Systems level investigation of the genetic basis of bovine muscle growth and development

A thesis submitted to the National University of Ireland for the degree of
Doctor of Philosophy

Presented by:
Anthony Gerard Doran

NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

This research was conducted at the Animal and Bioscience Research Department, Animal & Grassland Research and Innovation Centre, Teagasc, Grange, Dunsany, Co. Meath, Ireland and the National University of Ireland Maynooth, Department of Biology, NUI Maynooth, Maynooth, Co. Kildare, Ireland.

Teagasc
Agriculture and Food Development Authority
Walsh Fellowship Funded Research

October 2013

Supervisors:	Dr. Christopher Creevey B.Sc., Ph.D. (**Teagasc**) Dr. Donagh Berry B.Agr. Sc., Ph.D. (**Teagasc**) Professor James McInerney B.Sc., Ph.D., D.Sc. (**NUIM**)
Head of Department:	Professor Paul Moynagh, B.A., Ph.D.

Table of Contents

For my mother, brothers and sisters

Acknowledgments

Firstly, I would sincerely like to thank Dr. Christopher Creevey for all of the support and guidance over the last four years. It has been an absolute pleasure to work with you. This would not have been possible without all of your advice, guidance and expertise. I really feel extremely lucky to have had such a great supervisor.

I would also like to thank Teagasc and the Walsh Fellowship scheme for providing me the opportunity to carry out the research undertaken throughout this thesis. Also, I would like to thank the Irish Cattle Breeding Federation for providing the raw data used for a section of this thesis.

Also, to my other Teagasc supervisor, Dr. Donagh Berry, I would like to thank you for all of the advice and support you have given. Your feed back and guidance throughout is greatly appreciated.

I would also like to thank Professor James McInerney. Your advice, in particular towards the end of the fours years, has been really helpful. In addition, to the three girls that finished in your lab last year, particularly to Aoife, I would like to say thanks for the help, advice and good times.

To everyone in Grange, past and present; the staff, researchers and students, I would like to say thanks. At one stage or another, many of you will have contributed comments and suggestions to my work and presentations. In particular, thanks to Dr. Sinead Waters for advice on various aspects of my work, Dr. Sarah Keady, with whom I collaborated with on a section of this thesis, and Dr. Matt McCabe.

Thanks to the guys in the office, Bruce and Francesco. Sharing an office with both of you for the past few years has been great. The discussions, debates, conferences, meetings and drinks throughout have always been better with you both. I can only hope that any office mates I have in the future are half as good as either of you.

To all of my friends in Teagasc, Maynooth (especially Stephen and Liam), Wexford, abroad and anywhere else, I really want to say thanks for everything. Your support and encouragement has been immense. To the friends that I made along the way, my four years have been improved by meeting and getting to know you. In particular, Keelin for all of your help when I moved to Cork, and Eunan for making my two stays in Trim far more enjoyable.

Finally, thank you to my Family. Anytime I ever needed anything, you were there for me. This thesis would not have been possible without your support. To my mother (Mam), Mary, I am eternally grateful for the support and encouragement you have always shown me. To my two brothers, Seán and James, and my two sisters, Aileen and Tríona, thanks for all of your support and understanding throughout.

Declaration

This thesis has not been submitted in whole, or in part, to this, or any other university for any other degree and is, except where otherwise stated, the original work of the author.

Signed: ___________________________________

Anthony Gerard Doran

Publications

Doran, A.G. and Creevey, C.J. (2013). SNPdat: Easy and rapid annotation of results from de novo snp discovery projects for model and non-model organisms. *BMC bioinformatics*, 14:45.

Doran, A.G., Berry, D.P. and Creevey, C.J. Whole genome association study identifies regions of the bovine genome and biological pathways involved in carcass trait performance in Holstein-Friesian cattle. *BMC genomics* (manuscript in review).

Keady, S.M.*, **Doran, A.G.***, Kenny, D.A., Creevey, C.J. and Waters, S.M. Transcriptional response to reduced energy intake and subsequent compensatory growth of M. longissimus thoracis et lumborum in Aberdeen Angus steers (manuscript in preparation).

Doran, A.G. and Creevey, C.J. Convergence diagnostics for a Bayesian model used in genetic prediction (manuscript in preparation).

Doran, A.G.*, Mullen, M.P.*, Waters, S.M., McCabe, M.S., Meade, K.G. and Creevey, C.J. Divergent evolution of somatotropic axis genes in dairy and beef (Bos taurus) cattle (manuscript in preparation).

*SHARED FIRST AUTHOR

Abbreviations

ADF	Acid Detergent Fibre
BH	Benjamini-Hochberg
bp	Base Pair
BP	Before Present
CDS	Coding Sequence
CFAT	Progeny Carcass Fat
CONF	Progeny Carcass Conformation
CG	Compensatory Growth
CP	Crude Protein
CULL	Cull Cow Carcass Weight
CWT	Progeny Carcass Weight
DM	Dry Matter
DMD	Dry Matter Digestibility
DNA	Deoxyribonucleic Acid
DPBS	Dulbecco's Phosphate Buffered Saline
EBV	Estimated Breeding Value
ED	Euclidean Distance
FDR	False Discovery Rate
FWER	Family Wise Error Rate
gEBV	Genomic Estimated Breeding Value
Gbp	Giga Base Pair (one billion base pairs)
GEO	Gene Expression Omnibus
GFF	General Feature Format
GTF	General Transfer Format
GWAS	Genome-Wide Association Study

ICBF	Irish Cattle Breeding Federation
kb	kilobase (one thousand base pairs)
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LD	Linkage Disequilibrium
LMM	Linear Mixed Model
Mb	megabase (one million base pairs)
MCMC	Markov Chain Monte Carlo
MK	McDonald-Kreitman
mRNA	Messenger RNA
mtDNA	mitochondrial DNA
NDF	Neutral Detergent Fibre
NGS	Next generation sequencing
nm	Nanometre
PCD	Primary Ciliary Dyskinesia
PCR	Polymerase Chain Reaction
PP	Posterior Probability
pSSR	Proportion of non-significant SNPs estimated from the SSR
PTA	Predicted Transmitting Ability
QTL	Quantitative Trait Loci
RNA	Ribonucleic Acid
R^2	Coefficient of Determination
rs#	Reference SNP ID number (dbSNP identifier)
SAE	sum of all absolute SNP effects
SDE	Significantly differentially expressed
SNP	Single Nucleotide Polymorphism
SSE	sum of all SNP effects

ss#	Submitted SNP ID number (dbSNP identifier)
SSR	Single SNP Regression
UCSC	University of California Santa Cruz
URL	Uniform Resource Locator
UTR	Untranslated Region
VCF	Variant Calling Format
^{o}C	Degrees Celsius
€	Euro (currency)

Index of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Index of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Index of Equations

Index of Appendices

Index of Electronic Appendices

Chapter 2

Chapter 3

Chapter 5

Chapter 6

Abstract

Skeletal muscle growth is an economically and biologically important trait for livestock raised for meat production. As such, there is great interest in understanding the underlying genomic architecture influencing muscle growth and development. In spite of this, relatively little is known about the genes or biological processes regulating bovine muscle growth. In this thesis, several approaches were undertaken in order to elucidate some of the mechanisms which may be controlling bovine muscle growth and development. The first objective of this thesis was the development of a novel software tool (SNPdat) for the rapid and comprehensive annotation of SNP data for any organism with a draft sequence and annotation. SNPdat was subsequently utilised in chapters 3 and 6 to facilitate the identification of candidate genes and regions involved in bovine muscle growth. In chapter 4, a number of metrics were explored for their usefulness in assessing convergence of a Markov Chain using a Bayesian approach used in genetic prediction. The need to adequately assess convergence using multiple metrics is addressed and recommendations put forward. These recommendations were then implemented in chapter 3. In addition, three separate investigations of bovine muscle growth and development were performed. In chapter 3, a genome-wide association study was performed to identify regions of the bovine genome associated with four economically important carcass traits. This was followed by an examination of the transcriptional responses in muscle tissue of animals undergoing dietary restriction and compensatory growth (chapter 5). Finally, using high-throughput DNA sequencing, a candidate list of 200 genes was interrogated to identify genes which may be evolving at different rates, and under evolutionary selection pressure, in beef compared to dairy animals (chapter 6). A number of genes and biological pathways were found to be involved in traits related to bovine muscle growth, several of which were identified in more than one study.

Chapter 1: Introduction

1.1 Establishment of Cattle for domestic purposes

The domestication of cattle from the extinct wild aurochs (*Bos primigenius*) occurred in the early Holocene period between 8,000 and 10,000 years ago (Willham, 1986; Vigne, 2011). It is generally accepted that modern cattle originated from two independent domestication events, in the Near East and the Indian subcontinent, giving rise to the two major taxa observed today; *Bos taurus* and *Bos indicus*, respectively (Loftus et al., 1994; Troy et al., 2001; Achilli et al., 2009; Ramey et al., 2013). This hypothesis was confirmed by genetic analyses of mitochondrial DNA (mtDNA) between modern *Bos taurus* and *Bos indicus* haplotypes, which indicated their emergence from two genetically and geographically divergent populations. (Bruford et al., 2003; Edwards et al., 2010; Zeyland et al., 2013).

