Editing in Caenorhabditis elegans Using CRISPR-Cas9 Ribonucleoprotein Complexes. *Genetics*, *201*(1), 47–54. http://doi.org/10.1534/genetics.115.179382

Parma, P., Curik, I., Greppi, G. F., & Enne, G. (2005). Caprine a s1 -Casein Polymorphism : Characterisation of A , B , E and F Variants by Means of Various Biochemical and Molecular Techniques, *43*(2), 123–132.

Patrick, H. D., Eric, L. S., & Zhang, F. (2014). Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell*, *157*(6), 1262–1278. http://doi.org/10.1016/j.cell.2014.05.010.Development

Paul, D. S., Soranzo, N., & Beck, S. (2014). Functional interpretation of non-coding sequence variation: concepts and challenges. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *36*(2), 191–9. http://doi.org/10.1002/bies.201300126

Petersen, B., & Niemann, H. (2015). Molecular scissors and their application in genetically modified farm animals. *Transgenic Research*. Springer International Publishing. http://doi.org/10.1007/s11248-015-9862-z

Prosser, C. G., Davis, S. R., Farr, V. C., Moore, L. G., & Gluckman, P. D. (1994). Effects of close-arterial (external pudic) infusion of insulin-like growth factor-II on milk yield and mammary blood flow in lactating goats. *Journal of Endocrinology*, *142*, 93–9.

Rachagani, S., & Gupta, I. D. (2008). Bovine kappa-casein gene polymorphism and its association with milk production traits, *897*, 893–897.

Ramlee, M. K., Yan, T., Cheung, A. M. S., & Chuah, C. T. H. (2015). High-throughput genotyping of CRISPR / Cas9-mediated mutants using fluorescent PCR-capillary gel electrophoresis. *Nature Publishing Group*, (October), 1–13. http://doi.org/10.1038/srep15587

Ran, F. A., Hsu, P. D. P., Wright, J., Agarwala, V., Scott, D. a, & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308. http://doi.org/10.1038/nprot.2013.143.Genome

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nat. Protocols*, *8*(11), 2281–2308. http://doi.org/10.1038/nprot.2013.143\rhttp://www.nature.com/nprot/journal/v8/n11/abs/nprot.2013.143.html#supplementary-information

Raven, L.-A., Cocks, B. G., Goddard, M. E., Pryce, J. E., & Hayes, B. J. (2014). Genetic variants in mammary development, prolactin signalling and involution pathways explain considerable variation in bovine milk production and milk composition. *Genetics, Selection, Evolution : GSE*, *46*(1), 29. http://doi.org/10.1186/1297-9686-46-29

Raven, L.-A., Cocks, B. G., & Hayes, B. J. (2014). Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics*, *15*(1), 62. http://doi.org/10.1186/1471-2164-15-62

Raven, L.-A., Cocks, B. G., Kemper, K. E., Chamberlain, A. J., Vander Jagt, C. J., Goddard, M. E., & Hayes, B. J. (2015). Targeted imputation of sequence variants and gene expression profiling identifies twelve candidate genes associated with lactation volume, composition and calving interval in dairy cattle. *Mammalian Genome*. http://doi.org/10.1007/s00335-015-9613-8

Raychaudhuri, S. (2011). Mapping rare and common causal alleles for complex human diseases. *Cell*, *147*(1), 57–69. http://doi.org/10.1016/j.cell.2011.09.011

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics*, *13*(5), 278–289. http://doi.org/http://doi.org/10.1016/j.gpb.2015.08.002

Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L., & Corn, J. E. (2016). Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nature Biotechnology*, *34*(3), 339–44. http://doi.org/10.1038/nbt.3481

Rijnkels, M., Kooiman, P. M., de Boer, H. A., & Pieper, F. R. (1997). Organization of the bovine casein gene locus. *Mammalian Genome*, *8*(2), 148–152. http://doi.org/10.1007/s003359900377

Ron, M., & Weller, J. I. (2007). From QTL to QTN identification in livestock - Winning by points rather than knock-out: A review. *Animal Genetics*. http://doi.org/10.1111/j.1365-2052.2007.01640.x

Rouet, P., Smih, F., & Jasin, M. (1994). Introduction of Double-Strand Breaks into the Genome of Mouse Cells by Expression of a Rare-Cutting Endonuclease. *Molecular and Cellular Biology*, *14*(12), 8096–8106.

Rudolph, M. C., McManaman, J. L., Phang, T., Russell, T., Kominsky, D. J., Serkova, N. J., … Neville, M. C. (2007). Metabolic regulation in the lactating mammary gland: a lipid synthesizing machine. *Physiological Genomics*, *28*(3), 323–36. http://doi.org/10.1152/physiolgenomics.00020.2006

Rupp, R., Senin, P., Sarry, J., Allain, C., Tasca, C., Ligat, L., … Bouchez, O. (2015). RESEARCH ARTICLE A Point Mutation in Suppressor of Cytokine Signalling 2 (Socs2) Increases the Susceptibility to Inflammation of the Mammary Gland while Associated with Higher Body Weight and Size and Higher Milk Production in a Sheep Model, *2*, 1–19. http://doi.org/10.1371/journal.pgen.1005629

Sanger, F., Nicklen, S., & Coulson, a R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–7. http://doi.org/10.1073/pnas.74.12.5463

Sanjana. (2016). High-resolution interrogation of functional elements in the noncoding genome. *Science*.

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, *489*(7414), 109–13. http://doi.org/10.1038/nature11279

Schaeffer, L. R. (2006). Strategy for applying genome wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, *123*(4), 218–223.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., & Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Research*, *22*(9), 1748–1759. http://doi.org/10.1101/gr.136127.111

Schennink, a., Bovenhuis, H., Léon-Kloosterziel, K. M., Van Arendonk, J. a M., & Visker, M. H. P. W. (2009). Effect of polymorphisms in the FASN, OLR1, PPARGC1A, PRL and STAT5A genes on bovine milk-fat composition. *Animal Genetics*, *40*(6), 909–916. http://doi.org/10.1111/j.1365-2052.2009.01940.x

Schennink, a, Stoop, W. M., Visker, M. H. P. W., Heck, J. M. L., Bovenhuis, H., van der Poel, J. J., … van Arendonk, J. a M. (2007). DGAT1 underlies large genetic variation in milk-fat composition of dairy cows. *Animal Genetics*, *38*(5), 467–73. http://doi.org/10.1111/j.1365-2052.2007.01635.x

Shalem, O., Sanjana, E. N., Hartenian, E., & Zhang, F. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science*, *343*(January), 84–88. http://doi.org/10.1038/nbt.2647

Sheffield, N. C., Thurman, R. E., Song, L., Safi, A., Stamatoyannopoulos, J. a, Lenhard, B., … Furey, T. S. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. *Genome Research*, *23*(5), 777–88. http://doi.org/10.1101/gr.152140.112

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. http://doi.org/10.1038/nbt1486

Smith, R. M., Webb, A., Papp, A. C., Newman, L. C., Handelman, S. K., Suhy, A., … Sadee, W. (2013). Whole transcriptome RNA-Seq allelic expression in human brain. *BMC Genomics*, *14*(1), 571. http://doi.org/10.1186/1471-2164-14-571

Song, L., & Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor Protocols*, *2010*(2), pdb.prot5384. http://doi.org/10.1101/pdb.prot5384

Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., … Furey, T. S. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, *21*(10), 1757–1767. http://doi.org/10.1101/gr.121541.111

Spain, S. L., & Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human Molecular Genetics*, *24*(R1), R111–R119. http://doi.org/10.1093/hmg/ddv260

Spelman, R. J., Hooper, J. D., Stanley, G., Kayis, S. A., & Harcourt, S. (2004). Friesian Jersey crossbred trial: Generating phenotypes for the discovery of quantitative trait loci. *Proceedings of the New Zealand Society of Animal Production*, *64*, 92–95.

Spelman, R. J., Miller, F. M., Hooper, J. D., Thielen, M., & Garrick, D. J. (2001). Experimental design for QTL Trial involving New Zealand Friesian and Jersey breeds. *Proceedings of the 14th AAABG Conference*, *14*, 393–396. http://doi.org/10.1017/CBO9781107415324.004

Steinmetz, L. M., Sinha, H., Richards, D. R., Spiegelman, J. I., Oefner, P. J., McCusker, J. H., & Davis, R. W. (2002). Dissecting the architecture of a quantitative trait locus in yeast. *Nature*, *416*(6878), 326–30. http://doi.org/10.1038/416326a

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., … Dermitzakis, E. T. (2007). Expression Phenotypes. *Recherche*, *315*(February), 848–853.

Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, *187*(2), 367–383. http://doi.org/10.1534/genetics.110.120907

Strucken, E. M., Laurenson, Y. C. S. M., & Brockmann, G. a. (2015). Go with the flowâ€"biology and genetics of the lactation cycle. *Frontiers in Genetics*, *6*(March), 1–11. http://doi.org/10.3389/fgene.2015.00118

Szyda, J., Morek-Kopeć, M., Komisarek, J., & Zarnecki, A. (2011). Evaluating markers in selected genes for association with functional longevity of dairy cattle. *BMC Genetics*, *12*, 30. http://doi.org/10.1186/1471-2156-12-30

Takeuchi, K., & Reue, K. (2009). Biochemistry, physiology, and genetics of GPAT, AGPAT, and lipin enzymes in triglyceride synthesis. *American Journal of Physiology - Endocrinology and Metabolism*, *296*(6), E1195–E1209. http://doi.org/10.1152/ajpendo.90958.2008

Tamma, N., Sartelet, A., Druet, T., Michaux, C., Fasquelle, C., Ge, S., … Charlier, C. (2012). A Splice Site Variant in the Bovine RNF11 Gene Compromises Growth and Regulation of the Inflammatory Response, *8*(3). http://doi.org/10.1371/journal.pgen.1002581

Tellmann, G. (2006). The E-Method: a highly accurate technique for gene-expression analysis. *Nature Methods*, *3*, i–ii.

Terns, M. P., & Terns, R. M. (2011). CRISPR-based adaptive immune systems. *Current Opinion in Microbiology*, *14*(3), 321–327. http://doi.org/10.1016/j.mib.2011.03.005

Terunuma, A., Shiba, K., & Noda, T. (1997). A novel genetic system to isolate a dominant negative effector on DNA-binding activity of Oct-2, *25*(10), 1984–1990.

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., … Sabeti, P. C. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*, *165*(6), 1519–1529.

http://doi.org/10.1016/j.cell.2016.04.027

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., … Stamatoyannopoulos, J. a. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75–82. http://doi.org/10.1038/nature11232

Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., … Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. http://doi.org/10.1038/nbt.1621

Tsai, S. Q., Joung, J. K., Capecchi, M., & Evans, M. (2016). Defining and improving the genome- wide specificities of CRISPR – Cas9 nucleases. *Nature Publishing Group*, *17*(5), 300–312. http://doi.org/10.1038/nrg.2016.28

Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., … Sankaran, V. G. (2016). Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell*, *165*(6), 1530–1545. http://doi.org/10.1016/j.cell.2016.04.048

Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S., & Gregory, P. D. (2010). Genome editing with engineered zinc finger nucleases. *Nat Rev Genet*, *11*(9), 636–646. Retrieved from http://dx.doi.org/10.1038/nrg2842

Utsunomiya, Y. T., do Carmo, A. S., Carvalheiro, R., Neves, H. H. R., Matos, M. C., Zavarez, L. B., … Garcia, J. F. (2013). Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. *BMC Genetics*, *14*, 52. http://doi.org/10.1186/1471-2156-14-52

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., … Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, *5*(9), 829–834. http://doi.org/10.1038/nmeth.1246

Van Dyck, F., Declercq, J., Braem, C. V, & Van de Ven, W. J. M. (2007). PLAG1, the prototype of the PLAG gene family: versatility in tumour development (review). *International Journal of Oncology*, *30*(4), 765–74. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17332914

van Marle-Köster, E., Visser, C., & Berry, D. P. (2013). A review of genomic selection - implications for the south african beef and dairy cattle industries. *South African Journal of Animal Sciences*, *43*(1), 1–17. http://doi.org/10.4314/sajas.v43i1.1

VanRaden, P. M., Tooker, M. E., O'Connell, J. R., Cole, J. B., & Bickhart, D. M. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genetics Selection Evolution*, *49*(1), 32. http://doi.org/10.1186/s12711-017-0307-4

Vergnes, L., Beigneux, A. P., Davis, R., Watkins, S. M., Young, S. G., & Reue, K. (2006). Agpat6 deficiency causes subdermal lipodystrophy and resistance to obesity. *Journal of Lipid Research*, *47*(4), 745–54. http://doi.org/10.1194/jlr.M500553-JLR200

Viitala, S., Szyda, J., Blott, S., Schulman, N., Lidauer, M., Mäki-Tanila, A., … Vilkki, J. (2006).