The move toward domestication and control of captive animals marked an important step in early human culture (Vigne, 2011). Domesticated cattle provided resources such as milk, meat and draft ability (Sherratt, 1983; Evershed et al., 2008). Consequently, early cattle were subject to simple selective breeding for favourable traits based on morphologically obvious traits such as milk yield and size. (Vigne, 2011). For millennia this “soft selection” continued, until the emergence of the breed concept 200 years ago (Figure 1.1) (Taberlet et al., 2011). Selection pressure rapidly increased, resulting in the development of individual breeds specialised in, for example, milk yield and growth potential (Gibbs et al., 2009; Taberlet et al., 2011). Today, over 1,000 cattle breeds have been established, many of which are specialised in a particular trait of economic importance. Although much of the genomic architecture involved in complex traits has yet to be revealed, cattle represent a significant scientific resource for investigating and understanding the genetics of observed phenotypic variation (Andersson and Georges, 2004; Elsik et al., 2009).

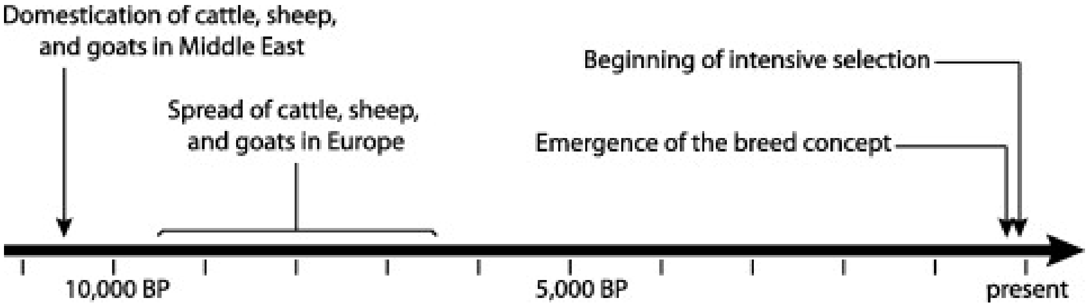

Domestication of cattle, sheep, and goats in Middle East
Spread of cattle, sheep, and goats in Europe
Beginning of intensive selection
Emergence of the breed concept
10,000 BP
5,000 BP
present

Figure 1.1 The main events in cattle, sheep and goat domestication. (Adapted from (Taberlet et al., 2011)). Events are displayed in years before present (BP).

1.1.1 Beef production in Ireland

The domestication of animals has allowed humans to utilise materials of relatively little value (for example, low quality forage and natural pasture) to produce products of high nutritional and commercial value. Domesticated cattle are an important source of food, meat, hides, hair etc. In Ireland, there are currently over 6 million cattle, consisting of both dairy and beef breeds (Department of Agriculture Food and the Marine, 2012). Among the variety of different production systems throughout Europe, there exist two prominent systems; pasture-based systems, predominately in the west, and cereal-based systems, mostly in central-eastern parts of Europe (Bernues et al., 2011). Combinations of these systems are also commonly found. Ireland has a climate ideal for the former, thus beef production within Ireland is predominately a grass based system (Drennan and Mcgee, 2009).

Within beef production systems, breeding programmes have been developed to select for more productive and efficient animals, leading to the use of animals with higher genetic merit and increased economic value (Barwick and Henzell, 2005). Historically, beef producers have been provided with breeding values to be used as selection tools, however, it had been left to the decision of the producer to decide the economic value of each trait and ultimately the use of those breeding values (Bourdon, 1998). In Ireland, selection indexes for the beef industry began to emerge in the early 2000's (Amer et al., 2001). Selection indexes provide producers with clear breeding objectives that can guide overall genetic change (Enns and Nicoll, 2008; Garrick and Golden, 2009). Sophisticated statistical methods have successfully been applied in animal breeding, resulting in considerable genetic gains in traits that were selected on (for a review see Golden et al., 2009). These methods are based on knowledge of population genetic parameters, without any knowledge of the genetic architecture of the trait or the number of genes affecting the trait (Berry et al., 2011). Recent advances in sequencing technology and bioinformatics approaches are facilitating the dissection of complex traits, which could potentially be included in future breeding programmes to further increase genetic gain. However, key genes and pathways regulating growth traits will need to be elucidated before this information can be successfully integrated into breeding programmes (Dekkers and Hospital, 2002).

1.1.2 Genetic improvement

In Ireland, the main body responsible for generating genetic information for cattle breeding is the Irish Cattle Breeding Federation (ICBF). Prior to the establishment of the ICBF in 1997, animal records on breeding and performance were submitted by farmers to multiple organisations such as the Department of Agriculture. This led to problems in national comparisons and duplicate data. To increase the rate of progress within the industry, an agreement to centralise this data into one single repository was reached. With this the ICBF came into existence (Wickham and Durr, 2011).

The ICBF assumed the responsibilities of the Department of agriculture, including the collection and maintenance of data, recording of breeding and the estimation and release of animal genetic evaluations. For each trait currently recorded in Ireland, the ICBF calculates a measure of genetic merit for all Irish dairy and beef cattle. The phenotype of an animal is a combination of both genetic and environmental effects. Genetic effects are the result of genes inherited from parents, whereas environmental effects are the result of conditions experienced by the animal (e.g. nutrition and temperature). Measures of genetic merit, such as a predicted transmitting ability (PTA) or an estimated breeding value (EBV), account for the additive genetic effects that are responsible for performance of an animal and its progeny (Vanraden et al., 1990). Measures of genetic merit are predicted from performance information of the animal, its progeny and ancestors. The genetic evaluation for each trait also includes an estimated reliability. The reliability measures the confidence in the estimated measure of genetic merit (scale between 0% - 99%). As more information is included in an animal's genetic evaluation, the reliability of the evaluation will increase. As the reliability increases, there is a decreasing chance that the evaluation will change in the future as more information is included. The ICBF continually updates estimates of genetic merit and reliability as more performance information is obtained for each animal or its relations.

1.1.3 Selection indexes

The ICBF also implemented a change in breeding indexes, introducing a multi-trait index for dairy breeding in 2001 called the Economic Breeding Index (EBI) (Veerkamp et al., 2002). Previous breeding indexes selected aggressively for commercially important production traits, but failed to incorporate functional traits such as fertility and health (Miglior et al., 2005). However, selection for production traits alone may have negative effects on functionally important traits (e.g. health (Heringstad et al., 2003) and reproduction (Veerkamp et al., 2001)). The EBI represented a broader and more balanced selection index that incorporated both functional and production traits. Breeding objectives for beef cattle in Ireland were outlined by Amer *et al.* (2001). Currently, two main indexes, which include both functional and production traits, are in use in beef breeding in Ireland; the "maternal index" and the "terminal index".

1.1.3.1 The maternal and terminal indexes

Both the maternal and terminal indexes are indicative of an animal's genetic merit. The maternal index specifically relates to an animal's genetic merit to produce profitable daughters, whereas the terminal index is related to an animal’s genetic merit to produce cattle for slaughter or sale. Each index contains its own sub-indexes, each with its own emphasis on different functional and production traits (Table 1.1 and Table 1.2). Animals are given a single figure, calculated from the genetic merit of the animal for each sub-index and the relative emphasis on that sub-index. The idea is that a single value related to the overall satisfaction with an animal can be achieved, and that this value can describe the relationship of the animal to profitability on either the maternal or the terminal index. As such, the index figures are expressed in Euros, relative to the performance of the base population (Campion et al., 2009) i.e. an animal with a maternal index of €100 is expected to produce daughters that are €100 more profitable.

Table 1.1 The maternal index and relative weightings on sub-indexes.

Maternal Index	
Sub-index	**% Emphasis**
Calving	24
Beef	43
Fertility	19
Milk	11
Docility	3

Table 1.2 The terminal index and relative weightings on sub-indexes.

Terminal Index	
Sub-index	**% Emphasis**
Calving	29
Beef	68
Docility	3

1.1.4 Limitations of traditional quantitative genetics

Despite considerable genetic gain in recent decades, traditional selection approaches based on quantitative genetics methodology have a number of limitations. For example, large and expensive breeding schemes are needed to accurately estimate an animal's genetic merit or the animal may need to be sacrificed in order to obtain the phenotype. Additionally, traits that are antagonistically correlated cannot be easily resolved in a traditional manner (Berry et al., 2011). Improved understanding of the genetic architecture, through functional studies or available interaction data, of complex traits may be incorporated into future breeding programmes leading to improved accuracy of selection (Berry et al., 2011; Snelling et al., 2013). Identifying key processes under selection may also help resolve problems of antagonistic genetic correlations between traits (Berry et al., 2011). However, many of the mechanisms affecting complex traits, such as growth, have yet to be elucidated and warrant further investigation.

1.2 Skeletal muscle

Skeletal muscle is a form of striated muscle tissue attached to the skeleton. The entire muscle is surrounded by the epimysium, a type of connective tissue. Skeletal muscle tissue is composed of complex bundles of muscle fibres (myofibres) called fascicles, which are in turn surrounded by the perimysium. Myofibres are long, multinucleated structures that are bound by a cell membrane, the sarcolemma (Figure 1.2). Each myofibre contains cylindrical bundles of contractile filaments known as myofibrils. Myofibrils consist of an ordered arrangement of longitudinal myofilaments which are grouped into two types; thin filaments (diameter 6–10 nm composed of actin) and thick filaments (diameter 14–16 nm and composed of myosin) (Hooper et al., 2008). Each thin filament is composed of two strands of actin. Tropomyosin (filamentous) and troponin (globular) are thin filament regularity proteins found in striated muscle. Tropomyosin blocks actin binding sites at rest, but moves away in the presence of Ca^{++} (Hooper et al., 2008). Thick filaments are comprised of myosin, which is used to produce force by engaging actin filaments (Hooper et al., 2008). Together, thin and thick filaments produce movement by contraction through the sliding filament model (Huxley and Niedergerke, 1954; Huxley and Hanson, 1954).