The role of the bovine growth hormone receptor and prolactin receptor genes in milk, fat and protein production in Finnish Ayrshire dairy cattle. *Genetics*, *173*(4), 2151–2164. http://doi.org/10.1534/genetics.105.046730

Visscher, P. M., McEvoy, B., & Yang, J. (2010). From Galton to GWAS: quantitative genetics of human height. *Genetics Research*, *92*(5–6), 371–379. http://doi.org/10.1017/S0016672310000571

Visscher, P. M., & Yang, J. (2016). A plethora of pleiotropy across complex traits. *Nature Genetics*, *48*(370), 40133–40141. http://doi.org/10.1038/ng.3604

Vockley, C. M., Guo, C., Majoros, W. H., Nodzenski, M., Scholtens, D. M., Hayes, M. G., … Reddy, T. E. (2015). Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort, 1–9. http://doi.org/10.1101/gr.190090.115.8

Voz, M. L., Agten, N. S., Van de Ven, W. J., & Kas, K. (2000). PLAG1, the main translocation target in pleomorphic adenoma of the salivary glands, is a positive regulator of IGF-II. *Cancer Res*, *60*(1), 106–113. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10646861

Wagner, J. C., Platt, R. J., Goldfless, S. J., Zhang, F., & Niles, J. C. (2014). Efficient CRISPR-Cas9 – mediated genome editing in Plasmodium falciparum. *Nature Methods*, *11*(9). http://doi.org/10.1038/nmeth.3063

Wang, H., Russa, M. La, & Qi, L. S. (2016). CRISPR / Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry*, *85*(22), 1–38. http://doi.org/10.1146/annurev-biochem-060815-014607

Wang, T., Wei, J., Sabatini, D., & Lander, E. (2014). Genetic screens in human cells using the CRISPR/Cas9 system. *Science*, *343*(6166), 80–84. http://doi.org/10.1126/science.1246981.Genetic

Wang, X., Wurmser, C., Pausch, H., Jung, S., Reinhardt, F., Tetens, J., … Fries, R. (2012). Identification and dissection of four major QTL affecting milk fat content in the German Holstein-Friesian population. *PloS One*, *7*(7), e40711. http://doi.org/10.1371/journal.pone.0040711

Wang, Y., Shang, W., Lei, X., Shen, S., Zhang, H., Wang, Z., … Zhang, C. (2013). Opposing functions of PLAG1 in pleomorphic adenoma: A microarray analysis of PLAG1 transgenic mice. *Biotechnology Letters*, *35*(9), 1377–1385. http://doi.org/10.1007/s10529-013-1213-7

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949280&tool=pmcentrez&rendertype=abstract

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids. *Nature*.

http://doi.org/10.1097/BLO.0b013e3181468780

Westra, H., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., … Metspalu, A. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Publishing Group*, *45*(10), 1238–1243. http://doi.org/10.1038/ng.2756

Wickramasinghe, S., Rincon, G., Islas-trejo, A., & Medrano, J. F. (2012). Transcriptional profiling of bovine milk using RNA sequencing.

Wong, J. J. L., Au, A. Y. M., Ritchie, W., & Rasko, J. E. J. (2016). Intron retention in mRNA: No longer nonsense: Known and putative roles of intron retention in normal and disease biology. *BioEssays*, *38*(1), 41–49. http://doi.org/10.1002/bies.201500117

Wray, N. R., Yang, J., Hayes, B. J., Price, A. L., Goddard, M. E., & Visscher, P. M. (2013). Pitfalls of predicting complex traits from SNPs. *Nature Publishing Group*, *14*(7), 507–515. http://doi.org/10.1038/nrg3457

Wright, J. B., & Sanjana, N. E. (2016). CRISPR Screens to Discover Functional Noncoding Elements. *Trends in Genetics*, *32*(9), 526–529. http://doi.org/10.1016/j.tig.2016.06.004

Wright, S. (1921). Systems of mating. I. The biometric relations between parent and offspring. *Genetics*, *6*(111).

Wu, X., Scott, D. A., Kriz, A. J., Chiu, A. C., Hsu, P. D., Dadon, D. B., … Sharp, P. A. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat Biotechnol*, *32*(7), 670–676. http://doi.org/10.1038/nbt.2889

Yan, Q. C., Kuo, M. S., Li, S., Bui, H. H., Peake, D. A., Sanders, P. E., … Cao, G. (2008). AGPAT6 is a novel microsomal glycerol-3-phosphate acyltransferase. *Journal of Biological Chemistry*, *283*(15), 10048–10057. http://doi.org/10.1074/jbc.M708151200

Yu, X., Liang, X., Xie, H., Kumar, S., Ravinder, N., Potter, J., … Chesnut, J. D. (2016). Improved delivery of Cas9 protein/gRNA complexes using lipofectamine CRISPRMAX. *Biotechnology Letters*, *38*(6), 919–929. http://doi.org/10.1007/s10529-016-2064-9

Zhang, J.-L., Zan, L.-S., Fang, P., Zhang, F., Shen, G., & Tian, W.-Q. (2008). Genetic variation of PRLR gene and association with milk performance traits in dairy cattle. *Canadian Journal of Animal Science*, *88*, 33–39.

Ziemann, M., Kaspi, A., Lazarus, R., & El-Osta, A. (2013). Motif analysis in DNAse hypersensitivity regions uncovers distal cis elements associated with gene expression. *Bioinformation*, *9*(4), 212–5. http://doi.org/10.6026/97320630009212

Zuris, J. A., Thompson, D. B., Shu, Y., Guilinger, J. P., Bessen, J. L., Hu, J. H., … Liu, D. R. (2015). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nature Biotechnology*, *33*(1). http://doi.org/10.1038/nbt.3081

Appendix I

1. CRISPR-Cas9 reagent details
2. qPCR primer sequences and probe numbers

Table 1. CRISPR-Cas9 reagent details including target variant details, crRNA HDR and primer sequences and primer annealing temperature

Chr	bp	Target variant	Forward primer	Reverse primer	crRNAs	HDR template	PCR temp (^{o}C)
5	93944937	T/C	AGACAGTGCAGTGTGGTTGG	TGTGCATTACCTTACATCATTTCT	CTTGGGTTCTTCTCCCAGTG	TAAGACTATAATCCTCATGAGTATATCTATTCTACAGATGAAAAAAAATGGAAATTAGAGATATTGAGCAGCCACACTGGGAGAAGAACCCAAGTCTCTGATCTCACATCTCATAA	56
5	93945655	T/G	CAACCCCCATGATGTTCAAG	AGCCAGTTTTGCCAGTTTTC	AAGATTCTCATAGAATCAGA	CTCCACAGGTCAGTAGTCTTGAAAGCAAGTATGATAGATTCTGCCAGCTAAAGACTGTCACTTGCCATCTGATTCTATGAGAATCTTTATCATGCCCTGAAAGGAGTTCAGAGTTTATCTGAAAGAAGTC	60
5	93945738	T/C	AGACTGTCACTTGCCATCTGA	GGTTGCTGCTGGTTATGAGG	ATGAGAAGATACAATAAATC	GAAAGAAGTCACCCTGTGTTCCAGGAAAACTGGCAAAACTGGCTTTCAGATAGTTAGACATTTTCCGGGAGAAACTTTTATGAGCCAGATTTATTGTATCTTCTCATACCTAGAAAAGCACTAAAATCAT	62
5	93946027	T/A	ACTGACTTCTCCATCTACCTCT	CAGCAACAGTTGGGAAGAAAA	TTATCTTGCACTGAGAAATG	TGAGGGGCTGTCTCCTGGACTACAGTCCTCAGTAAGCCTCAACAAAAACTGAACTCATAGCTCTCACATTATGTTTTATTCAGTCCACATTTCTCAGTGCAAGATAATATTATACTTTTAGTAAC	60
5	93946548	G/C	GAACATTGGAGTGGGTCGC	TCCAGAGACATAGGATTTAG	GTGCACTGTGAAGTCGGAGA	TTCTCCAATGCATGAAAGTGAAAAGTGAAACTGAAGTCGCTCAGTAGTGTCCAACCCTCAGCGACCCCATGGACTGCAGCCTTCCAGGCTC	60

				GGA		CTCCGTCCATGGGATTTCCGAGGCAAGAGTACTGGAGTGGGGTGCCATTGCCTTCTCCGACTTCACAGTGCACTTGGAGTAATTAGCTT	
5	93947761	C/T	GTCCCTTTCATTGGTTGGCA	TCCCCACTTTATGCTCTATTCAC	TTTATTAACCTCATGTTGCA	CCTTTACAAAGGACAAAGACAGGCTCCAGAACTAGATGTAGTGGATCCAATTCTGTCATTTACCAGCTGTGATTAATGAATCTTCCCTGCAACATGAGGTTAATAAAACCTCCTACCTCACAAAACTCTT	60
5	93947989	T/A	GCTATGCCTCGGTACAAAATAAA	TGGGTCTCAGTTTCGTTATATGT	GTAAGTGCTAGGTAAGTATT	GCCTTTGAAAACTCTGAAATTGTATCCTCAATGCTTAGAACAGTGCCAAGGTAAGGGCTTATTTATTCACCAAATACTTACCTAGCACTTACCAGATCCCTCTGCCTTTCTTTTTTCATTTTTTTATTTCAAACTT	60
5	93948357	C/T	CAAATTCACAGTGGAGGGCC	AGAAGCGGGCTATTTCTCAGT	GGTGGGGGTGGGATTCTAGG	CCATCAAGGAGAATTCTGGACTACCCTGGTCTTCTGGGGAATTCTTTCTGTCAAATCCCCTTCCTGTTTTATCCCCCTCACTCTCCCCCTAGAATCCCACCCCCACCCCTTATTGTCTATCACAG	60
5	93948646	C/G	GCGATTTCAGACCTTCTTAAAGC	GGTCAGCATAGTTTCCTGAGC	AAAGAGAAAAGACAGTTCAG	ACCTGGGCCCTGAAGCATAGGGGGCCATGTCTGCAGTCACTCAGTCTTTTCCATAAATTCTACAACTCACAAACAGCAAGGAGACCACTGAACTGTCTTTTCTCTTTATCTCACGGGCAGTGTTC	60
5	93948718	G/C	TCAGTGGTCTCCTTGCTGTT	CTGAGGGTTTGAGAGTGGAGT	CTTCAGGGCCCAGGTGTTCG	CTGAGACTAGAAAAGGAGGTCAGCATAGTTTCCTGAGCATAGTCCCCCTCCCCCCACAGCCCAGCAGCATTCCCCGCACCCTGACCCCGAACACCTGGGCCCTGAAGCATAGGGGGCCATGTCTG	60
5	93948804	T/C	CCTATGCTTCAGGGCCCA	TCCAAAGTAAGAGGCCTGTGT	TTTTTCTGAGGGTTTGAGAG	AATGCTGCTGGGCTGTGGGGGGAGGGGGACTATGCTCAGGAAACTATGCTGACCTCCTTTTCTAGTCTCAGCATCAAAGAAACTCCACTCTCAAACCCTCAGAAAAACCGCCCCTGCCACAGAAA	62
5	93949810	G/A	CCCAAAGCTCAACTGGCG	TCTCTGCCGACCTGAAACTT	TTGGCTTGAGAATTCAAAGT	ACCCCCCAACCCCGTAGATAAACTGTCAAACTCCAGGATTAAAACTTGAGAGAATAATTCTTTTTGTGTTTAAGTGATGATTTACCTACTTTGAATTCTCAAGCCAAGGGAGCATCCACTTGGAA	60
5	93954748	T/C	AGACCGTTCTTGGATATTTCAGT	AAATGCCAGGGACAGAGGAG	TAATCTTACAAAGATTATTG	AGTTCATGGGGTCTTAAAAGAGTTGGACATGACTTAGTGATTAAACAACAACAACACAATACCTTTAAAAAGTGTATCATAATTCCACAATAATCTTTGTAAGATTAGAATGCTGATTGTTTTTCTTTCACTTTTTGCTTCACATGATCTTAAG	60

27	36198117	TGGC/T	GACGAGAGGGTC ACGTCAAG	AGCCCCGCTA GAGGTTCAT	TTACGCACGCCTG GGGCTGG	GAAGCGGCCGGCAGCGTACGACTCACCCAGCGCGAGGCTCCGGGC GCCAAATCCCTCCGCGCACCGTCCGTCCCGCCCTCGCCGCCGCCGC CGCCAGCCCCAGGCGTGCGTAAGAACGTGCGCGCGCCCGCCCTTT	60
27	36198117	TGGC/T	GACGAGAGGGTC ACGTCAAG	AGCCCCGCTA GAGGTTCAT	TTACGCACGCCTG GGGCTGG	GAAGCGGCCGGCAGCGTACGACTCACCCAGCGCGAGGCTCCGGGC GCCAAATCCCTCCGCGCACCGTCCGTCCCGCCCTCGCCGCCGCCGC CAGCCCCAGGCGTGCGTAAGAACGTGCGCGCGCCCGCCCTTT	60
27	36200888	T/C	CTGAAGACATGG GGTGGTTG	ATGACACTTG CTTAGGTGCC	TGTGCTGGAGAA TATGGGCC	TGATACAGGAAGACGTAAATGGCCTGCTATGATGAGGTTCTGTTTT ACCTAAAGAAACAAATCTGCARTACAGCTCAAGCTCGACCAGGCC CATATTCTCCAGCACACAGGACCTTTGCCCCTCA	60
27	36200968	T/C	GAGAATATGGGC CTGGTCGA	CCCCAAGAAC TGATAATGCT CC	TTACGTCTTCCTG TATCATT	ACTAAAATCCATTCCTCATGCTACAATAAGCGTTATTATACAGGGC ACTGATGACACTTGCTTAGGTGCCCACCCCRCCATGAGCCTAATGA TACAGGAAGACGTAAATGGCCTGCTATGATGAG	60
27	36202188	T/A	CAGACAGCAAGC TTCCACTG	GCTGTCTGAC CACGTGTAAC	TGAGCTGTAAAA ACAGACAC	CAGTAATGTATATGTTTCAGTAACATACAAGTTCTTAATGAGATCTT TTACATTCCTCTTTTTTCACCCTWAATCACTGAAATTCCAGTGTCTG TTTTTACAGCTCATTTCAATTTCTACTGTCC	60
27	36202636	GT/G	TCTTCTCAGACCA GGACCCT	TGTGAAACAT GGGACTTCTT TGT	TGTGCCGTCAGG GAAGTTTG	TCAACTTGCCACAAGGTTGAAACAACCCTAAATATGCAGAGAAAT CACTGAAACAGGAAAGAAAAAAAAAAAACTCATGACCAAAACCAC AAACTTCCCTGACGGCACAGTGGATGGGAATCTGACT	62
27	36202636	GT/G	TCTTCTCAGACCA GGACCCT	TGTGAAACAT GGGACTTCTT TGT	TGTGCCGTCAGG GAAGTTTG	TCAACTTGCCACAAGGTTGAAACAACCCTAAATATGCAGAGAAAT CACTGAAACAGGAAAGAAAAAAAAAAACTCATGACCAAAACCACA AACTTCCCTGACGGCACAGTGGATGGGAATCTGACT	62
27	36203904	G/C	GCTGTAAGCGTAT CCCTGGA	GGCATGCACA CAGAGGAAAT	TGTAAGAAACTT GCTTGAGT	AGCTCAGGTTTGCTTCACCCAGCAGCTCTGGGCTGCCCATCTCCTTG GAAATGTGTTTACACCTTTCCTCSTGAGCAGCAGCATCCTACTCAAG CAAGTTTCTTACACTCCACTCCTCTGACACA	60
27	36204066	T/C	TGACACACTCGA CAGCAGAT	GAGGCAGGA CTACATCTCA GA	CCTGGGCTCTATT TTGCTCT	CATGCCTTTATTTACCATCACATCAAAATGGGGGCATTTGCTGGTCT GTGACCTGACTGAGTAAATYGGAAAAGTGGATACCTGCCCAGAGC AAAATAGAGCCCAGGGCACAGAAAGACCCCTAG	62