Signals penetrate the sarcolemma through T-tubules, activating the sarcoplasmic reticulum leading to muscle contraction. Repeating units known as sarcomeres are the basic units of muscle. A sarcomere consists of light bands (known as the I-band) and dark bands (known as the A-band) giving muscle the characteristic striation observable by light microscopy. Sarcomeres are bounded by thin filaments called Z-lines that bisect the I-band such that one half of the I-band belongs to one sarcomere and the other half to the adjacent sarcomere (Hooper and Thuma, 2005). At the centre of the sarcomere is the H-band, comprising only thick filaments. The H-band is bisected by the M-line. Contraction of the sarcomere, and consequently the entire muscle, happens when the Z-lines move closer together (Davies and Nowak, 2006). The complexity and design of this bundle within a bundle organisation throughout muscle tissue is key to the strength and coordination of muscle contraction (Blandin et al., 2013).

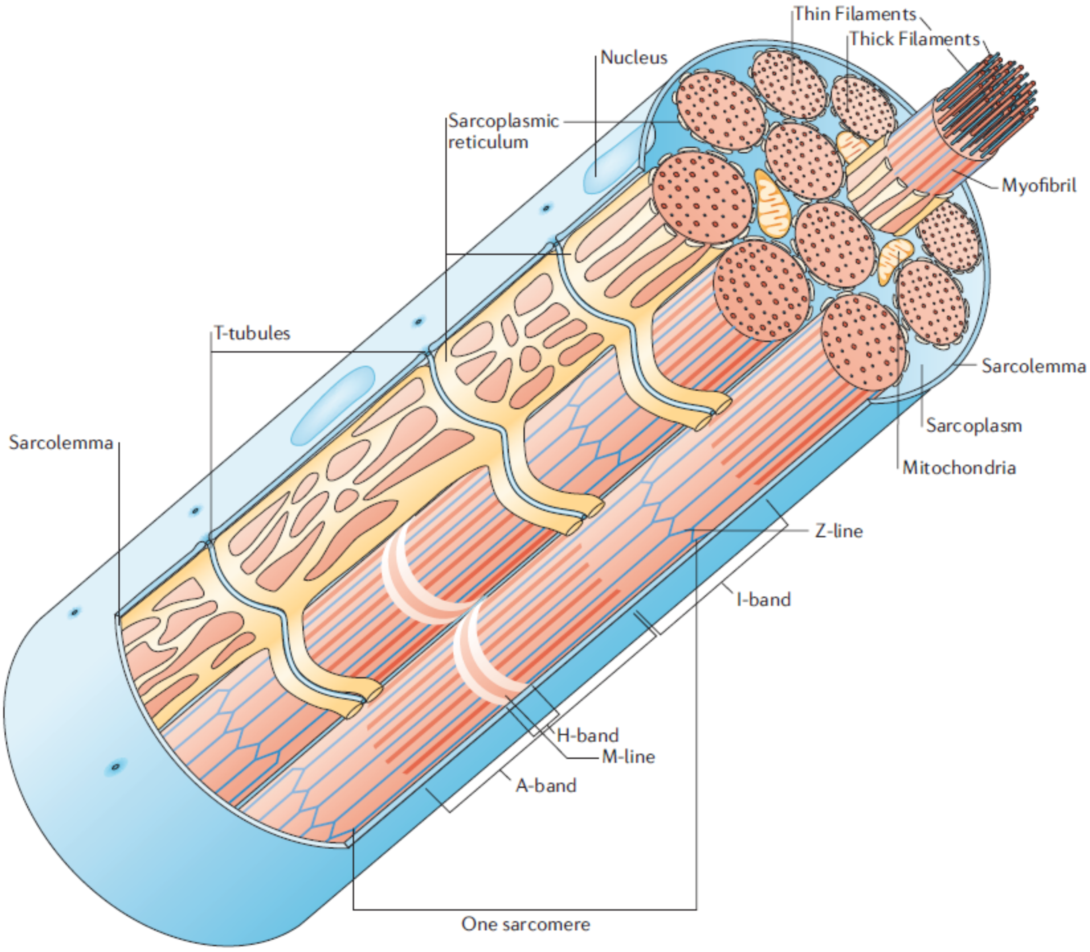


Figure 1.2 The structure of a skeletal muscle fibre. (Adapted from (Davies and Nowak, 2006)).

1.2.1 Cattle breed muscle characteristics

1.2.1.1 Holstein-Friesian

In Ireland, there are over 6 million cattle, of which approximately 2.3 million are Holstein-Friesian (Department of Agriculture Food and the Marine, 2012). Holstein-Friesian cattle are a popular breed of cow known for their ability to produce large amounts of milk (Gibbs et al., 2009). Although Holstein-Friesian cattle are used primarily in dairy production systems, they are also an important source of meat for beef production and export.

1.2.1.2 Aberdeen Angus

There are about 3.5 million beef cattle in Ireland, of which approximately 16% are Aberdeen Angus animals (Department of Agriculture Food and the Marine, 2012). Approximately 13.4 % of all dairy cows are bred to Aberdeen Angus bulls in Ireland. The Aberdeen Angus is an early maturing breed that was developed in north-eastern Scotland (Gibbs et al., 2009; McTavish et al., 2013). Aberdeen Angus animals are known for their characteristic black coat and greater levels of intramuscular fat marbling which is associated with increased meat quality (Kuber et al., 2004; Gibbs et al., 2009; Keady et al., 2013).

1.2.2 Compensatory growth

Following long periods of reduced feed intake, many organisms have the capacity to rapidly increase growth rate following re-alimentation to a higher energy diet (Figure 1.3). This phenomenon, commonly referred to as compensatory growth, allows organisms to achieve a genetically pre-determined inherent size following periods of restricted energy intake (Connor et al., 2010). Examples of compensatory growth have been observed in fish (Ali et al., 2003), cattle (Lehnert et al., 2006) and even humans (Ashworth, 1969). An animal experiencing compensatory growth is characterised by a significantly faster growth rate and increased efficiency of energy use compared to animals that have not experienced reduced feed intake (Hornick et al., 2000; Ali et al., 2003; Connor et al., 2010). The ability of an animal to compensate is dependant on several factors including; the severity and duration of reduced feed intake, the age and sex of the animal and the quality of the re-alimentation diet (Ali et al., 2003).

A number of studies have tried to elucidate the mechanisms regulating compensatory growth in several livestock species basis (Johansen and Overturf, 2006; Lehnert et al., 2006; e.g. Picha et al., 2008; e.g. Connor et al., 2010). In spite of this, the mechanisms underlying compensatory growth have remained concealed. For example, Lehnert *et al.* (2006) identified only a single gene as differentially expressed during re-alimentation compared to controls, although gene expression was examined quite late in to the re-alimentation period (84 days) when animals had entered normal growth trajectories.

In cattle production systems, compensatory growth is gaining attention from many producers as a means to offset high feed costs. In beef cattle production systems, animal feed costs are highest during Winter and lowest in Spring. Because of this, compensatory growth is seen as a potential management strategy to offset high costs over Winter until feed is available as cheap Spring pasture. Additionally, compensatory growth presents an interesting model to study the effects of reduced feed intake on muscle tissue and the subsequent acute response observed upon re-alimentation.