27	36204680	CAG/ATC	TTTGGCAGGTGTTGTTGAAC	TGTGTGTGCTCTGCTCTTCA	TAACAGACTGGGCTTCGCAG	GTGGACTACAGCCCACCAGGSWKCTCTGTCCCTGGGATTCCCCAGGCAAGAAGACTAGAGTGGGCTGCCATTTCCTCCTCCAGGGGATCTTCCCAACCCAGGGATCGAACTCATCTTCTGTACTGGCAGGCAGATTCTTTACCACTCACCCCGCTGCGAAGCCCAGTCTGTTATTTTGAG	60
27	36206783	C/A	GGGGTTGAAGAGTCTCATTAGC	GAAGCAGCGGAAGTCATCAG	AGACCACCTTCCCTCCCGAA	TTCAGCCAAATCCAGAGTCACTTGGACCAACGTTCCCCCAGGAGTCCAGCCAGTCCAAGGAAACCCGAAGAACGAAGTKCTGAGCCTTTCGGGAGGGAAGGTGGTCTCATCACAGGAAGCTGATG	60
27	36209319	T/G	CGTCAACCAACACCAGCTTG	CCCTGTGGTAGAAGTGCTGA	AAAGTGGCCAGAAAGGCTGG	GTTCTAACCCCTGGACCACCAGGGAATTCCCAGAAGCACAGTTTAGTTTTACAAGATACCGTCAMATTTTCTTGGTAGATAACTCCTCCAGCCTTTCTGGCCACTTTTTTTTTTTTTTTAAGATT	60
27	36211257	GA/T	GCAGGAGCGATTCCTAAC	ATATATGGACACAAGACACC	GCACACTCCAAGGAGAAGAT	CGTCTCAGGAAGCATCCGGAGTGTCCTAATGTTGGGGCTGCTTCTGCGGCCCAGAGCTCCAGGCAGTGGGGTCAGTGAGGAGGCCCATCTTCTCCTTGGAGTGTGCCCTCTTTATCTCTTGAAA	56
27	36211257	GA/T	GCAGGAGCGATTCCTAAC	ATATATGGACACAAGACACC	GCACACTCCAAGGAGAAGAT	CGTCTCAGGAAGCATCCGGAGTGTCCTAATGTTGGGGCTGCTTCTGCGGCCCAGAGCTCCAGGCAGTGGGGGACAGTGAGGAGGCCCATCTTCTCCTTGGAGTGTGCCCTCTTTATCTCTTGAAA	56
27	36211708	T/C	CCTATTAGAAAAGTGTGAGTGGC	ATGCGACACAAACGGCAC	AAACCTGGATGAAACGCCTG	AGGGCCAGGGCCAGCTCCTCTCATCCCCAAGGTCGAGGGGACCAGGCAGCGCACACAGGCARCACATGCGGGCGTGGACGAAAACCGCAGGCGTTTCATCCAGGTTTCAACCCCGATGGTTTAAG	60
27	36212352	G/A	TGGCAATGACAGACCTTCAG	CAGAGGGTGAGAGCTGAAGG	GCTCTTGGGCAGGAGATACA	TCAGGCCAGGTGGGGCAGCCCAGCCAGTGGGTGGCCCGGGCGCACTCTGGGCTCTGTTCCGCYGGCACTGCTTCCAGAAGTTTCCCTGTATCTCCTGCCCAAGAGCATTTAGCAGATAAATCTGT	56
14	1802265	GC/AA	AAGGCCAAGGCTGGTGAG	GGGGCGAAGAGGAAGTAGTA	CGCTTGCTCGTAGCTTTGGC	CAGTCCCCCCAGCCCCCGGCAGGATCCTCACCGCGGTAGGTCAGGTTGTCGGGGTAGCTCACGGTGCGCTGGGCAGCTCCCCCGTTGGCCTTCTTACCTGCCAAAGCTACGAGCAAGCGGCAGGGGGCGGGTCGGGGGTGAG	64.5
11	103301781	G/A	AGCCATGAAGTG	GATTTGTCAG	ATTGTCACCCAG	CACAGCCTCCCTTGGTCTCTGAGGCCCAGCTCCCCTGCCTGCCCTGC	60

	CCTCCTG	GCGGCTCTAG	ACCATGAA	AACTCACCACCCACCCGGGCACCCTCGAACCTTCTGGATATCYAGGCCCTTCATGGTCTGGGTGACAATGAGGGCCTGGGCGCCACAAGTGA

All forward and reverse primers had Nextera adapters attached: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG and GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG, respectively.

Table 2. qPCR primer and probe numbers for four genes targeted by CRISPR-Cas9 genome editing

Gene	Forward primer	Reverse primer	Probe
MGST1	GCTTCGGCAAAGGAGAAA	CGATGTTTTCAAGGTCATTCAAG	142
AGPAT6	GCTCCGAAGTGAAGGATCG	GCTTTTATCCTGCACATGCTC	49
DGAT1	GGGAGGGCGGGGCTAGCA	ACTTGCCTCGGGACCGGCAG	66
LGB	CCCCCTGAGAGTGTATGTGG	GAGCACACTCACCGTTCTCC	165

Optimisation of Cas9 plasmid-mediated CRISPR-Cas9 genome editing

Experiment 1: Optimisation of Lipofectamine transfection reagent for the transfection of Cas9 plasmids

Lipofectamine® RNAiMAX (ThermoFisher) is recommended for the transfection of RNP and gRNA complexes (http://sg.idtdna.com/). Lipofectamine® LTX (ThermoFisher) has been highly successful at delivering plasmid DNA into the MAC-T cell line (see Chapter 2 and 5). The use of both of these transfection reagents was investigated for the delivery of Cas9 expression plasmids into these cells. Transfections using Lipofectamine® LTX and Lipofectamine® RNAiMAX were conducted to determine the transfection reagent and concentration that would achieve maximum delivery of plasmid DNA into the mammary cell line. Initial transfection conditions were based on the optimal transfection parameters derived from transfection of other plasmids into MAC-T cells (described in Chapter 2).

Cells were plated in a 24-well plate and grown for 24 hours in proliferation media to achieve ~70% confluency. For transfection, a 1 µL volume of transfection reagent was gently mixed with 24 µL Opti-MEM reduced serum media (Invitrogen). One thousand ng of plasmid DNA was diluted in 25 µL Opti-MEM and combined with the transfection reagent and incubated at room temperature for 5 minutes. Then, the 50 µL transfection mix was added to each well. After 24 hour incubation, cells were visualised on the Nikon Ti-E inverted light microscope to assess transfection efficiency. As the PX459 Cas9 expression plasmid does not contain a fluorescent marker, cells were co-transfected with pMAXGFP (Lonza) in a 1:1 ratio i.e. 500 ng of both PX459 and pMAXGFP.

Experiment 2: Reverse and forward transfections of gRNA into FACS sorted and unsorted cells

Experiments were conducted to optimise the transfection protocol that would achieve maximum delivery of Cas9 plasmid DNA into cells, and achieve the highest cleavage efficiency at the chr27:36198117T>TGGC target locus. To this end, the amount of PX459 plasmid was titrated by co-transfecting in a 1:1 ratio and 2:1 ratio with pMAXGFP, and subjecting the cell population to FACS before gRNAs were transfected by either forward or reverse transfection. At the same time, the unsorted cells (i.e. those that weren't subjected to FACS) were also transfected by either a forward and reverse transfection of gRNA (Figure 3.3).

Cells were plated at 7.5 x 10^5 cells/well in a 6-well cell culture plate in 3 mL of complete proliferation media (refer to General Methods for culture conditions). Transfection was undertaken 24 hours after plating (~70% confluency) using 5 µL Lipofectamine® LTX transfection reagent diluted in 300 µL Opti-MEM. For 1:1 ratio transfections, 2 µg each of PX549 and pMAXGFP was diluted in 300 µL Opti-MEM along with 5 µL PLUS reagent, while 3.35 µg PX459 and 1.65 µg pMAXGFP was diluted in the same conditions for the 2:1 ratio transfection. Each transfection of plasmid DNA was conducted in duplicate.

After 24 hours of incubation following transfection, cells were visualised using the Nikon TE Inverted microscope. Following visualisation, the media was removed, replaced with 750 µL trypsin-EDTA and incubated for 5 minutes. Following incubation, the trypsin was deactivated by adding 2.25 µL of full proliferation media and gently mixed. The cell suspensions were transferred to a 15 mL falcon tube and centrifuged for 5 minutes at 1300 g. Each duplicate was pooled at this point to ensure a cell yield sufficient for FACS. The media was carefully removed and the cell pellet was washed in 500 µL pre-warmed PBS and resuspended in 400 µL FACS Pre-Sort buffer (BD Biosciences) supplemented with 50 ng DAPI immediately prior to cell sorting.

Cells were gated based on GFP fluorescence, with DAPI-positive cells excluded from the sort. Cells were sorted into 15 mL falcon tubes containing 1 mL of complete proliferation media. Following sorting, cells were counted using a haemocytometer. For the reverse transfection of the gRNA complex, an aliquot of these cells was diluted to 400,000 cells/mL in antibiotic free proliferation media. For the forward transfection of the gRNA complex, an aliquot of these cells was diluted to 150,000 cells/mL in complete proliferation media. Similarly, two aliquots of unsorted cells were diluted in the same way for the reverse and forward transfection of gRNA complexes.

Reverse transfection of gRNA complex into sorted and unsorted cells

For reverse transfection of the gRNA complex, 30 nM gRNA complex for the chr27:36198117 target variant was incubated in the well of a 96-well plate with 0.75 µL

RNAiMAX, in a final volume of 50 µL Opti-MEM for 20 minutes. Following the 20 minute incubation, 100 µL of the diluted sorted (2:1 and 1:1 PX459:pMAXGFP) and unsorted cells (2:1 and 1:1 PX459:pMAXGFP) were added to the transfection mix, and incubated at 37°C for 48 hours.

Forward transfection of gRNA complex into sorted and unsorted cells

For forward transfection of the gRNA complex, 100 µL of the diluted sorted (2:1 and 1:1 PX459:pMAXGFP) and unsorted cells were added to the wells of a 96-well plate and incubated at 37°C, 5% CO_2 overnight. The next day, 30 nM gRNA complex for the chr27:36198117 target variant was incubated with 0.75 µL RNAiMAX in a final volume of 50 µL Opti-MEM for 20 minutes. During this incubation, each well was washed using 100 µL pre-warmed 1xPBS, and replaced with 100 µL antibiotic-free proliferation media. The transfection mix was then added to cells and incubated at 37°C, 5% CO_2 for 48 hours.

For both the forward and reverse transfected cells, media was replaced 24 hours post-transfection with full proliferation media. Following the 48 hour incubation, DNA extraction and genome editing detection was carried out as described in 3.4.3.

Optimisation of Cas9 protein-mediated CRISPR-Cas9 genome editing

Experiment 1: Optimisation of Cas9 RNP concentration

The IDT CRISPR-Cas9 protocol recommends 3 nM – 30 nM final concentration of RNP in conjunction with 0.5 µL – 2 µL Lipofectamine® RNAiMAX. The first round of optimisation experiments were conducted using these recommendations as a guide, using 0.5 µL, 1 µL and 2 µL Lipofectamine® RNAiMAX and 3 nM, 10 nM and 20 nM RNP complex for the chr27:36198117T>TGGC target. The second round of optimisation experiments were conducted using 2 µL Lipofectamine® RNAiMAX and 10 nM, 20 nM, 30 nM, 40 nM and 80 nM RNP complex for the chr27:36198117 target. Transfections were conducted using a forward transfection of these RNP complexes.

Immediately prior to transfection, the above RNP complexes were formed by incubation, and mixed with Lipofectamine® RNAiMAX in a final volume of 50 µL. During complex preparation, each cell colony was washed with 100 µL pre-warmed PBS and replaced with 100 µL of antibiotic-free proliferation media. Following the 20 minute incubation, 50 µL transfection complexes were added to each well and incubated at 37°C, 5% CO_2.

Experiment 2: Testing efficiency of NHEJ and HDR editing for gRNAs of target variants

To investigate the efficiency of NHEJ editing at each of the target loci in Table 3.1, transfections of RNP complexes were completed using a forward transfection, 2 µL Lipofectamine® RNAiMAX, and 20 nM RNP complex. Additional transfections were completed to also include 10 nM of HDR template in these incubations. The sequences for the HDR templates for the targets in Table 3.1 are presented in Appendix I.

Immediately prior to transfection, RNP complexes were formed by incubation as described in 3.4.5.1. For those transfections conducted with the addition of HDR template, 15 minutes into the 20 minute incubation in Lipofectamine® RNAiMAX, 10 nM ssODN was added to the transfection mix. Each colony was washed and transfected as described for Cas9 RNP Experiment 1 (3.4.5.1).

Experiment 3: Optimisation of HDR template concentration

To investigate the optimal ssODN concentration that would result in the highest rate of HDR at the chr5:93946027T>A target locus, the amount of template was titrated from 0 nM to 20nM. Transfections were conducted using a forward transfection, 2 µL Lipofectamine® RNAiMAX, and 20 nM RNP complex.

Immediately prior to transfection, RNP complexes were formed by incubation as described in 3.4.5.1. After 15 minutes of the 20 minute incubation in Lipofectamine® RNAiMAX, 3.33 nM, 10 nM and 20 nM ssODN were added to the transfection mix. Each colony was washed and transfected as described for Experiment 1 (3.4.5.1).