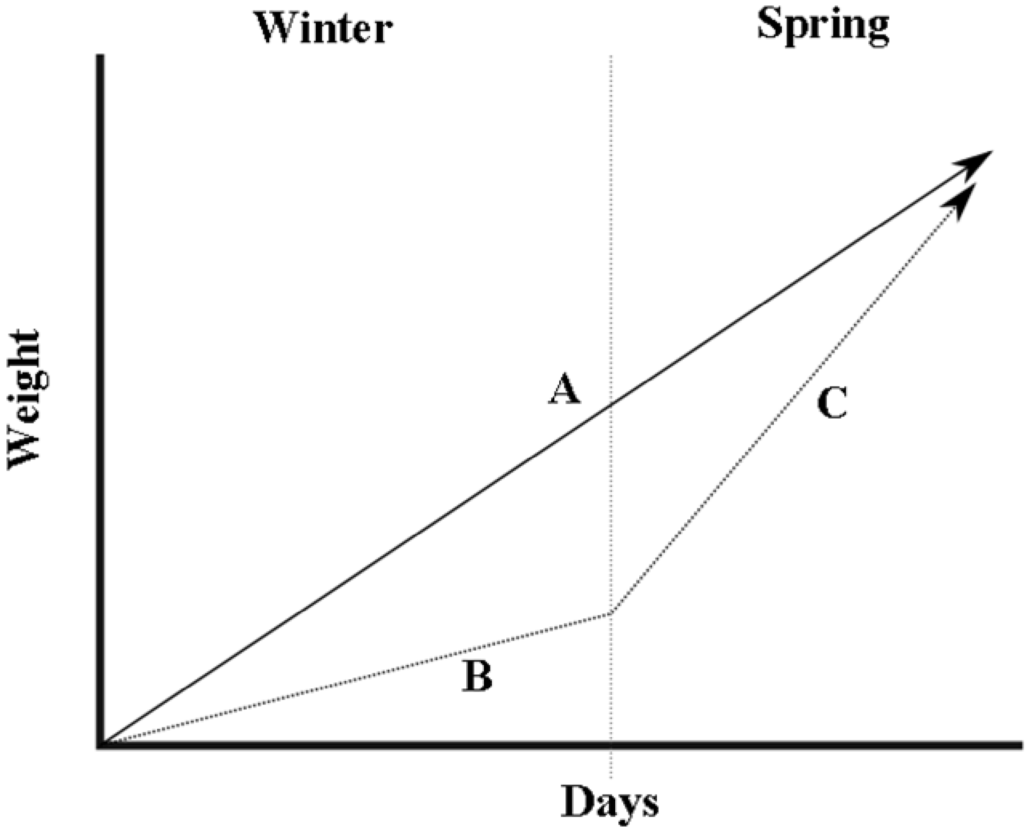


Figure 1.3 Growth trajectories associated with compensatory growth. Implementation of compensatory growth in animal production may include a restricted feeding period over Winter, and a recovery period during Spring. (A) = The expected growth trajectory of animals experiencing normal conditions. As the number of days increases, the weight of the animal will continue to steadily increase. (B) = the depressed growth trajectory experienced during a restricted feeding period. (C) = A rapid increase in growth rate experienced by animals during compensatory growth. Adapted from (Hornick et al., 2000).

1.2.3 Genes involved in regulating muscle growth

1.2.3.1 Growth hormone

Growth hormone (GH), an anterior pituitary produced hormone, is a key regulator of metabolism and growth in mammals (Mullen et al., 2010). In cattle, administration of GH has also been shown to stimulate muscle growth by increasing the size of myofibres (Vann et al., 2001; Jiang and Ge, 2013). The effects of GH are mediated by binding to its receptor (GHR), which in turn stimulates the release of insulin-like growth factor 1 (IGF1), an important circulating growth factor (Bonaldo and Sandri, 2013). GHR expression has been reported in several tissues in cattle including liver, fat and skeletal muscle (Jiang and Ge, 2013). Disruption of GH, or GHR, results in mice that are 50% smaller than their wild-type littermates (Zhou et al., 1997; Jiang and Ge, 2013). Additionally, a number of studies have reported associations between single nucleotide polymorphisms, in GH and GHR, and a number of performance traits in cattle including growth and animal size (Mullen et al., 2010; Mullen et al., 2011b; Waters et al., 2011).

1.2.3.2 Insulin-like growth factor 1

Insulin-like growth factor 1 (IGF1) is an important circulating growth factor involved in stimulating proliferation and differentiation of muscle cells (Otto and Patel, 2010; Keady et al., 2011; Schiaffino and Mammucari, 2011). IGF1 is expressed in a number of tissues including skeletal muscle, although expression is most abundant in liver (Jiang and Ge, 2013). However, in the circulation, most IGF1 is found as part of a complex with one of 6 IGF binding proteins (IGFBP), which increase the half-life of IGF1 in blood (Boisclair et al., 2001; Jiang and Ge, 2013). Over-expression of IGF1 in mice has shown increased muscle hypertrophy as well as resistance to induced muscle atrophy (Musaro et al., 2001; Schulze et al., 2005; Bonaldo and Sandri, 2013).

Additionally, binding of IGF to IGF1R activates several signalling cascades, including the MEK-ERK pathway and the PI3K-AKT pathway (Glass, 2003; Duan et al., 2010; Glass, 2010). Indeed a number of studies have implicated both of these pathways in regulating muscle size and inhibition of protein degradation (Sartori et al., 2009; Ge et al., 2013).

1.2.3.3 Myostatin

Myostatin (also called growth differentiation factor 8, or GDF8), which is a member of the transforming growth factor β (TGFβ) family of proteins, is primarily expressed in skeletal muscle and acts as a negative regulator of muscle growth (McPherron et al., 1997; Allen and Unterman, 2007). Mutations in the myostatin gene have been shown to induce increased muscle mass leading to the double-muscling phenotype (Esmailizadeh et al., 2008; Kollias and McDermott, 2008) (Figure 1.4). Double-muscling is characterised by both an increase in the number of cells (hyperplasia) and an increase the size of individual cells (hypertrophy). This phenotype, due to disruption of myostatin, has been reported in a number of species including mice (McPherron et al., 1997), cattle (McPherron and Lee, 1997) and humans (Schuelke et al., 2004). The double-muscling phenotype has been reported in many cattle breed including Friesian, however, it is most commonly observed in Belgian Blue and Piedmontese breeds (Kambadur et al., 1997).

The inhibitory effects of myostatin on muscle are mediated by binding to activin type II receptor B, which in turn leads to phosphorylation of R-Smads 2 and 3 to initiate numerous signalling cascade (Trendelenburg et al., 2009; Burks and Cohn, 2011). Myostatin treatment has been shown to block the IGF1-PI3K-AKT signalling pathways, inhibiting muscle growth and differentiation (Trendelenburg et al., 2009; Bonaldo and Sandri, 2013).

The role of myostatin, and the effect of myostatin inactivation have gained considerable interest in medical science as a potential target for therapeutic treatment of numerous muscle wasting disorders in humans (Roth and Walsh, 2004; Zhou et al., 2010; Han et al., 2013). However, in cattle doubling-muscling is often considered undesirable due to increased problems associated with calving such as dystocia, and unfavourable meat characteristics (Cuvelier et al., 2006; Allais et al., 2010; Kolkman et al., 2010).

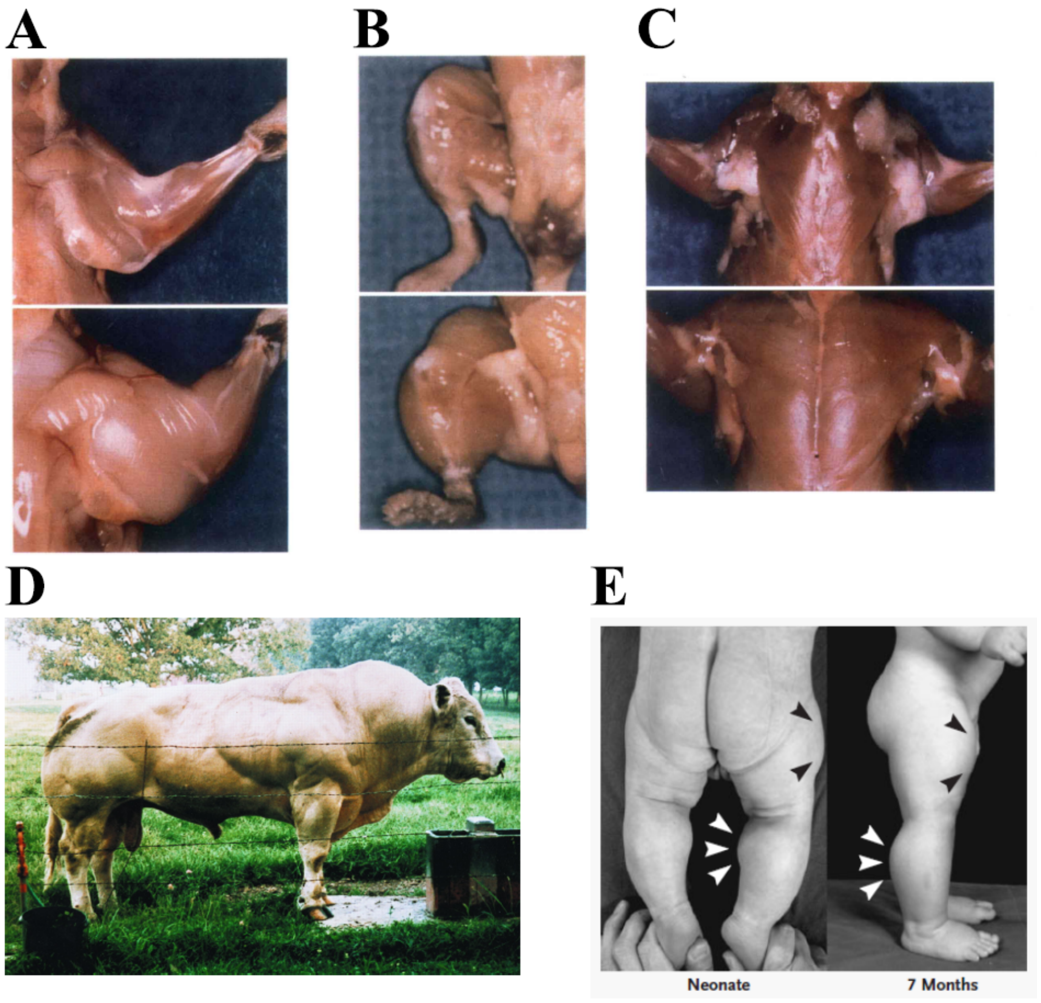


Figure 1.4 Examples of muscle hypertrophy caused by myostatin inactivation. A-C increased muscle mass in myostatin null mice (lower panels) compared to wild-type littermates (top panels) in the upper limb (A), lower limb (B) and pectoral muscles (C) of skinned animals (McPherron et al., 1997). (D) a Belgian Blue bull exhibiting the double muscling phenotype (McPherron and Lee, 1997). (E) photographs of a child at 6 days (left) and 7 months (right). Arrows indicate protruding muscles at the child's thigh and calf caused by muscle hypertrophy (Schuelke et al., 2004).

1.3 The bovine genome

1.3.1 The Bovine Genome Project

Following the publication of the human genome in 2001 (Lander et al., 2001), sequencing technologies have rapidly and dramatically improved. In the last decade, decreasing costs and increasing output associated with sequencing has culminated in the publication of full genome sequences for an increasing number of animal species (Eggen, 2012)(Table 1.3). The bovine genome was published in 2009 (Elsik et al., 2009), which was appropriately, the Chinese year of the Ox. The bovine genome contains about 22,000 genes on 29 autosomes and 2 sex chromosomes. The bovine genome was based primarily on the DNA sequence of a single female Hereford animal, L1 Dominette. The Bovine Genome Project represented a large collaborative effort between multiple groups and funding from the Australia, Canada, France, New Zealand, the United Kingdom and the United States (Burt, 2009; Elsik et al., 2009).

The unique biology of ruminants and, in particular, the importance of cattle as a major source of nutrition for humans were some of the main reasons for sequencing the bovine genome (Elsik et al., 2009; Tellam et al., 2009). Also, the bovine genome was the first complete high coverage genome sequence from the Cetartiodactyl order of eutherian mammals which first appeared approximately 60 million years ago (Burt, 2009; Tellam et al., 2009). The cattle genome sequence provides an important scientific resource for exploring the impact of genetic variation, understanding mammalian evolution and genomic studies related to cattle breeding (Elsik et al., 2009; Berry et al., 2011). Two different genome assemblies, using different approaches, have been generated based largely on the same sequence data generated from the Bovine Genome Project.

Table 1.3 Sequenced genomes for animal species. (Adapted from (Eggen, 2012)).

Species	**Genome length (assembly) (Gbp)**	**Year**
Dog (*Canis familiaris*)	2.4	2003
Chicken (*Gallus gallus*)	1.05	2004
Cat (*Felis catus*)	1.64	2006
Sheep (*Ovis aries*)	2.78	2008
Cattle (*Bos taurus*)	2.91	2009
Horse (*Equus caballus*)	2.47	2009
Pig (*Sus scrofa*)	2.2	2009
Rabbit (*Oryctolagus cuniculus*)	2.67	2009
Turkey (*Meleagris gallopavo*)	1.08	2009
Dromedary (*Camelus dromedarius*)	2.2	2011

1.3.2 Baylor College of Medicine assembly

Baylor college of medicine published their first assembly of the bovine genome in 2009 (Liu et al., 2009), prefixed as Btau. Btau4.0 was assembled by combining whole-genome shotgun reads, and bacterial artificial chromosomes sequences leading to a 2.87 Gbp assembled genome (2.91 Gbp including unplaced contiguous sequences (contigs)). In the Btau4.0 assembly, 90% of the total genome was placed on 29 autosomes and the X chromosome. The N50 size (50% of the genome is in contigs of this size or larger) for the assembly was

81.6 kbs. Although, this initial assembly did not assign any sequence to the Y chromosome, the current version (Btau4.6.1) contains 43.3 Mb of sequence assigned to the Y chromosome.

1.3.3 University of Maryland assembly

A second, independent, assembly of the bovine genome was carried out at the University of Maryland and published with the prefix UMD (Zimin et al., 2009). UMD used paired-end sequence data, mapping data and the human genome to create an assembly of 2.86 Gbps (Figure 1.5). This also included a partial assembly of the bovine Y chromosome. The current UMD assembly (UMD3.1) also has a slightly higher N50 contig size of 96.955 kbs compared to 83.68 kbs for the current Btau assembly.

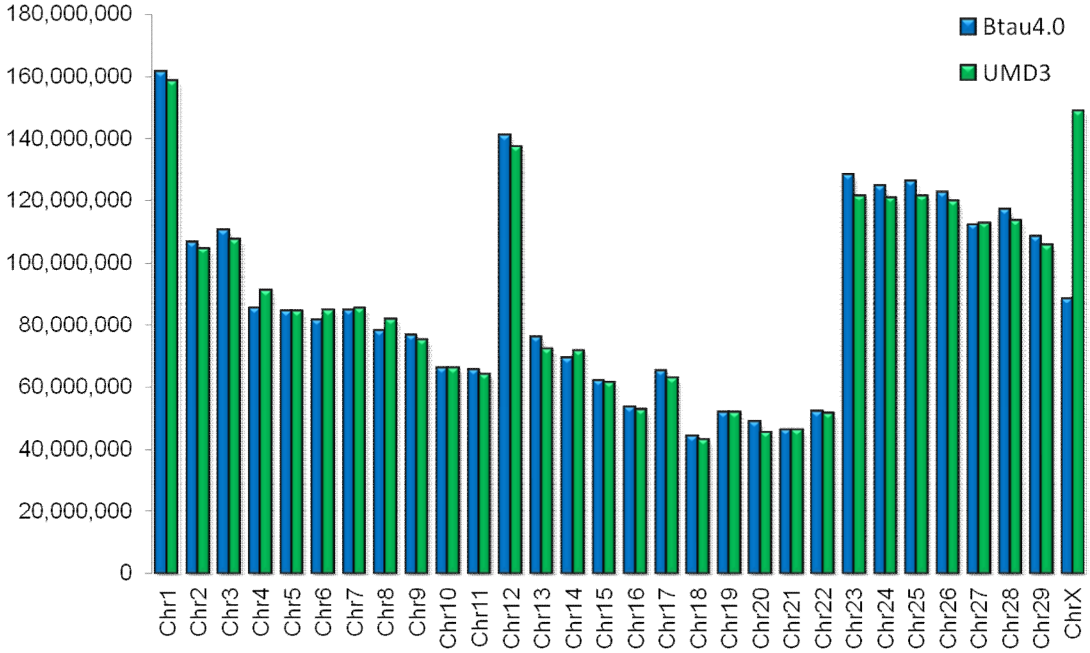


Figure 1.5 Chromosome lengths based on the UMD3 assembly of the bovine genome. Chromosome (Chr) lengths are displayed in base pairs on the y-axis. (Adapted from (Zimin et al., 2009)).

1.3.4 Gene orthology

Orthologs for more than 75% of the predicted bovine genes have been identified in dog, human, mouse or rat, opossum and platypus (Elsik et al., 2009). There are numerous examples of individual genes having similar functional impact across species (Goddard and Hayes, 2009). For example, mutations in the myostatin gene are known to cause the double muscling phenotype in humans (Schuelke et al., 2004), mice (McPherron et al., 1997) and cattle (McPherron and Lee, 1997; Martinez et al., 2010). In spite of this, across species there is little known on the conservation of genes with low to moderate effects on a phenotype (Elsik et al., 2009). However, there is growing evidence for conservation of gene classes between species (Pryce et al., 2011). For example, sequencing of the bovine genome revealed that most genes involved in metabolism are generally highly conserved in mammals, although in cattle a few gene losses compared with other mammalian species were identified (Elsik et al., 2009; Tellam et al., 2009).

Gene interaction data and pathway annotations for humans are of a higher quality and contain greater detail than bovine annotations. This is an important consideration for analyses involving bovine gene data (or non-model organism). Because of this, it is common practice to map genes from bovine to human orthologs for downstream analysis such as identifying gene interactions and affected pathways.

1.3.5 The Bovine HapMap project

One of the main goals of the Bovine HapMap project was to shed light on the genomic structure of cattle through the identification and analysis of over 37,000 single nucleotide polymorphisms (SNPs). Data from 19 different breeds of both *Bos taurus* and *Bos indicus* cattle was used to investigate breed diversity, patterns of linkage disequilibrium, population history and effective population sizes (Gibbs et al., 2009).

1.3.5.1 Variation in the bovine genome

Single nucleotide polymorphisms (SNPs) are the most common genetic variant found in vertebrates and invertebrates (Brookes, 1999; Cohuet et al., 2008). Identifying DNA sequence variation in domestic livestock species is of particular importance in breeding and genetic improvement studies (Gao et al., 2012). A by-product of the bovine genome project and subsequent HapMap project is the availability of massive numbers of genetic markers in the form of SNPs. This massive increase in marker numbers spurred the development of high-density genotyping arrays, such as the BovineSNP50 beadchip (Matukumalli et al., 2009), and has subsequently enabled scanning of the genome to identify markers associated with particular phenotypes of interest. At present, there are almost 10 million bovine SNPs reported in dbSNP (Sherry et al., 2001), although this is quite low compared to species such as mice (approximately 16 million) or humans (approximately 60 million). As such, much of the variation in the bovine genome has yet to be discovered and characterised.

1.3.5.2 Linkage disequilibrium and effective population size

In recent years effective population size in cattle has dramatically declined with the development of advanced breeding techniques used to identify elite animals that are then used to sire large numbers of progeny (de Roos et al., 2008; Gibbs et al., 2009; Muers, 2009). Although the total number of domesticated cattle continues to rise, the genetic diversity of those animals has decreased (Gibbs et al., 2009). This decline has had an effect on the pattern of linkage disequilibrium (LD) in cattle, which is the non-random association between two or more regions of DNA that occurs when they are inherited together (Khatkar et al., 2008). The pattern of LD in a species reflects historical rates of recombination between loci and evolutionary forces (Feder et al., 2012). Two commonly used measures of LD are *D'* and r^2. Each of these have different statistical properties but both range from 0 (no disequilibrium) to 1 (perfect disequilibrium) (for a review see Pritchard and Przeworski, 2001; Balding, 2006).