Experiment 4: Targeting more than one locus in a single transfection

To investigate if more than one locus could be targeted in a single transfection, gRNAs with targets in each of the *AGPAT6*, *MGST1*, *DGAT1* and *LGB* genes were multiplexed together. Transfections were conducted containing 5 nM, 10 nM and 20 nM of two different RNP multiplexes, comprising complexes of the chr27:36198117, chr5:93946027, chr14:1802265 and chr11:103301781 loci, and chr27:36212352, chr5:93945738, chr14:1802265 and chr11:103301781 loci. These multiplexes resulted in total RNP concentrations of 20 nM, 40 nM and 80 nM as the sum of the four locus-specific-complexes

Immediately prior to transfection, RNP complexes were formed by incubation as described in 3.4.5.1. Similarly, each colony was washed and transfected as described for Experiment 1 (3.4.5.1).

Experiment 5: GFP-enrichment of RNP-transfected cells

To investigate if CRISPR-Cas9 RNP complexes could be co-transfected with a selectable marker to enrich for edited cells within the cell population, transfections were conducted to include the pMAXGFP plasmid. For these experiments, transfections targeted three variants (chr27:36198117 T>TGGC, chr27:36211257 GA>T and chr27:36212352G>A), in conjunction with both forms of their corresponding ssODN templates. Cells were plated at 2 x 10^5 cells per well of a 12-well plate. The following day, when the cells were ~70% confluent, the RNP complexes were formed by incubation, and mixed with Lipofectamine® RNAiMAX in a final volume of 250 µL. After 15 minutes of the 20 minute incubation in Lipofectamine® RNAiMAX, 1.5 nM ssODN and 1500 ng pMAXGFP were added to the transfection mix. During this incubation, each cell colony was washed with 1 mL PBS and replaced with 1 mL of antibiotic-free proliferation media. Following the 20 minute incubation, 250 µL transfection complexes were added to each well and incubated at 37^{o}C, 5% CO_2 for 24 hours. Then, the media was changed to full proliferation media and the cells were visualised on the Nikon Ti-E inverted light microscope.

Appendix II

1. Mammary *DGAT1* expression association statistics for top WGS-derived variants not included in Table 5.12
2. Mammary *DGAT1* expression association statistics for top WGS-derived variants conditioned on *DGAT1* K232A not included in Table 5.13

Table 1. Mammary *DGAT1* expression association statistics for top WGS-derived variants not included in Table 5.12

Variant	Chr 14 pos	Parameter Est	P-value
rs208091850	1722033	0.1961(±0.0164)	1.29x10^{-27}
rs207577324	1725282	0.1961(±0.0164)	1.29x10^{-27}
rs137202508	1725536	0.1961(±0.0164)	1.29x10^{-27}
rs470414367	1728355	0.1961(±0.0164)	1.29x10^{-27}
rs133100921	1728858	0.1961(±0.0164)	1.29x10^{-27}
rs133257289	1729977	0.1961(±0.0164)	1.29x10^{-27}
rs378872350	1732068	0.1961(±0.0164)	1.29x10^{-27}
rs110577193	1735582	0.1961(±0.0164)	1.29x10^{-27}
rs133534450	1735779	0.1961(±0.0164)	1.29x10^{-27}
rs137491588	1735896	0.1961(±0.0164)	1.29x10^{-27}
rs134432442	1736599	0.1961(±0.0164)	1.29x10^{-27}
rs210909150	1738945	0.1961(±0.0164)	1.29x10^{-27}
rs378415895	1739675	0.1961(±0.0164)	1.29x10^{-27}
rs208827625	1739677	0.1961(±0.0164)	1.29x10^{-27}
rs135442643	1739725	0.1961(±0.0164)	1.29x10^{-27}
rs211605023	1741516	0.1961(±0.0164)	1.29x10^{-27}
rs133739752	1741650	0.1961(±0.0164)	1.29x10^{-27}
rs134533294	1741900	0.1961(±0.0164)	1.29x10^{-27}
rs209542297	1742529	0.1961(±0.0164)	1.29x10^{-27}
rs109050667	1745016	0.1961(±0.0164)	1.29x10^{-27}
rs135551752	1745431	0.1961(±0.0164)	1.29x10^{-27}
rs208657440	1745504	0.1961(±0.0164)	1.29x10^{-27}
rs210230767	1745777	0.1961(±0.0164)	1.29x10^{-27}
rs207855353	1746284	0.1961(±0.0164)	1.29x10^{-27}
rs209925040	1746291	0.1961(±0.0164)	1.29x10^{-27}
rs207795387	1750107	0.1961(±0.0164)	1.29x10^{-27}
rs135576599	1750824	0.1961(±0.0164)	1.29x10^{-27}
rs209058440	1752281	0.1961(±0.0164)	1.29x10^{-27}
rs134187064	1754238	0.1961(±0.0164)	1.29x10^{-27}
rs384226556	1755742	0.1961(±0.0164)	1.29x10^{-27}
rs210324747	1755898	0.1961(±0.0164)	1.29x10^{-27}
rs208417762	1756075	0.1969(±0.0166)	2.38x10^{-27}
rs209745231	1757801	0.1969(±0.0166)	2.38x10^{-27}
rs137617619	1759054	0.1969(±0.0166)	2.38x10^{-27}
rs133338261	1759353	0.1969(±0.0166)	2.38x10^{-27}
rs211560120	1759592	0.1969(±0.0166)	2.38x10^{-27}
rs208857025	1759620	0.1969(±0.0166)	2.38x10^{-27}
rs211021755	1759667	0.1969(±0.0166)	2.38x10^{-27}
rs207512673	1760544	0.1969(±0.0166)	2.38x10^{-27}
rs383052767	1760852	0.1969(±0.0166)	2.38x10^{-27}
rs136089752	1762331	0.1969(±0.0166)	2.38x10^{-27}
rs133788084	1763138	0.1969(±0.0166)	2.38x10^{-27}
rs135017891	1763380	0.1969(±0.0166)	2.38x10^{-27}
rs137071126	1765835	0.1969(±0.0166)	2.38x10^{-27}
rs208030891	1767385	0.1969(±0.0166)	2.38x10^{-27}
rs209595231	1769367	0.1969(±0.0166)	2.38x10^{-27}
rs211403999	1772560	0.1969(±0.0166)	2.38x10^{-27}

rs209907620	1779083	0.1969(±0.0166)	2.38x10^{-27}
rs133269088	1784505	0.1969(±0.0166)	2.38x10^{-27}
rs383392423	1787580	0.1969(±0.0166)	2.38x10^{-27}
rs384957047	1793616	0.1969(±0.0166)	2.38x10^{-27}
rs109162116	1804647	0.1969(±0.0166)	2.38x10^{-27}
rs211282745	1805963	0.1969(±0.0166)	2.38x10^{-27}
rs135258919	1808145	0.1969(±0.0166)	2.38x10^{-27}
rs135805021	1817975	0.1969(±0.0166)	2.38x10^{-27}
rs383356863	1818125	0.1969(±0.0166)	2.38x10^{-27}
rs208211113	1819475	0.1969(±0.0166)	2.38x10^{-27}
rs208113678	1825125	0.1969(±0.0166)	2.38x10^{-27}
rs135458711	1724688	0.1906(±0.0167)	1.10x10^{-25}
rs135433386	1726659	0.1906(±0.0167)	1.10x10^{-25}
rs136630297	1728264	0.1906(±0.0167)	1.10x10^{-25}
rs133014040	1735969	0.1906(±0.0167)	1.10x10^{-25}
rs136307654	1737473	0.1906(±0.0167)	1.10x10^{-25}
rs110825388	1739885	0.1906(±0.0167)	1.10x10^{-25}
rs135443540	1754287	0.1906(±0.0167)	1.10x10^{-25}
rs136875432	1762435	0.1919(±0.0169)	1.59x10^{-25}
rs448296125	1762504	0.1919(±0.0169)	1.59x10^{-25}
rs134364612	1765055	0.1919(±0.0169)	1.59x10^{-25}
rs135423283	1773053	0.1919(±0.0169)	1.59x10^{-25}
rs110982468	1775397	0.1919(±0.0169)	1.59x10^{-25}
rs132699547	1783521	0.1919(±0.0169)	1.59x10^{-25}
rs208317364	1800399	0.1919(±0.0169)	1.59x10^{-25}
rs209876151	1800439	0.1919(±0.0169)	1.59x10^{-25}
rs109421300	1801116	0.1919(±0.0169)	1.59x10^{-25}
rs109234250	1802265	0.1919(±0.0169)	1.59x10^{-25}
rs109326954	1802266	0.1919(±0.0169)	1.59x10^{-25}
rs135718911	1807139	0.1919(±0.0169)	1.59x10^{-25}
rs136783505	1807140	0.1919(±0.0169)	1.59x10^{-25}
rs133931291	1810124	0.1919(±0.0169)	1.59x10^{-25}
rs17870386	1810779	0.1919(±0.0169)	1.59x10^{-25}
rs209421707	1815678	0.1919(±0.0169)	1.59x10^{-25}
rs378247861	1816568	0.1919(±0.0169)	1.59x10^{-25}
rs136559790	1819667	0.1919(±0.0169)	1.59x10^{-25}
rs137672484	1822293	0.1919(±0.0169)	1.59x10^{-25}
rs133921340	1823757	0.1919(±0.0169)	1.59x10^{-25}

Table 2. Mammary *DGAT1* expression association statistics for top WG-derived variants conditioned on DGAT1 K232A not included in Table 5.13

Variant	Chr14 pos	Parameter Est	P-value
rs472613236	1721117	0.1013(±0.0242)	3.58x10^{-05}
rs434305476	1723909	0.1013(±0.0242)	3.58x10^{-05}
rs437352354	1728797	0.1013(±0.0242)	3.58x10^{-05}
rs479549292	1730155	0.1013(±0.0242)	3.58x10^{-05}
rs480237342	1731184	0.1013(±0.0242)	3.58x10^{-05}
rs463280458	1750500	0.1013(±0.0242)	3.58x10^{-05}
rs449834076	1752623	0.1013(±0.0242)	3.58x10^{-05}
rs479061541	1755801	0.1013(±0.0242)	3.58x10^{-05}
rs452083184	1758299	0.1013(±0.0242)	3.58x10^{-05}
rs477220846	1759997	0.1013(±0.0242)	3.58x10^{-05}
rs377930443	1760411	0.1013(±0.0242)	3.58x10^{-05}
rs432569830	1760697	0.1013(±0.0242)	3.58x10^{-05}
rs482439868	1778715	0.1013(±0.0242)	3.58x10^{-05}
rs464747584	1779138	0.1013(±0.0242)	3.58x10^{-05}
rs465770218	1797137	0.1013(±0.0242)	3.58x10^{-05}
rs460540024	1797980	0.1013(±0.0242)	3.58x10^{-05}
rs477342233	1799190	0.1013(±0.0242)	3.58x10^{-05}
rs382076865	1799567	0.1013(±0.0242)	3.58x10^{-05}
rs470788835	1807459	0.1013(±0.0242)	3.58x10^{-05}
rs455181695	1812094	0.1013(±0.0242)	3.58x10^{-05}
rs378952641	1820110	0.1013(±0.0242)	3.58x10^{-05}
rs439963934	1820256	0.1013(±0.0242)	3.58x10^{-05}
rs478390748	1831055	0.1013(±0.0242)	3.58x10^{-05}
rs456163226	1832315	0.1013(±0.0242)	3.58x10^{-05}
rs461148966	1839075	0.1013(±0.0242)	3.58x10^{-05}
rs472155081	1842827	0.1013(±0.0242)	3.58x10^{-05}
rs469678631	1845938	0.1013(±0.0242)	3.58x10^{-05}
rs454305447	1847561	0.1013(±0.0242)	3.58x10^{-05}
rs476736066	1848954	0.1013(±0.0242)	3.58x10^{-05}
rs379497765	1855915	0.1013(±0.0242)	3.58x10^{-05}
rs437144332	1856154	0.1013(±0.0242)	3.58x10^{-05}
rs384066610	1870287	0.1013(±0.0242)	3.58x10^{-05}
rs385302400	1874534	0.1013(±0.0242)	3.58x10^{-05}
rs384076576	1875806	0.1013(±0.0242)	3.58x10^{-05}
rs133347173	1882765	0.1013(±0.0242)	3.58x10^{-05}
rs476469426	1885986	0.1013(±0.0242)	3.58x10^{-05}
rs380971435	1887734	0.1013(±0.0242)	3.58x10^{-05}
rs481946110	1888221	0.1013(±0.0242)	3.58x10^{-05}
rs439241480	1891651	0.1013(±0.0242)	3.58x10^{-05}
rs383105805	1428907	0.1002(±0.0240)	3.88x10^{-05}
rs382180412	1429654	0.1002(±0.0240)	3.88x10^{-05}
rs379367933	1430529	0.1002(±0.0240)	3.88x10^{-05}
rs382233810	1432831	0.1002(±0.0240)	3.88x10^{-05}
rs385805021	1433601	0.1002(±0.0240)	3.88x10^{-05}
rs381356840	1434812	0.1002(±0.0240)	3.88x10^{-05}
rs383838904	1435823	0.1002(±0.0240)	3.88x10^{-05}
rs384691198	1436461	0.1002(±0.0240)	3.88x10^{-05}
rs381585464	1437537	0.1002(±0.0240)	3.88x10^{-05}
rs137587412	1438890	0.0975(±0.0243)	7.08x10^{-05}
rs383530306	1442055	0.0975(±0.0243)	7.08x10^{-05}
rs133283507	1443060	0.0975(±0.0243)	7.08x10^{-05}
rs132806157	1444046	0.0975(±0.0243)	7.08x10^{-05}
rs134459514	1445273	0.0975(±0.0243)	7.08x10^{-05}

Appendix III

FJXB F2 dam *AGPAT6* VNTR genotypes derived from GeneScan

Animal Key	Genotype
18158358	T/TGGC
18015888	T/T
18015882	T/TGGC
18203406	T/TGGC
18124266	T/T
18094350	T/TGGC
18012549	T/TGGC
18142746	N/A
18011233	T/TGGC
17584448	T/TGGC
18071394	TGGC/TGGC
17332581	T/TGGC
18032909	T/T
18143581	T/T
17195480	T/TGGC
18143570	N/A
18187042	T/T
18240060	T/TGGC
18143566	TGGC/TGGC
18240063	T/TGGC
17144784	T/TGGC
18143602	T/T
17150670	T/TGGC
17141955	T/T
18121607	T/TGGC
17195481	T/TGGC
17150699	T/TGGC
17118428	T/T
17108813	TGGC/TGGC
18186567	TGGC/TGGC
17195473	T/T
18375816	T/T
18240061	TGGC/TGGC
18186526	T/T
18187039	T/TGGC
18061259	T/TGGC
18083409	T/T
17145306	T/TGGC
18056917	TGGC/TGGC
17108825	T/TGGC
18049123	TGGC/TGGC
18108413	TGGC/TGGC
18102192	TGGC/TGGC
17192655	T/T
18143613	TGGC/TGGC
18143568	T/T
18240062	TGGC/TGGC
18240064	T/T
18083301	T/T
18158366	T/T
18186508	TGGC/TGGC
18094314	TGGC/TGGC
18271905	T/TGGC
18158369	T/TGGC
18015886	T/T
18068926	T/TGGC
18187041	T/TGGC
18094309	T/TGGC
18029080	T/TGGC
17584462	T/TGGC
18057002	T/TGGC
17090215	T/TGGC
17192636	T/TGGC
17141004	T/TGGC
17090216	T/T
17106126	T/TGGC
17250849	T/TGGC
18143603	T/TGGC
17195466	T/TGGC
18083410	T/TGGC
18216376	T/TGGC
17150702	T/TGGC
17141947	T/TGGC
17141971	T/T
18071391	T/T
18101632	T/TGGC
18015890	T/TGGC
17078588	T/TGGC
18057047	T/TGGC
18158356	T/T
18041337	T/T
18203409	T/TGGC
17225533	T/T
17929749	T/TGGC
17118425	T/TGGC
18094400	T/TGGC
18186502	T/TGGC
17181746	T/TGGC
17181747	T/T
18186500	T/TGGC
17108957	T/TGGC
17074192	T/T
17141907	TGGC/TGGC
17078582	TGGC/TGGC
18143572	TGGC/TGGC
18318000	T/T
17074197	T/TGGC
16989640	T/TGGC
17144783	T/TGGC
18143621	TGGC/TGGC
17077380	TGGC/TGGC
17077388	T/TGGC
17141913	TGGC/TGGC
18158365	T/TGGC
17078580	T/TGGC
17145716	T/TGGC
17192630	TGGC/TGGC
17396939	TGGC/TGGC
18158367	T/TGGC
18083364	T/TGGC
18143635	TGGC/TGGC
18083365	T/T
18216388	T/TGGC

18094305	T/TGGC
18102189	T/TGGC
18033109	T/TGGC
18186561	T/TGGC
18240374	T/TGGC
18143574	T/TGGC
17150700	T/TGGC
18375815	T/TGGC
18094310	T/TGGC
17107841	T/T
17074198	TGGC/TGGC
18032897	T/TGGC
17141915	T/T
18186463	T/T
18186999	T/TGGC
17192619	TGGC/TGGC
17309016	T/T
17029694	T/TGGC
17194537	T/TGGC
17108826	TGGC/TGGC
18143622	T/T
17195482	TGGC/TGGC
17331676	T/TGGC
18240367	T/TGGC
18093578	T/TGGC
17150685	T/TGGC
17118416	TGGC/TGGC
18143615	T/TGGC
18143611	TGGC/TGGC
18093560	TGGC/TGGC
18093564	T/TGGC
18187040	TGGC/TGGC
18143569	T/T
18094308	T/T
18095384	T/TGGC
18083080	T/TGGC
17107862	T/TGGC
18057124	T/T
17192618	TGGC/TGGC
17192617	TGGC/TGGC
17228393	TGGC/TGGC
17192616	T/TGGC
17078585	TGGC/TGGC
17192629	T/TGGC
17520833	TGGC/TGGC
17107839	T/T
17396957	TGGC/TGGC
18045801	T/TGGC
18045798	T/TGGC
18071405	T/TGGC
18094313	T/T
17195471	T/TGGC
17107842	T/TGGC
18240371	T/TGGC
18083085	T/TGGC
18083289	T/TGGC
18083291	TGGC/TGGC
18318002	T/T
18090774	T/TGGC
17994433	T/TGGC
17994437	T/T
18071917	TGGC/TGGC
18046083	N/A
18268430	T/TGGC
18144223	TGGC/TGGC
18268426	TGGC/TGGC
18268427	T/T
18268425	T/TGGC
18071393	T/TGGC
18143601	TGGC/TGGC
18144224	T/TGGC
18071392	T/TGGC
17068023	T/TGGC
18318004	T/T
18136034	T/TGGC
17213113	T/TGGC
17141927	TGGC/TGGC
17213110	TGGC/TGGC
17150671	TGGC/TGGC
17181745	TGGC/TGGC
18186510	T/TGGC
17074199	T/T
18045768	T/TGGC
17831766	T/T
17018782	T/TGGC
18178737	TGGC/TGGC
17095184	T/TGGC
18071431	T/T
18075261	T/TGGC
18143617	T/TGGC
18083408	TGGC/TGGC
18143618	T/T
18098425	T/TGGC
18335738	T/TGGC
18143616	T/T
18182220	T/TGGC
18291371	T/TGGC
18271903	TGGC/TGGC
18158336	T/TGGC
18045782	T/TGGC
18288188	T/T
18094348	T/TGGC
18071426	T/TGGC
18158359	T/TGGC
18240059	T/TGGC
18094357	T/TGGC
17092083	T/TGGC
17195477	T/TGGC
17250851	T/TGGC
17192624	T/TGGC
17192622	T/TGGC
18029138	T/TGGC
17069099	T/TGGC
17088768	TGGC/TGGC
18057120	TGGC/TGGC
17250863	TGGC/TGGC
17141969	TGGC/TGGC
18032894	N/A
17222800	T/TGGC
18102207	T/TGGC
18158331	N/A
17078579	TGGC/TGGC
18083565	TGGC/TGGC
17584432	TGGC/TGGC
17141968	TGGC/TGGC
18186991	T/TGGC

17368530	TGGC/TGGC
18288198	T/TGGC
18094353	T/TGGC
17144781	TGGC/TGGC
17228507	T/TGGC
17228481	T/TGGC
17228484	N/A
17018051	T/TGGC
17090209	TGGC/TGGC
17396963	T/TGGC
17150677	T/TGGC
18033091	T/TGGC
18095140	TGGC/TGGC
17228457	T/TGGC
17195474	TGGC/TGGC
17074196	TGGC/TGGC
17192634	TGGC/TGGC
17192653	T/TGGC
17978329	TGGC/TGGC
18045796	T/TGGC
18083575	T/TGGC
18071396	T/TGGC
18143600	T/TGGC
17228350	T/TGGC
17467500	TGGC/TGGC
17181737	T/TGGC
18186990	T/TGGC
17195472	TGGC/TGGC
18158347	T/TGGC
18158345	TGGC/TGGC
17082049	T/TGGC
17195465	TGGC/TGGC
17150703	TGGC/TGGC
17107838	TGGC/TGGC
17070016	TGGC/TGGC
18045795	T/TGGC
18012547	TGGC/TGGC
18158371	T/TGGC
18094352	T/TGGC
17966669	TGGC/TGGC
18032903	TGGC/TGGC
18032901	TGGC/TGGC

18046070	TGGC/TGGC
18046069	TGGC/TGGC
18046071	TGGC/TGGC
18083298	TGGC/TGGC
18083300	T/TGGC
17228511	T/TGGC
17141950	T/TGGC
17332588	TGGC/TGGC
17181735	T/TGGC
18035912	TGGC/TGGC
18056841	TGGC/TGGC
18052431	TGGC/TGGC
18094395	T/TGGC
17090002	T/TGGC
17917556	TGGC/TGGC
18186568	T/TGGC
17917567	T/TGGC
18016971	T/TGGC
18102212	T/TGGC
18045789	TGGC/TGGC
18064960	TGGC/TGGC
18071408	T/TGGC
18071407	T/TGGC
17146001	TGGC/TGGC
17228342	TGGC/TGGC
18186992	TGGC/TGGC
17181736	TGGC/TGGC
17107844	TGGC/TGGC
18143577	T/TGGC
17250864	T/TGGC
17396954	TGGC/TGGC
17396964	TGGC/TGGC
18052353	TGGC/TGGC
17250871	T/TGGC
18143607	T/TGGC
17068027	T/TGGC
18046078	TGGC/TGGC
18046080	T/TGGC
18015885	T/TGGC
17966662	T/TGGC
17973340	T/TGGC
18187010	T/TGGC

17396965	T/TGGC
17982032	T/TGGC
18032899	T/TGGC
18015887	TGGC/TGGC
18057119	TGGC/TGGC
18032895	TGGC/TGGC
18083389	TGGC/TGGC
17181734	TGGC/TGGC
17090004	TGGC/TGGC
18033093	T/TGGC
17074195	T/TGGC
17090210	TGGC/TGGC
17090223	TGGC/TGGC
18083387	TGGC/TGGC
18216379	T/TGGC
18158344	TGGC/TGGC
18158346	TGGC/TGGC
17396933	N/A
17584456	TGGC/TGGC
17396931	T/TGGC
18083335	T/TGGC
18011234	T/TGGC
18029140	T/TGGC
18039526	TGGC/TGGC
18033094	T/TGGC
18124278	TGGC/TGGC
18124277	TGGC/TGGC
18056916	T/TGGC
18029083	N/A
17990190	TGGC/TGGC
18029079	TGGC/TGGC
18029081	TGGC/TGGC
18029082	T/TGGC
18029077	TGGC/TGGC
18124272	T/TGGC
18021792	T/TGGC
18065207	TGGC/TGGC
18071395	TGGC/TGGC
18056838	TGGC/TGGC
18288196	TGGC/TGGC
18288779	T/TGGC
18032904	N/A

18024861	T/TGGC
18104820	T/TGGC
17092078	T/T
17917553	T/T
17092077	T/TGGC
17396956	T/TGGC
17150693	T/TGGC
17141948	TGGC/TGGC
17141961	T/TGGC
17195479	T/TGGC
18033088	T/TGGC
17584463	T/TGGC
17090212	T/T
18093592	T/T
18158334	TGGC/TGGC
18154050	T/TGGC
18057045	TGGC/TGGC
17185950	T/TGGC
17226492	T/T
17118407	TGGC/TGGC
18071422	T/TGGC
17118418	TGGC/TGGC
18094345	T/T
18071423	T/TGGC
17228340	T/T
17141926	T/TGGC
17250859	TGGC/TGGC
18071411	T/T
18015880	T/T
17141930	T/TGGC
17288675	T/TGGC
18171923	T/T
17250854	T/TGGC
17195475	T/T
17150681	T/T
17141936	T/T
18124569	T/TGGC
17250824	T/TGGC
18057046	TGGC/TGGC
17090207	TGGC/TGGC
17226491	TGGC/TGGC
18045794	T/TGGC
17114949	T/T
18186495	T/TGGC
18057121	T/TGGC
18029144	TGGC/TGGC
18071390	T/TGGC
17003813	T/TGGC
17192652	T/T
18094344	T/T
17090206	T/TGGC
17150696	T/TGGC
17331527	T/TGGC
18019875	T/T
17192635	TGGC/TGGC
17090221	T/TGGC
17067121	T/TGGC
17144798	T/TGGC
17092086	TGGC/TGGC
17074200	T/TGGC
17092074	TGGC/TGGC
18158362	T/TGGC
17396944	T/T
17090211	T/T
17108820	T/TGGC
17141967	TGGC/TGGC
17141965	T/T
17181739	T/TGGC
18071409	T/T
18071410	T/T
18158333	TGGC/TGGC
17945249	T/TGGC
18032898	T/TGGC
17962627	T/TGGC
17107858	TGGC/TGGC
17118406	T/T
17288668	T/TGGC
18032891	T/T
17067122	TGGC/TGGC
18124570	TGGC/TGGC
18143590	T/TGGC
18158329	T/TGGC
18030299	TGGC/TGGC
17090217	TGGC/TGGC
17090218	T/TGGC
17228333	TGGC/TGGC
17078594	TGGC/TGGC
17078595	T/TGGC
17108816	TGGC/TGGC
17083184	TGGC/TGGC
17228341	T/TGGC
17141934	T/T
17192647	T/T
17141932	T/TGGC
17192646	T/TGGC
17141933	T/TGGC
17092076	T/TGGC
18158354	N/A
17195484	T/T
18059608	T/TGGC
17195483	T/TGGC
17941505	TGGC/TGGC
17150675	T/T
18186462	TGGC/TGGC
17141962	TGGC/TGGC
17108822	T/TGGC
17108818	T/TGGC
17070513	T/T
17070511	TGGC/TGGC
17090228	T/TGGC
17192648	T/T
17141920	TGGC/TGGC
17192651	T/TGGC
17150680	T/T
18187003	T/TGGC
17962584	TGGC/TGGC
18186997	T/TGGC
18186498	TGGC/TGGC
18094408	T/T
18135754	T/TGGC
18071424	T/TGGC
18102221	T/T
18102216	T/TGGC
18151431	TGGC/TGGC
18035897	T/TGGC
17092082	T/TGGC

17141935	TGGC/TGGC
17090225	T/TGGC
17141924	T/TGGC
17118421	T/TGGC
17170495	T/TGGC
17150698	T/T
18187001	T/TGGC
17250855	TGGC/TGGC
17192639	T/TGGC
18187000	TGGC/TGGC
18062704	T/TGGC
18032892	T/TGGC
18187046	TGGC/TGGC
18151397	T/TGGC
18020405	T/TGGC
18158364	T/TGGC
18083074	T/TGGC
18151420	N/A
18015879	TGGC/TGGC
18094343	TGGC/TGGC
18057044	TGGC/TGGC
18216389	T/TGGC
17195485	T/T
17396960	T/T
17141917	T/T
17228353	T/T
17144786	T/T
17228357	T/T
17332586	T/T
17144789	T/T
17228358	T/T
17144788	T/T
17584434	T/T
17228478	T/T
17396959	T/T
17073217	T/TGGC
17228392	T/T
17078587	T/TGGC
17038877	T/T
17288672	T/T
17288671	N/A
17181742	T/T