Unlike in species such as humans, where LD is found only up to tens of kb (Tenesa et al., 2007), LD patterns in cattle have been reported at low but nonzero levels of up to 1,000 kb (de Roos et al., 2008; Khatkar et al., 2008; Gibbs et al., 2009) (Figure 1.6). In fact, the extent of LD has been shown to be much greater in numerous livestock species compared to humans (e.g. pigs (Nsengimana et al., 2004), sheep (McRae et al., 2002; Meadows et al., 2008) and cattle (Farnir et al., 2000; Khatkar et al., 2008; Gibbs et al., 2009)). This is an important consideration for analyses in cattle that rely of the pattern of LD (such as genome-wide association studies), as the number of markers (usually SNPs) required to for sufficient coverage of the genome is substantially less than that required in humans (Matukumalli et al., 2009).

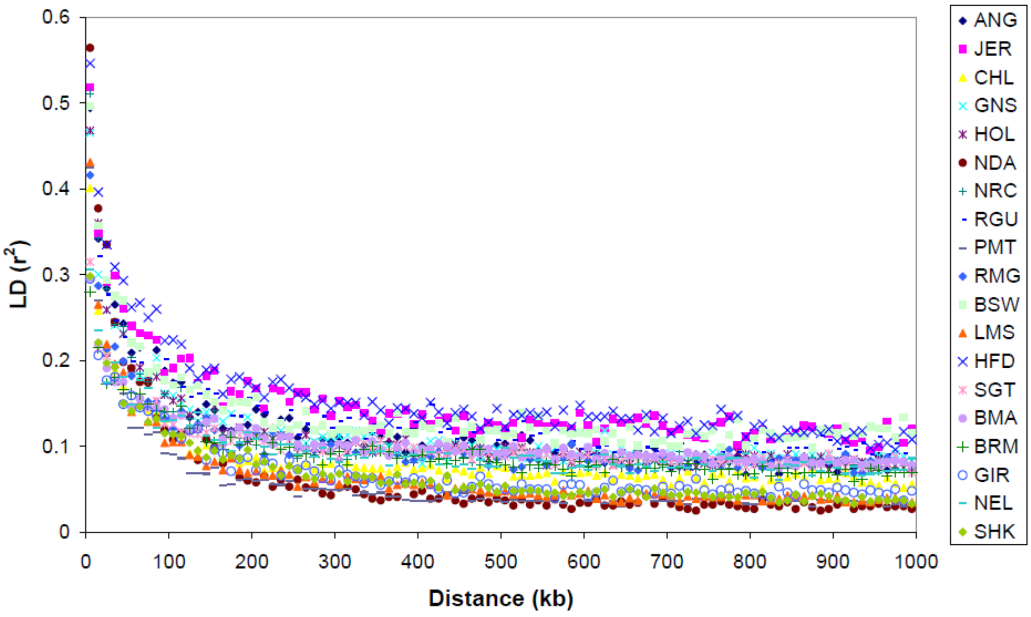


Figure 1.6 Changes in linkage disequilibrium (r^2) between marker pairs with increasing distance for all breeds analysed as part of the bovine HapMap project. Values are the genome-wide average r^2 values within 10 kb bins (Adapted from (Gibbs et al., 2009)). ANG = Angus, JER = Jersey, CHL = Charolais, GNS = Guernsey, HOL = Holstein, NDA = N'Dama, NRC = Norwegian Red, RGU = Red Angus, PMT = Piedmontese, RMG = Romagnola, BSW = Brown Swiss, LMS = Limousin, HFD = Hereford, SGT = Santa Gertrudis, BMA = Beefmaster, BRM = Brahman, GIR = Gir, NEL = Nelore, SHK = Sheko.

1.4 Systems biology

The mechanisms regulating many traits are undoubtedly complex, and as such require the integration of data from many sources to fully comprehend. Systems biology is an integrative approach whose constituent disciplines include genomics, transcriptomics, proteomics, metabolomics and bioinformatics (van Ommen and Stierum, 2002; Berry et al., 2011). Systems biology seeks to exploit the growing amount of information (particularly from next-generation sequencing) from these disciplines in order to understand the dynamic outcomes of molecular interactions at the cell, pathway and organism levels (Auffray et al., 2003; Smith and Bolouri, 2005; Bruggeman and Westerhoff, 2007). Unlike analysis of individual components of a system (such as the response of a single cell type to a single stimulus), systems biology concentrates on all of the components, the interactions between them and the resulting outcomes (Hood, 2003; Aderem, 2005; Berry et al., 2011). Approaches from the genomics, transcriptomics and bioinformatics disciplines were used in this thesis and as such are discussed in the following sections. Additional disciplines that may be exploited within the systems biology context are reviewed by Smith and Bolouri (2005) and Berry *et al.* (2011).

1.4.1 Genomics

To help us better understand biological systems, genomic technologies aim to dissect the structure, variation and intra-genomic relationships within the genome. In 1977, the first complete genome (bacteriophage φX174) was sequenced by Sanger *et al.* (1977). With this landmark paper the field of genomics emerged. Twenty-four years later, following a large collaborative effort, the first draft human genome was published and later finalised in 2003 (Lander et al., 2001; IHGSC, 2004). An increasing demand for low cost sequencing throughout the 2000s spurred the development of high-throughput massively parallel sequencers (next-generation sequencing (NGS)). Since then, the cost of sequencing entire genomes has fallen dramatically prompting the launch of the $1,000 genome challenge in 2005 (Bennett et al., 2005; Mardis, 2006; Sboner et al., 2011). These developments triggered an era of genome sequencing projects for several animal species including the bovine genome which was published in 2009 (Elsik et al., 2009). The publication of the bovine genome has provided an invaluable resource enabling the examination of genes, discovery of mutations and the testing of evolutionary hypotheses (Gibbs et al., 2009). A number of different NGS platforms are available (for a review of the various strategies see Mardis (2008), Shendure and Ji (2008), Ansorge (2009), Voelkerding *et al.* (2009), Metzker (2010)) but essentially, to sequence a genome, DNA is broken into many short fragments which are then amplified. The nucleotide bases composing these fragments are identified (called “sequencing”), and through multiple rounds of re-arranging these fragments (called “assembly”), sequences representing the DNA of each chromosome from the genome is obtained (known as “contigs”). The quality of the genome is assessed based on the number and length of the assembled contigs, usually the fewer the number and the longer the length of the contigs, the better (Schatz et al., 2010; Yandell and Ence, 2012).

1.4.2 Transcriptomics

Transcriptomics refers to the study of the transcriptome, which is the set of all RNA transcripts including mRNA, rRNA, tRNA, miRNA and small RNAs. Transcriptomics enables global understanding of the molecular changes in gene expression that controls the synthesis of proteins within the cell. Examination of gene expression within tissues can provide insights as to the genes and pathways that are functionally important in regulating animal performance. To date, most gene expression studies in cattle have been performed using microarrays (Lehnert et al., 2006; Sadkowski et al., 2009; Connor et al., 2010). However, there are several limitations with microarray technologies, such as reliance on existing knowledge about genome sequence, limited dynamic range of detection and issues surrounding reproducibility (Wang et al., 2009b).

Much of the limitations of microarrays have been overcome with the advent of NGS technologies. RNA-seq enables deep-sequencing of the transcriptome and the quantification of expression through calculation of digitally defined counts of reads that align to a transcript. RNA-seq is a relatively new approach to transcriptome sequencing but has several advantages and novel applications over microarrays (Wang et al., 2009b). For example, this technology is highly sensitive to the detection of all expressed genes, has virtually no background noise and has a higher range of detection. Also, the digital readout of sequence from RNA-seq has enabled the detection of novel transcripts, alternative splicing, transcript fusions and SNP discovery. Because of these advantages, RNA-seq has quickly been adopted by a large number of researchers and employed in the analysis of several complex traits (e.g. negative energy balance (McCabe et al., 2012), infection (Foley et al., 2012; Vegh et al., 2013) and stress response (O'Loughlin et al., 2012)).

1.4.3 Bioinformatics

Bioinformatics can be described as the use of computer algorithms to process, manage and analyse biological data. Bioinformatics covers a wide range of disciplines including software development, database design, genome annotation, mathematical modelling and graphical display of biological data (Chicurel, 2002; Hood, 2003; Melham, 2013). Bioinformatics allows the efficient and rapid analysis of huge datasets in a high-throughput manner. Bioinformatics has a synergistic relationship with the omics disciplines, in that as data generated from the omics has become increasingly large, bioinformaticians have been tasked with the job of developing efficient algorithms for the analysis of this data. Similarly, the improvement of bioinformatic techniques has enabled more difficult questions and new hypotheses to be tested (Smith and Bolouri, 2005). Within systems biology, bioinformatics plays an important role in bringing together information from various sources to address biological questions (Kitano, 2002). Examples of this can seen with the availability of numerous visualisation and analysis resources such as InnateDB (Lynn et al., 2008; Breuer et al., 2013), STRING (Franceschini et al., 2013) and various genome browsers (e.g. UCSC (Meyer et al., 2013)). A key challenge for bioinformatics will be the continued development of novel, more efficient and, importantly, user-friendly tools that can be utilised by the wider community of researchers.