17144797	T/T
17107849	T/T
18015883	T/T
17994656	T/T
17118427	T/T
17118426	T/TGGC
17228479	T/T
17092079	T/TGGC
17150705	T/T
18094347	T/TGGC
18077788	T/T
17252013	T/T
17150687	T/TGGC
17192644	T/T
17038873	T/T
17037692	T/T
17069100	T/T
17332583	T/TGGC
17228356	T/T
17332587	T/TGGC
17107851	T/T
17144790	T/T
17853219	T/TGGC
17144791	T/T
17228391	T/TGGC
17181744	T/T
17181741	T/T
17181740	T/T
18172178	T/T
18172172	T/T
18018470	T/TGGC
17067621	T/TGGC
18187014	T/TGGC
17108814	T/T
17252021	T/TGGC
17250873	T/TGGC
17150674	T/T
17994655	T/T
18033049	T/T
18083566	N/A
17195486	T/TGGC
18057115	T/TGGC

17280015	T/T
18033092	T/TGGC
17078592	T/T
17108831	T/T
18186993	T/TGGC
17228348	T/T
17288673	T/TGGC
18083574	T/TGGC
18093571	T/T
17839010	T/TGGC
18040513	T/T
17107860	T/TGGC
18083313	T/T
18083332	T/TGGC
17141919	T/T
18083315	T/T
18083318	T/T
18143589	T/TGGC
17090219	T/TGGC
17053865	T/TGGC
17053866	T/TGGC
18033694	T/T
17107845	T/TGGC
18143576	T/T
17144782	T/T
17974612	T/T
17192654	T/T
18143583	T/T
18143585	T/T
18143584	T/TGGC
17144794	T/TGGC
17998551	T/TGGC
17998552	T/T
17137650	T/T
18187005	T/TGGC
18239813	T/TGGC
18124268	T/TGGC
17853334	N/A
17853356	T/TGGC
18124267	T/TGGC
18057116	T/TGGC
18240354	T/TGGC

18240056	T/T
18045778	T/TGGC
18029148	T/T
18172180	T/T
18158339	T/TGGC
18158338	T/T
18071398	T/TGGC
18158342	T/TGGC
18071399	T/T
18158340	T/T
18083076	T/T
18187043	N/A
18158341	T/TGGC
18071401	T/T
18192028	T/TGGC
18586748	T/TGGC
18083398	T/T
18089767	T/T
18089759	T/TGGC
18186569	T/T
18215578	T/TGGC
17150692	T/T
18186998	T/TGGC
17331679	T/T
18612074	T/T
18057043	T/T
18123833	T/T
17331677	T/T
18102193	T/T
17255313	N/A
17250866	T/TGGC
17250869	T/T
17250867	T/TGGC
17199739	T/TGGC
17277648	T/T
17396937	T/TGGC
18101636	T/TGGC
17090220	N/A
17141909	T/T
18186573	T/T
18124269	T/T
17396934	T/TGGC
17250870	T/TGGC
18158335	T/T
18280911	T/TGGC
18057127	T/T
17195487	T/T
17897284	T/TGGC
18407513	T/TGGC
17192672	T/TGGC
18407512	N/A
17250860	T/T
17332580	T/TGGC
17150672	T/TGGC
17331675	T/T
17009145	T/T
17331674	T/TGGC
17141952	T/T
18187006	T/TGGC
17118405	T/TGGC
18216380	T/TGGC
18056814	T/T
18216387	T/T
18093545	T/TGGC
18158361	T/TGGC
17181733	T/TGGC
18216382	T/TGGC
18240058	T/T
17250872	T/T
17250868	T/T
17250865	T/T
18071421	T/TGGC
18158368	T/TGGC
17118404	T/T
17192631	T/TGGC
18280914	T/T
18280913	T/T
18335742	T/T
17141970	T/TGGC
17074193	T/TGGC
17092084	T/T
18158330	T/T
17277638	T/T
17078591	T/TGGC
17031634	T/T
17141953	T/TGGC
17351044	T/TGGC
17396936	T/T
17195463	T/TGGC
17963562	T/TGGC
17228451	N/A
18035079	T/TGGC
18407505	T/TGGC
18186345	T/T
18143571	T/T
18318196	T/TGGC
17150684	T/TGGC
18187017	T/TGGC
16881434	T/T
17068673	T/TGGC
18240057	T/T
18187045	T/TGGC
18158328	T/TGGC
18186549	N/A
17192645	T/T
17195467	T/T
18216381	T/T
17195489	N/A
18216375	T/TGGC
18186995	T/T
18335749	T/T
17141949	T/TGGC
18187008	N/A
18121868	N/A
18121857	T/TGGC
18121887	T/T
18083079	T/T
18071420	T/T
18057117	N/A
18017228	T/T
18123769	T/T
18123813	T/TGGC
18143619	N/A
18143586	T/TGGC
18093567	T/TGGC
18047784	T/TGGC

18143564	T/TGGC
18586747	T/T
18318198	T/TGGC
18186343	T/TGGC
18187019	T/TGGC
10932752	T/T
18143595	TGGC/TGGC
18094346	T/TGGC
17110985	T/TGGC
17192627	T/TGGC
11767718	TGGC/TGGC
17078584	T/T
13868100	T/T
18094304	T/TGGC
15541340	T/T
13022886	T/T
18029141	TGGC/TGGC
17396942	TGGC/TGGC
12151017	N/A
12199110	T/TGGC
11748989	TGGC/TGGC
17991265	T/TGGC
18124265	T/T
17090214	T/T
17195478	TGGC/TGGC
12791606	N/A
13974179	T/TGGC
14430781	T/TGGC
10204606	N/A
15660220	T/T
15683229	N/A
17962601	TGGC/TGGC
17141002	TGGC/TGGC
12698568	TGGC/TGGC
17150683	T/TGGC
14732918	T/T
15028885	T/TGGC
11690029	N/A
18041655	T/T
17141923	TGGC/TGGC
12688205	T/TGGC
17181732	TGGC/TGGC
14627888	T/TGGC
18143580	T/T
18073190	T/TGGC
12736076	TGGC/TGGC
13676200	T/TGGC
17274373	T/T
12958676	T/TGGC
17078586	N/A
15845313	TGGC/TGGC
18094307	N/A
17288670	T/T
17192650	TGGC/TGGC
17107859	T/TGGC
17212358	T/TGGC
17067120	T/TGGC
17228380	TGGC/TGGC
17228379	T/TGGC
17396932	N/A
18187013	T/TGGC
16971539	T/TGGC
18057157	T/TGGC
18094419	N/A
17584433	T/T
17078583	T/T
18017292	T/TGGC
18045783	T/TGGC
18158337	T/T
18094422	T/T
17154523	TGGC/TGGC
17107861	TGGC/TGGC
18216377	N/A
17118410	T/TGGC
18218067	T/T
18280912	T/TGGC
18158325	T/TGGC
18056857	T/T
18056843	T/TGGC
18143631	T/T
18083413	T/T
18143633	T/T
18143632	T/TGGC
17023806	N/A
17261496	N/A
18186501	T/T
18031034	T/TGGC
18240364	TGGC/TGGC
17141972	T/TGGC
18083391	T/T
18102507	T/TGGC
18263159	T/TGGC
18186348	T/TGGC
17908065	N/A
18240358	T/TGGC
18158360	T/TGGC
18094349	T/TGGC
18203378	TGGC/TGGC
7543119	T/TGGC
359064	N/A
7932279	N/A
10601498	N/A
12151323	T/TGGC
13560102	N/A
18158332	T/TGGC
17404420	T/TGGC
11107124	T/TGGC
7006926	N/A
7435283	N/A
15525685	N/A
14459809	N/A
15015704	N/A
15485396	T/TGGC
17015137	T/T
9086725	T/TGGC
8814639	T/T
8499654	T/T
16109405	T/TGGC
13216263	T/TGGC
12759259	N/A
15572172	T/TGGC
18034197	T/TGGC
9797431	T/TGGC
7154444	TGGC/TGGC
8405816	N/A
16120004	N/A

14771476	N/A
18083083	T/TGGC
18187038	T/T
13873891	N/A
9951726	T/TGGC
6823525	N/A
6232846	N/A
10253821	N/A

11749007	N/A
18083082	N/A
16989634	N/A
18083280	T/T
5553901	T/T
9099	T/TGGC
6795478	N/A
5869403	N/A

12882341	TGGC/TGGC
18186524	N/A
17228370	N/A
15325484	N/A

Appendix IV

Linkage disequilibrium statistics (R^2 values) between rs109815800 and DGAT1 K232A/BovineHD panel markers in chromosome 14 body weight locus

Marker	Position	R^2
rs109234250	1802265	0.063
rs109637592	24008839	0.141
rs109925810	24014579	0.465
rs110489692	24018803	0.162
rs136508017	24044381	0.328
rs42545204	24047418	0.164
rs41660107	24048952	0.328
rs109602517	24049812	0.328
rs109318512	24051093	0.328
rs109705035	24051987	0.150
rs109284285	24053137	0.328
rs110845339	24057354	0.328
rs42545192	24065280	0.163
rs133032517	24067610	0.212
rs109422239	24072137	0.212
rs110821373	24074220	0.212
rs42545182	24075714	0.162
rs110390285	24089516	0.070
rs109643003	24092123	0.212
rs136339222	24096532	0.132
rs42545165	24099094	0.162
rs134304712	24099719	0.132
rs137341544	24102024	0.132
rs109528593	24106396	0.212
rs110864751	24114365	0.433
rs42545145	24115422	0.054
rs42544420	24132456	0.321
rs42544424	24133627	0.217
rs42544430	24138878	0.217
rs109682353	24143265	0.321
rs41660101	24145838	0.321
rs110006364	24147525	0.433
rs42544418	24150127	0.217
rs42544401	24158787	0.219
rs42544357	24161697	0.050
rs42544377	24167861	0.217
rs42544383	24172479	0.323
rs42544386	24175600	0.324
rs42544392	24179150	0.207
rs42544395	24181858	0.324
rs42544396	24182406	0.207
rs42544400	24185058	0.324
rs42544356	24187772	0.324
rs42544349	24191959	0.324
rs42544348	24193383	0.324
rs134319614	24206232	0.023
rs109890494	24219041	0.435
rs42544336	24220070	0.000
rs110993288	24221657	0.435
rs109174538	24222338	0.434
rs109976467	24225369	0.435
rs108982003	24226206	0.434
rs110482368	24227327	0.434
rs134751608	24229059	0.501
rs110010333	24235712	0.493
rs109645403	24237304	0.493
rs42543230	24243733	0.323
rs133885118	24258275	0.006
rs137434020	24260937	0.006
rs136017102	24263980	0.027
rs43004834	24266960	0.000
rs41581840	24275232	0.018
rs109963694	24276214	0.124
rs134518389	24278284	0.036
rs135639509	24280431	0.006
rs136703875	24281870	0.006
rs134312033	24285339	0.002
rs133296385	24291712	0.028
rs43003348	24304427	0.037
rs43002526	24321232	0.183
rs137785718	24323400	0.102
rs110104035	24324094	0.103
rs110383563	24326513	0.021
rs134955677	24329536	0.096
rs43003344	24330594	0.268
rs134193404	24333423	0.084
rs135748649	24335922	0.084
rs132711884	24336953	0.084
rs109341693	24348047	0.029
rs135409116	24359161	0.272
rs42648898	24361242	0.290
rs41615249	24363276	0.024
rs29020689	24365162	0.316
rs134293561	24369510	0.269
rs135459952	24373031	0.269
rs133288868	24376195	0.269
rs42648880	24378496	0.016
rs136831935	24384496	0.201
rs135164902	24385879	0.125
rs42648868	24391175	0.051
rs135413008	24395527	0.138
rs137782768	24396836	0.137
rs133719195	24400605	0.138
rs137364314	24404982	0.138
rs133138223	24406302	0.137
rs42648925	24411455	0.067
rs42649744	24412493	0.067
rs42649760	24413758	0.140
rs109119025	24418370	0.060
rs42649767	24419295	0.083
rs42649771	24420840	0.122
rs109185321	24425758	0.104
rs110340643	24429310	0.047
rs110717761	24431237	0.104