1.4.4 Databases and online resources

A large number of public databases now exist, maintaining information on genetic variation, genetic association, DNA sequences, gene interactions and expression data. Centralised databases and resources, with standardised storage structures, are essential to systems biology (Kitano, 2002; Cassman, 2005). Many of these resources form the basis of much research, facilitating analyses that would not be possible without them. Humans are the most represented organism in many databases, however the information available for many non-model organisms has rapidly grown in recent years. A detailed review of many databases and resources available to aid the analysis of both model and non-model organisms is presented by Helmberg (2012). A brief description of the resources and databases utilised in this thesis are outlined in the following sections.

1.4.4.1 The Single Nucleotide Polymorphism Database

The Single Nucleotide Polymorphism Database (dbSNP) (Sherry et al., 2001) is a public repository for molecular variation. Although the name dbSNP suggests that only SNPs are contained within the database, the term SNP was chosen as shorthand for variation, thus information on SNPs, short insertions and deletions, microsatellite markers and other forms of genetic variation are also represented (Sherry et al., 1999). Submissions from both public and private sources are accepted. As variation at the same location in the genome can be submitted from multiple sources, two types of records are maintained in dbSNP; submitted SNPs (ss accession number) and reference SNPs (rs accession number). An ss number is assigned to all submitted variations, and an rs number to all unique variation in an organisms' reference genome. Variation for approximately 100 organisms (including cattle) is currently maintained by dbSNP and this number continues to grow as more organisms' genomes are sequenced.

1.4.4.2 Ensembl

Ensembl is a joint project of the Wellcome Trust and EMBL-EBI, which provides genome resources and a genome browser (Flicek et al., 2012). Similar databases are available through the National Centre for Biotechnology Information (NCBI) (Jenuth, 2000) and University of California, Santa Cruz (UCSC) (Karolchik et al., 2013). The Ensembl database contains reference sequences and genome annotations for over 50 species including evidence-based gene annotations and comparative genomics resources such as homology, orthology and paralogy relationships. In addition, support for user data upload and visualisation is provided for BAM, GTF and VCF file formats (among others).

1.4.4.3 Kyoto Encyclopaedia of Genes and Genomes

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa et al., 2012) is an online database consisting of 15 sub-databases representing functional information (such as pathways), genomic information (such as genes) and chemical information (such as small molecule compounds). KEGG Pathway links known information on molecular interactions for human diseases, metabolic, cellular and organismal processes in manually drawn pathway maps. Users are provided with a visualisation tool linking the components of biological pathways that are present in the system under investigation. KEGG's pathway mapping tool can be used to connect molecular data (such as that derived from sequencing) to biological function, enabling improved understanding and interpretation of experimental outcomes.

1.4.4.4 Animal Quantitative Trait Loci database

The Animal Quantitative Trait Loci database (Animal QTLdb) (Hu et al., 2013) is an open repository housing publicly available QTL and genetic association data on numerous livestock species. Currently, information is stored for cattle (Cattle QTLdb), chicken (Chicken QTLd), pig (Pig QTLdb), sheep (Sheep QTLdb) and rainbow trout (Rainbow Trout QTLdb). Tools are also provided to enable comparison of QTL across within and across species. Data can also be retrieved in several formats including GFF3 and SAM. Since the inception of Animal QTLdb in 2005 (Hu et al., 2005), its popularity has steadily grown and there is now information on over 20,000 QTLs for over a thousand traits in 5 species contained within the database (Hu et al., 2013).

1.4.4.5 Gene Expression Omnibus

Gene Expression Omnibus (GEO) (Barrett et al., 2013) is a public repository containing functional genomics data derived from microarray, next-generation sequencing and other forms of functional genomics experiments. GEO hosts the raw data, processed data and metadata derived from gene expression, gene regulation and epigenomics experiments. GEO also provides tools to search for, identify, analyse and visualise data. Additionally, programmatic access to data contained within GEO is supported. Furthermore, a number of journals publishing studies of gene expression require that relevant data are made available through online repositories such as GEO.

1.4.5 Systems biology and animal science

Integrating the ‘omics’ technologies and bioinformatics under the systems biology umbrella is a promising strategy in animal science (Kadarmideen et al., 2006; Berry et al., 2011; Eggen, 2012). By utilising a more holistic approach, systems biology is capable of facilitating new understanding of the genetic control of economically important phenotypes (D'Alessandro and Zolla, 2013). In livestock species, systems biology is particularly appealing to quantitative geneticists and breeders. Systems biology could fundamentally change the way in which animal breeding is practiced, moving away from the traditional “black box” approach, where very little is known about the underlying genetic mechanism, toward an approach where the genetic regulatory networks and biological pathways underlying changes in the phenotype are considered (Zhu et al., 2009; Berry et al., 2011; Cole et al., 2013; Snelling et al., 2013).

1.5 Genome-wide association study

In very general terms, the aim of any genetic association study is to identify associations between a phenotypic variable and a genetic variable (Konig, 2011). Phenotypic variables may be binary outcomes, for example a disease status being infected or not infected (yes or no), or quantitative such as height (Visscher, 2008; Lango Allen et al., 2010). Genetic markers where the alleles and the location in the genome are known are used as the genetic variable. Most association studies utilise SNPs as genetic markers. Although these SNPs may not be responsible for the observed variation in a phenotype, with a sufficient density of SNPs throughout the genome, the effect of an unknown causal variant may be detected due to LD between the known SNP and the unknown causal variant (Balding, 2006). A genome wide association study (GWAS) is a large scale association study that assays genetic variants across the entire genome (Ramanan et al., 2012). In this way, GWAS can be a powerful tool for identifying candidate regions (or quantitative trait loci (QTLs)) harbouring genes or mutations influencing outcomes in a phenotype of interest (Manolio et al., 2009; Gao et al., 2012). For example, by examining genes that are found within candidate regions, important insights into the functional mechanisms underlying the trait being investigated can be elucidated. This can then lead to better understanding of the molecular processes that affect variation within the trait of interest (Wang et al., 2007; Holmans et al., 2009; Fontanesi et al., 2012). However, in order to detect the polymorphisms with small effects which underlie a complex trait, a large number of samples are needed to maintain statistical power (Balding, 2006; Zhang et al., 2010).

In the last number of years GWAS has become a popular tool for investigating the underlying genetic architecture associated with several phenotypes. For example, in human studies over 1,300 GWAS studies have been published since 2005 (Ramanan et al., 2012; Hindorff et al., 2013; Manolio, 2013) (Figure 1.7). Increasingly, GWAS are viewed as an initial screening tool to enable targeted downstream analysis for discovering genes and pathways relevant to a phenotype (Korte and Farlow, 2013). In cattle, the development of high density genotyping arrays and a decrease in costs of genotyping large pools of animals have made possible scanning of the entire genome to identify regions of the genome associated with phenotypes of interest.

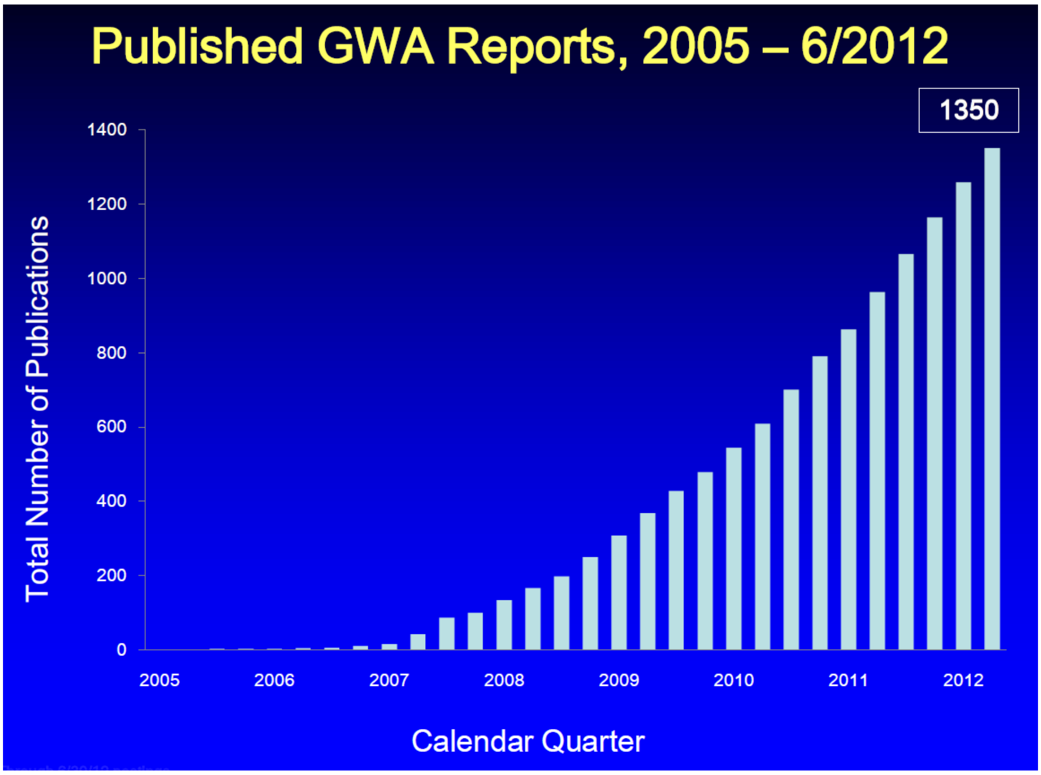


Figure 1.7 The number of published GWA studies in humans from 2005 – 2012. Adapted from (Manolio, 2013).