rs136321755	24434190	0.051
rs42649775	24437778	0.187
rs42649776	24440797	0.054
rs109425569	24445514	0.073
rs110395333	24448641	0.073
rs109828753	24453615	0.055
rs42649778	24455791	0.084
rs110727287	24459302	0.055
rs42649780	24466047	0.187
rs135958550	24470245	0.187
rs110543321	24471148	0.187
rs42646633	24472819	0.111
rs42646635	24473841	0.156
rs42646636	24475213	0.156
rs42646638	24476256	0.156
rs132820259	24478336	0.156
rs42646648	24487011	0.352
rs42646660	24524205	0.463
rs132924262	24532336	0.467
rs42646685	24536549	0.314
rs42646677	24545053	0.314
rs135646716	24553162	0.464
rs42646691	24556301	0.359
rs42646700	24562756	0.313
rs42646702	24563237	0.352
rs42646708	24573257	0.466
rs42646720	24590812	0.479
rs134188138	24595318	0.031
rs42646723	24598515	0.477
rs41724398	24621142	0.667
rs133020056	24633076	0.238
rs41723523	24639618	0.049
rs41724332	24643266	0.413
rs109080115	24656389	0.597
rs135008823	24699409	0.104
rs136704276	24706121	0.099
rs109748092	24710609	0.328
rs108959399	24716826	0.328
rs110451945	24718647	0.328
rs109055951	24720352	0.286
rs109151890	24724974	0.287
rs135573576	24729765	0.286
rs136860540	24736064	0.145
rs134726969	24737917	0.295
rs109028875	24758139	0.000
rs110334248	24761860	0.106
rs109490367	24763610	0.107
rs110321820	24765731	0.106
rs109783781	24769617	0.107
rs109003440	24772375	0.166
rs109172777	24779419	0.106
rs133640284	24783381	0.051
rs41627956	24787245	0.166
rs134998417	24797724	0.105
rs134309686	24805520	0.104
rs137425866	24808828	0.000
rs133467103	24824037	0.052
rs134623289	24828922	0.048
rs134435992	24864286	0.017
rs136166205	24869608	0.301
rs133846946	24874608	0.287
rs110367762	24892678	0.515
rs133921678	24894527	0.034
rs135626029	24897094	0.034
rs133142988	24900445	0.034
rs134735082	24902136	0.034
rs137627685	24906337	0.034
rs109116062	24909247	0.680
rs137044774	24911824	0.034
rs109110003	24913654	0.188
rs133409868	24915886	0.034
rs134649249	24922753	0.408
rs135388492	24931388	0.409
rs133480234	24933932	0.409
rs134848602	24939285	0.409
rs137780934	24941523	0.034
rs133714277	24944254	0.034
rs137200131	24952035	0.034
rs136544328	24956145	0.034
rs134174250	24958417	0.034
rs135735870	24961879	0.530
rs110243083	24973324	0.265
rs136888475	24975563	0.034
rs135318045	24980786	0.034
rs136631581	24989997	0.034
rs109636480	24998326	0.433
rs135404594	25001051	0.034
rs134286310	25009960	0.035
rs135538206	25012733	0.034
rs109815800	25015640	1.000
rs137204453	25018843	0.034
rs135852767	25021594	0.034
rs136828442	25026174	0.034
rs133840388	25036693	0.034
rs135401045	25050448	0.034
rs136474498	25054377	0.206
rs133704402	25056208	0.206
rs137401639	25058053	0.552
rs136288032	25060055	0.034
rs136982189	25066322	0.438
rs135114806	25069487	0.034
Chr14_25075542	25075542	0.034
rs134048394	25079291	0.034
rs135256588	25082358	0.034
rs136629079	25092241	0.034
rs41722894	25098364	0.145
rs137557469	25102663	0.034
rs133531622	25105265	0.034
rs41627954	25107556	0.117
rs41722912	25111082	0.217
rs41722915	25114769	0.248
rs41722918	25117469	0.214
rs134949970	25119622	0.034
rs41722865	25123059	0.242
rs136826029	25129005	0.388
rs41722872	25134787	0.173
rs135269914	25147967	0.379
rs43151427	25154132	0.062
rs43151429	25160597	0.015
rs135211309	25164603	0.001
rs43157016	25173600	0.015
rs43157018	25174741	0.151

rs43157020	25175950	0.015
rs43157023	25180411	0.126
rs133606272	25185357	0.015
rs41720594	25199512	0.015
rs41720598	25203669	0.002
rs137338872	25204467	0.092
rs41720621	25215027	0.003
rs41720613	25215969	0.003
rs136157418	25221694	0.142
rs41720630	25225097	0.003
rs41721832	25226878	0.003
rs41721841	25228905	0.410
rs41722127	25235472	0.090
rs132881564	25241366	0.169
rs41722847	25249177	0.000
rs134732749	25252182	0.217
rs41722855	25254540	0.090
rs41722039	25267800	0.077
rs136598772	25272140	0.083
rs41722053	25276491	0.345
rs41578094	25284162	0.345
rs137648164	25287012	0.123
rs133403697	25290225	0.074
rs134650233	25293271	0.459
rs135388393	25298972	0.486
rs41627953	25307116	0.626
rs136191791	25315687	0.453
rs41722103	25320421	0.244
rs42892600	25329035	0.621
rs42892592	25332510	0.359
rs42892582	25336906	0.001
rs42892571	25343470	0.001
rs42892565	25348919	0.019
rs42892557	25351733	0.001
rs109227633	25354674	0.019
rs137490412	25358895	0.034
rs132872540	25365895	0.034
rs135577401	25374602	0.034
rs133012258	25376827	0.034
rs134286113	25379505	0.034
rs135530224	25383331	0.034

rs133193436	25389001	0.034
rs134485671	25391747	0.000
rs110816518	25393163	0.368
rs109871644	25396031	0.063
rs109782798	25400645	0.200
rs41657755	25401722	0.371
rs110576056	25412013	0.004
rs110602867	25423298	0.000
rs41722033	25425357	0.000
rs133919865	25429707	0.034
rs109679763	25446793	0.000
rs137844185	25451730	0.034
rs110020650	25457504	0.000
rs41627950	25459674	0.163
rs41721289	25460952	0.163
rs41721294	25466175	0.163
rs134038032	25468481	0.034
rs137536818	25471170	0.047
rs135252751	25474456	0.034
rs109056763	25478810	0.071
rs109216574	25488048	0.406
rs110108793	25490226	0.574
rs110506327	25492467	0.702
rs110741347	25497146	0.575
rs41720428	25498881	0.702
rs132979341	25500235	0.702
rs110632518	25501417	0.126
rs136543212	25502915	0.702
rs41627948	25504073	0.574
rs137267491	25505663	0.574
rs41627946	25506575	0.574
rs136889989	25507730	0.574
rs135262614	25510859	0.574
rs133736127	25513599	0.574
rs41720387	25517111	0.032
rs41720383	25518123	0.032
rs41721322	25520749	0.577
rs109670294	25521888	0.576
rs132786957	25525225	0.577
rs41623108	25526683	0.577
rs136345290	25528516	0.702

rs133597776	25529645	0.574
rs134607819	25530203	0.574
rs136855191	25536019	0.574
rs134846474	25537252	0.574
rs137619218	25541189	0.574
rs133319071	25544079	0.407
rs134551820	25548323	0.068
rs136150430	25551817	0.034
rs133834218	25555197	0.034
rs134633991	25557034	0.034
rs42935363	25567411	0.126
rs135313923	25570551	0.034
rs109633597	25577162	0.084
rs109418171	25579888	0.126
rs43093130	25582834	0.007
rs134725247	25584875	0.025
rs136320412	25586296	0.021
rs109761795	25589517	0.324
rs110985019	25601303	0.329
rs109521494	25608094	0.324
rs29021334	25612510	0.232
rs29021333	25616884	0.237
rs42962539	25621782	0.007
rs108941421	25638580	0.012
rs42961225	25640190	0.069
rs42299113	25648989	0.069
rs137236027	25650993	0.034
rs109540593	25655658	0.015
rs134159539	25659050	0.034
rs42299126	25664934	0.006
rs135316058	25675568	0.051
rs136667611	25683113	0.034
rs134206288	25686207	0.034
rs110774011	25698286	0.030
rs133094347	25699163	0.030
rs134370861	25704807	0.030
rs135531050	25708285	0.030
rs42839864	25715320	0.004
rs42839872	25719951	0.001
rs134649406	25725057	0.033
rs42839876	25730129	0.000

rs42839873	25731992	0.001
rs133249705	25733426	0.001
rs42839886	25739308	0.001
rs109067020	25747953	0.000
rs137397008	25753013	0.033
rs43010099	25759873	0.001
rs43010094	25762795	0.002
rs43010082	25766228	0.016
rs43010073	25767656	0.002
rs43010065	25769988	0.002
rs41665905	25771436	0.002
rs135254559	25776037	0.040
rs43757985	25779560	0.003
rs137551408	25786908	0.033
rs133519399	25794261	0.032
rs43770985	25797331	0.016
rs135744414	25800191	0.033
rs133347432	25803886	0.033
rs134947467	25808557	0.033
rs43770972	25812326	0.001
rs133570825	25814803	0.033
rs43770969	25817300	0.150
rs137748068	25819872	0.033
rs133628406	25823040	0.033
rs137494880	25826189	0.033
rs135939284	25828312	0.033
rs42299100	25832112	0.048
rs133003803	25835618	0.033
rs136976295	25839257	0.033
rs135202659	25842735	0.033
rs42299083	25846511	0.008
rs134006862	25849150	0.000
rs42299080	25851646	0.075
rs41665281	25857110	0.213
rs134601995	25860105	0.000
rs136141080	25863924	0.000
rs133252286	25866853	0.000
rs137336582	25869266	0.039
rs135734725	25871315	0.076
rs136755107	25873843	0.000
rs134567839	25877586	0.003
rs136146069	25882396	0.000
rs41665280	25887784	0.039
rs133829973	25894793	0.000
rs134624153	25899663	0.080
rs42299038	25909536	0.091
rs41665273	25913294	0.093
rs137492335	25915281	0.033
rs132878276	25917492	0.033
rs42298501	25921382	0.032
rs42298505	25926527	0.041
rs133124974	25930713	0.066
rs134858624	25933870	0.032
rs136316897	25938065	0.044
rs42299032	25946432	0.086
rs110721536	25947476	0.084
rs110797611	25956248	0.005
rs42298481	25959516	0.406
rs110342609	25962036	0.242
rs42298477	25964134	0.406
rs109372952	25966829	0.206
rs135375478	25972263	0.032
rs42298471	25979073	0.079
rs109341059	25980137	0.022
rs42298470	25982072	0.079
rs110558178	25983064	0.022
rs110982026	25985624	0.036
rs42298467	25986431	0.237
rs29017100	25987996	0.027
rs29017103	25991165	0.174
rs42306917	25992595	0.079
rs137839813	25999691	0.016

Appendix V

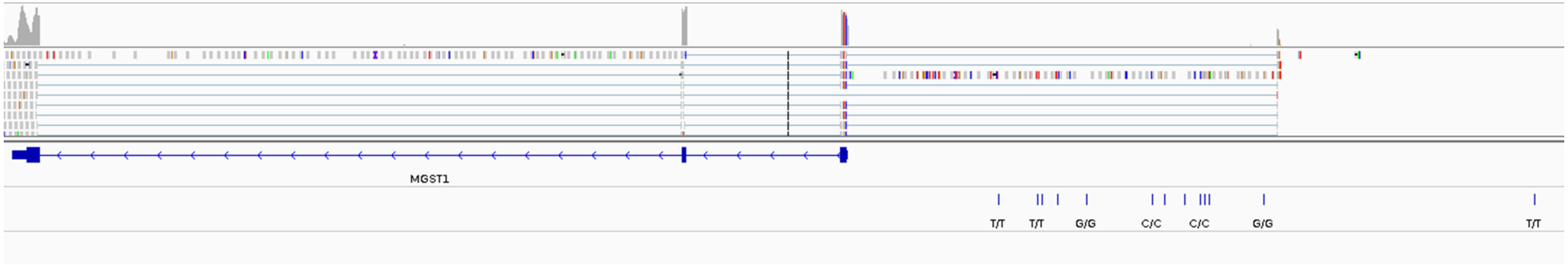


Figure 1. Schematic of *MGST1* gene (in blue) and candidate causative variants (below). The top track is mammary RNAseq data. There are a cluster of statistically identical variants that reside within 4 kb of the TSS of *MGST1*.

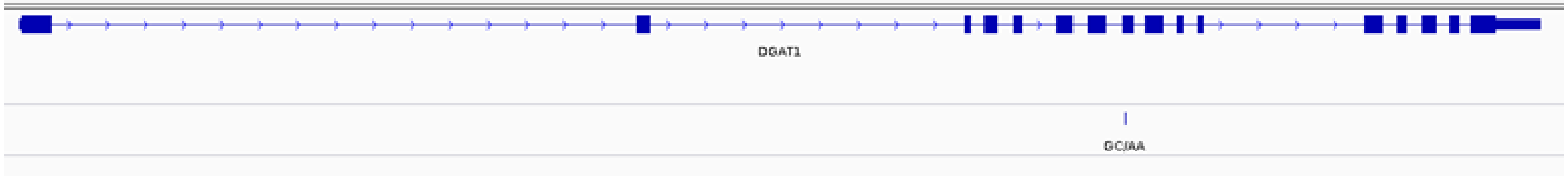


Figure 2. Schematic of *DGAT1* gene (in blue) with the dinucleotide substitution (AA>GC) responsible for DGAT1 K232A in exon 8 of the gene.

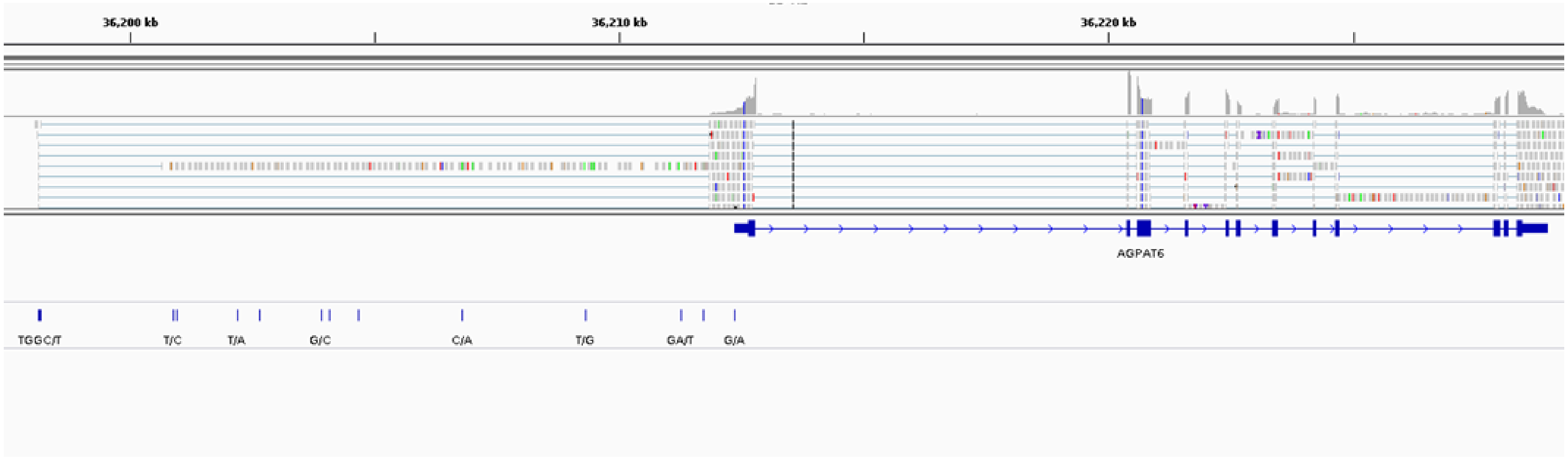


Figure 3. Schematic of *AGPAT6* gene (in blue) and candidate causative variants (below). The top track shows mammary RNAseq data indicating the previously unannotated first exon of the gene. The *AGPAT6* VNTR is the left most variant (TGGC>T) in the first exon.

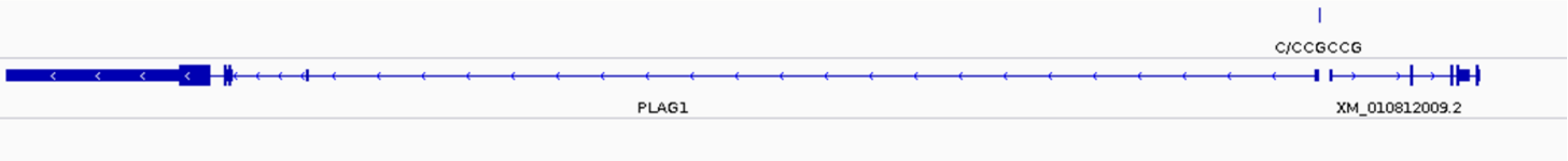
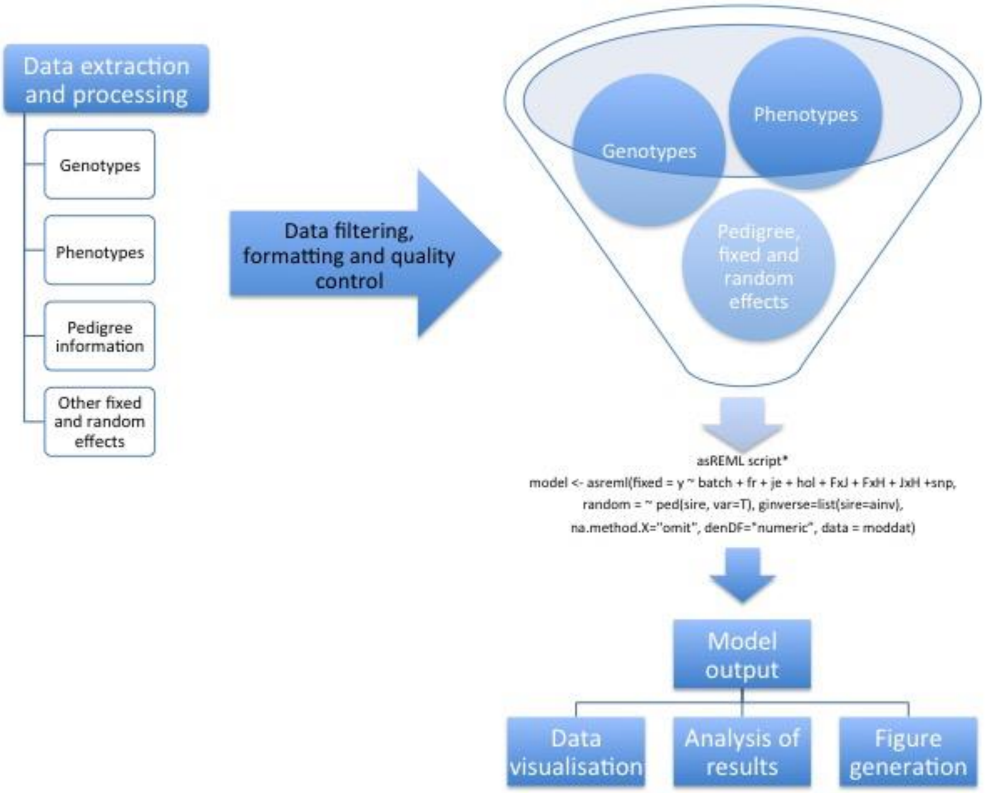


Figure 4. Schematic of *PLAG1* gene (in blue) and the candidate causative variant in the promoter of the gene (above). XM_010812009.2 is also known as *CHCHD7*.

Appendix VI

Each results chapter in this thesis describes the use of an asREML model for genetic association analysis for milk production and gene expression traits. This required the use of many bioinformatic scripts and tools including IGV tools, PLINK, R and R Studio and bash scripts. The primary asREML script was developed by Kathryn Sanders at LIC, and in conjunction with her I modified this existing script for inclusion in my analyses.


Data extraction and processing
Genotypes
Phenotypes
Pedigree information
Other fixed and random effects
Data filtering, formatting and quality control
Genotypes
Phenotypes
Pedigree, fixed and random effects
asREML script*
Model output
Data visualisation
Analysis of results
Figure generation

```
model <- asreml(fixed = y ~ batch + fr + je + hol + FxJ + FxH + JxH +snp,
random = ~ ped(sire, var=T), ginverse=list(sire=ainv),
na.method.X="omit", denDF="numeric", data = moddat)
```

Transformation and subsetting output files in sam, fastq, and vcf format

Figure 1. Genetic association analysis pipeline used in this thesis.

The first step in the genetic association analysis pipeline was the extraction and processing of the data which was held in LIC databases. This data handling was carried out by generating a list of animal keys, pedigree files, and extracting phenotypes and genotypes required for each analysis and transforming these into a format suitable for use in the model.

This involved the use of R, VCF tools and PLINK scripts. The next step was to run the asREML model which required the development of R scripts which called functions in the primary asREML script specific to each analysis (example presented below). The model then outputted the results (effect sizes and standard errors for each model) with summary statistics including p-values and F statistics, as well as the percentage phenotypic and genotypic variance explained in each model. The analysis of this data was conducted predominately in R and I wrote many scripts for this analysis and the generation of figures.

In addition to the bioinformatic scripts and tools used as part of this genetic analysis, additional scripts were developed for the analysis of sequence data generated as part of other experimental work described in this thesis. These scripts included the use of commands in samtools, bedtools, freebayes and GATK.

```
#Getting a list of sire animal keys for RNAseq animals


anmls <- read.csv("/data/seq/talaw0/MGST1_trans_eQTL/asreml/datafile.csv",
header=TRUE)[ ,2, drop=FALSE]
colnames(anmls) <- "anml_key"

source("/data/gs/kasan0/My_scripts/getAnimalPD01.r")
anml_dat <- getAnimalPD01(anmls)[ ,c("anml_key", "sire_anml_key")]

# generate a file of sire keys
sires <- unique(anml_dat$sire_anml_key[anml_dat$sire_anml_key != 0])
sires <- sort(sires)
write.table(sires, file="sires.txt", sep=" ", quote=FALSE,
                        row.names=FALSE, col.names=FALSE)
#Generating Pedigree file
pedfile = "/data/seq/talaw0/MGST1_trans_eQTL/asreml/pedigree.txt"
siresfile =”/data/seq/talaw0/MGST1_trans_eQTL/asreml/sire.txt"

system(paste("bash /data/ls/dajoh0/pedigree/getped.sh '", siresfile, "' '",
pedfile, "'", sep=""))

ped = read.table('/data/seq/talaw0/MGST1_trans_eQTL/asreml/pedigree.txt',
skip=1, sep="", strip.white=T, stringsAsFactors = FALSE)
```

```
ped = subset(ped, select=c(1:6))
colnames(ped) = c("anim", "sire", "dam", "anml_key", "sire_key", "dam_key")
peddams = ped[,c("anim", "sire")]
colnames(peddams) = c("dam", "mgs")
ped = merge(ped, peddams, by="dam", all.x=TRUE)
ped$mgs[is.na(ped$mgs)==TRUE] <- 0
ped = ped[order(ped$anim), c("anim", "sire", "mgs", "anml_key")]
colnames(ped) = c("sire", "pgsire", "pmgsire", "sire_anml_key")
ped = subset(ped, select=c(1:3))
write.table(ped, file="pedfile_upd.txt", sep=" ", quote=FALSE,
row.names=FALSE, col.names=TRUE)


ssh jazz
R
Library(asreml)


# Read in pedigree file
  ped =
read.table('/data/seq/talaw0/MGST1_trans_eQTL/asreml/pedfile_upd.txt',
header=TRUE, sep= "", stringsAsFactors=FALSE)


# Generate the a inverse matrix
   ainv <- asreml.Ainverse(ped, mgs=TRUE)$ginv


modeldat <-
asreml.read.table("/data/seq/talaw0/MGST1_trans_eQTL/asreml/datafile.csv",
sep=",", header=TRUE)


modeldat$MGST1 = as.numeric(modeldat$MGST1)


modout <- asreml(fixed = MGST1~1 + Batch + fr + je + hol + FxJ + FxH + JxH
+ Genotype, random = ~ ped(sire, var=T),
ginverse=list(sire=ainv),na.method.X="omit", denDF="numeric",
            data = modeldat, workspace=4e+08, pworkspace=4e+08)


# Fit the model with pedigree
   modout <- asreml(fixed = y ~ Batch + fr + je + hol + FxJ + FxH + JxH
+ Genotype, random = ~ ped(sire, var=T),
ginverse=list(sire=ainv),na.method.X="omit", denDF="numeric",
            data = modeldat, workspace=4e+08, pworkspace=4e+08)
```

```
phens <- c("myphen1", "myphen2")
snps <- snpslist

for (phen in 1:length(phens)) {

  for (j in 1:length(snpslist)) {

    phenName <- phens[i]
    snpName <- snps[j]

   moddattemp <- modeldat
    names(moddattemp) [names(moddattemp) == phenName] <- "y"
    names(moddattemp) [names(moddattemp) == snpName] <- "snp"

 moddattemp$snp <- as.numeric(moddattemp$snp)
 moddattemp <- moddattemp[is.na(moddattemp$y)==FALSE &
                          is.na(moddattemp$snp)==FALSE, ]

 if (!(nrow(moddattemp) > 0 & var(moddattemp$y) > 0 &
            var(moddattemp$snp) > 0)) {
         print(paste("No data for model available or no variance in",
phenName, "or", snpName))
 } else {

       modout <- asreml(fixed = y ~ Batch + fr + je + hol + FxJ + FxH + JxH
+
                                     snp,
                         random = ~ ped(sire, var=T),
ginverse=list(sire=ainv),
                         na.method.X="omit", denDF="numeric",
                         data = moddattemp, workspace=5e+08,
pworkspace=5e+08)
        ....

          If (i==1 & j==1) {outdat <- newoutdat} else {outdat <-
rbind(outdat, newoutdat)}
 }
      # Summarise model

      # p-values and F statistics
      aovdat <- wald.asreml(modout, Ftest = formula("~NULL"),
                            denDF="numeric", ssType="conditional")
      print(aovdat)
      aov <- as.data.frame(aovdat[1])

      pval <- aov[rownames(aov) == "snp", ][6]

      if (!(nrow(pval) > 0)) {
        print("Unable to estimate snp effect (all animals except 1 animal
in same genotype class).")
      } else {

      pval[pval == 0] <- 2.2251e-308
      colnames(pval) <- c("pval")
      print(pval)
      waldFstat <- aov[rownames(aov) == "snp", ][4]
      colnames(waldFstat) <- c("waldFstat")
      print(waldFstat)

      # identify snp genotype frequencies
      freqs <- as.data.frame(table(moddattemp[ ,"snp"]))
      freqs <- subset(freqs, Var1 != ".")
      n_0 <- max(freqs[freqs$Var1 == 0, "Freq"],
0)

      n_1 <- max(freqs[freqs$Var1 == 1, "Freq"], 0)
      n_2 <- max(freqs[freqs$Var1 == 2, "Freq"], 0)
      total <- sum(freqs$Freq)*2
      p <- (2*n_0+n_1)/total
      q <- (2*n_2+n_1)/total

      # estimated additive snp effect and std error - trend model
      estdat <- summary(modout, all=TRUE)$coef.fixed[row.names="snp",
      c("solution", "std error")]
      names(estdat) <- c("trend_est", "trend_se")
      trend_est <- estdat["trend_est"]
      trend_se <- estdat["trend_se"]

      # estimate ls means - trend model
      preddat <- predict(modout, classify='snp',
                         levels=list('snp'=c(0,1,2)), trace=FALSE)
      lsmean_0 <-
preddat$predictions$pvals$predicted.value[preddat$predictions$pvals$snp==0]
      lsmean_1 <-
preddat$predictions$pvals$predicted.value[preddat$predictions$pvals$snp==1]
      lsmean_2 <-
preddat$predictions$pvals$predicted.value[preddat$predictions$pvals$snp==2]
      se_0 <-
preddat$predictions$pvals$standard.error[preddat$predictions$pvals$snp==0]
      se_1 <-
preddat$predictions$pvals$standard.error[preddat$predictions$pvals$snp==1]
      se_2 <-
preddat$predictions$pvals$standard.error[preddat$predictions$pvals$snp==2]
      lsmeans <- cbind(n_0, lsmean_0, se_0, n_1, lsmean_1, se_1,
                       n_2, lsmean_2, se_2)

      # estimate variance accounted for by snp
      add_var <- 2*p*q*trend_est*trend_est

      # identify and summarise variance components
      print(summary(modout)$varcomp)
      varcompdat <- summary(modout)$varcomp[ ,"component"]
      genot_var <- varcompdat[1] * 4
      res_var <- varcompdat[2] - (varcompdat[1] * 3)
      tot_var <- genot_var + res_var + add_var
      pct_pheno_var_explan <- add_var/tot_var*100
      tot_gen_var <- genot_var + add_var
      pct_gen_var_explan <- add_var/tot_gen_var*100

      # estimate heritability
      h2dat <- pinfn(modout, anim.prop ~ (V1 * 4) / ( V1 + V2))
      h2 <- as.numeric(h2dat[1])
      h2_se <- as.numeric(h2dat[2])
      #h2 <- genot_var/(genot_var + res_var); h2 # estimate heritability

      # model mse
      mse <- modout$sigma2

      # consolidate output data
     newoutdat <- cbind.data.frame(phenName, snpName, p, q,
                      snpEff, trend_est, trend_se, pval, waldFstat,
                      lsmeans, add_var, genot_var, res_var, tot_var,
                      pct_pheno_var_explan, tot_gen_var,
pct_gen_var_explan,
                      h2, h2_se, mse)
```