1.5.1 Single SNP regression

GWAS allows the examination of a large number of genetic markers distributed across the entire genome to detect variants associated with a trait of interest. Most association studies to date have used single SNP regression approaches to quantify the association between SNPs and a quantitative phenotype (Ziegler et al., 2008). Linear mixed model (LMM) approaches analyse each SNP separately (include a single SNP at a time) for an association with a phenotype allowing for the inclusion of fixed and random effects within the model. In this type of analysis, the null hypothesis is that there is no association between the SNP and the trait (dependent variable). SNPs can be included as continuous variables, in which case an allelic effect will be estimated. Importantly, linear models in GWAS assume a linear relationship between the trait and genotype as well as a common variance at each genotype, which is not always a realistic assumption. Additionally, as many thousands of SNPs are included in a GWAS, errors arising from many thousands of individual tests must be controlled for (Balding, 2006).

1.5.1.1 Correcting for thousands of tests

Typically in GWAS, many thousands of SNPs are investigated for an association with a trait. As genotyping arrays have increased the number of SNPs assayed on a single array, the number of SNPs that can be included in an association study has grown dramatically. This poses a major challenge for GWAS. Not only is it computationally challenging to analyse such large numbers of SNPs in a reasonable amount of time but also, as many thousands of individual test are performed, controlling the number of false positives reported has become an important procedure in GWAS. Anytime the null hypothesis is rejected (i.e. the p-value is less than a certain threshold, commonly $p<0.05$), it is possible that this significant result was due to chance alone. A case in which the null hypothesis is rejected when in fact it should be accepted is called a false positive (or type I error). Conversely, when the null hypothesis is accepted when in fact the null hypothesis should be rejected is known as a false negative (or type II error). Several correction algorithms for controlling multiple testing errors are regularly employed in GWAS. The most common correction methods are Bonferroni, Benjamini-Hochberg and the q-value which are discussed below.

1.5.1.2 Bonferroni

Bonferroni correction is a commonly used approach to the multiple comparison problem, which seeks to control the number of type I errors (false positives). Bonferroni correction controls the family-wise error rate (FWER), which is the probability of making one or more false discoveries among all tests. For a given significance threshold, α, a test is significant (i.e. the null hypothesis of no association is rejected) if the p-value multiplied by the total number of tests is less than α. This ensures that the probability of rejecting at least one hypothesis when all are true is no greater than α. This approach assumes independence of SNPs, which is unlikely as many SNPs included in an analysis are likely to in LD with each other (Gao et al., 2010; Johnson et al., 2010). As a consequence, the probability of type II error rate (false negatives) is inflated as the number of tests increases (Streiner and Norman, 2011). In general, as the number of tests increases, statistical power to detect true positives decreases when controlling FWER (Storey, 2002).

1.5.1.3 Benjamini-Hochberg

An alternative approach to controlling the FWER is to control the false discovery rate (FDR). This approach allows a proportion of false positives to be contained within all significant results, and thus is much less conservative than the Bonferroni approach. Also, for the same set of p-values, the number of false negatives will be less than that of the Bonferroni approach. A commonly used FDR approach is the Benjamini and Hochberg (BH) (Benjamini and Hochberg, 1995). Although a similar approach was originally proposed by Simes (1986), Benjamini and Hochberg (1995) developed the proposal of Simes in greater detail. In order to perform the BH correction, all p-values are placed in ascending order and ranked from 1 (smallest) to the total number of tests, *k* (largest). A threshold for the study is then chosen (α) at which the FDR will be controlled. For each p-value a critical threshold is calculated by dividing its rank by the total number of tests and multiplying the result by α. A p-value is significant after BH correction if it is smaller than its critical threshold. All p-values less than the highest ranked that is significant are also significant. BH is less conservative than the Bonferroni correction, especially when many correlated tests are involved.

1.5.1.4 The q-value

Another approach for correcting for errors arising from multiple testing is the q-value (Storey, 2002; Storey, 2003; Storey and Tibshirani, 2003). The q-value is an extension of the FDR approach (Benjamini and Hochberg, 1995), and is analogous to the traditional p-value obtained from FDR correction methods. The q-value is a particularly useful statistic in extremely large datasets where dependence between tests may exist, such as in a GWAS (Storey, 2002; Storey and Tibshirani, 2003). This approach controls the number of significant results that will be false positives, unlike p-value methods that control the number of tests that result in false positives. For example, given 1,000 tests in which a p-value threshold of 0.05 is chosen. This means that 5% of all tests (50) will result in a false positive, whereas a q-value of 0.05 means that 5% of significant results will be false positives. So if 100 of the 1,000 tests are significant (either a p-value < 0.05 or a q-value < 0.05), a p-value of 0.05 means that up to half (50) of these may be false positives, whereas a q-value of 0.05 for the same results means that up to 5 results may be false positives. Storey and Tibshirani (2003) demonstrated that the q-value methodology also controls the number of false negatives more appropriately than the BH approach resulting in greater power, especially in scenarios where dependence between tests may exist such as in genome-wide studies.

1.5.2 Bayesian inference

Single SNP regression approaches analyses each SNP individually for association with the phenotype under investigation (McCarthy et al., 2008). However, many traits are complex in nature, and likely to be influenced by many genes. Models that analyse all SNPs simultaneously should provide more accurate results than models that analyse one or a few markers at a time (van den Berg et al., 2013). A Bayesian approach is advantageous compared to SSR as there is no need to correct for errors arising from many thousands of tests. Additionally, Bayesian approaches allow the incorporation of prior information which may be used parameter estimation. Bayes theorem (Equation 1.1) is used to combine prior beliefs with information from the data for inference. This may be defined as follows:

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)} \propto p(y \mid \theta)\, p(\theta) \qquad [1.1]$$

This equation calculates the posterior distribution of hypothesis *θ* (i.e. estimated SNP effect) given data *y* (i.e. prior information such as the genotypic data and phenotypic data). The denominator, p(*y*), is a normalising constant that is usually computationally intensive to calculate and can be dropped, which changes the relation of the posterior distribution from equals to proportional to. This results in a posterior distribution that is proportional to the probability of *y* given *θ* (note that this is the likelihood of *θ* given data *y*) times the prior probability of *θ*.

Markov Chain Monte Carlo (MCMC) sampling methods can then be used to provide an approximation of the true posterior distribution. The Markov chain, if properly constructed and after a sufficient amount of time, will eventually converge to a region of the posterior and remain in that region (Oszkiewicz et al., 2012)

Up until a few years ago, Bayesian approaches were difficult to implement mainly due to computational constraints (Robert and Casella, 2011). Nowadays, Bayesian approaches are routinely applied in many research fields including computational biology (Feng et al., 2011; Oldmeadow and Keith, 2011) and phylogenetic inference (dos Reis and Yang, 2011; Drummond et al., 2012). Bayesian inference approaches have also been applied to domestic livestock in genetic prediction studies (e.g. Bolormaa et al., 2011; Fan et al., 2011; Meredith et al., 2012; Saatchi et al., 2013).

1.5.2.1 Convergence assessment

Convergence assessment relates to the idea that a Markov chain, given a sufficient amount of time, will eventually convergence to a stationary distribution from which posterior inference can be made (Oszkiewicz et al., 2012). As early iterations in a Markov chain are often influenced by starting values, which are values used to initiate the algorithm (usually referred to as seeds), an initial portion of the Markov chain should be removed so as to reduce potential bias or pollution of posterior inferences caused by the starting values (Dodds and Vicini, 2004). Thus, all iterations before convergence should be discarded as “burn-in”. Once convergence has been achieved, all further samples should be concentrated from around the mode of the target distribution. Convergence assessment is not a new element in Bayesian inference, and several approaches have been extensively reviewed (see Cowles and Carlin, 1996; Brooks and Gelman, 1998; Brooks and Roberts, 1998).

However, implementation and interpretation of diagnostics for assessing convergence is often difficult requiring custom scripting and arbitrarily defined cut-offs. Graphical approaches are commonly employed as they can be easily implemented in different Bayesian approaches, and can provide an easy way to diagnose problems with lack of convergence (Nylander et al., 2008). Graphical approaches, such as trace plots (also called history plots), entail plotting a sampled parameter by the iteration number. In this way convergence can be identified by a plateau in the estimated parameter as the number of iterations increases. Importantly, no one diagnostic or metric can provide assurance of convergence, and in general several metrics should be used to assess convergence of all parameters and not just those of interest (Cowles and Carlin, 1996).

Despite the establishment of convergence diagnostics in several research fields (e.g. Nylander et al., 2008; Oszkiewicz et al., 2012), the use of convergence diagnostics in genetic prediction of livestock species is still relatively uncommon. In fact several authors fail to mention whether or not convergence was even assessed (e.g. Olsen et al., 2011; Purfield et al., 2013). As a consequence, the length of the burn-in and the total number of iterations to be sampled is often chosen *a priori*.

1.5.3 GWAS for muscle growth in cattle

One of the main aims of GWAS is to improve understanding of the genetic control of economically and biologically important traits. Unlike studies in humans, where association analysis have the primary purpose of identifying markers for disease (e.g. Klein et al., 2005), identification of QTL associated with a particular trait in livestock is primarily for the purpose of genomic prediction to improve selection for economically important traits (Goddard and Hayes, 2007; Zhang et al., 2013b). Most association studies involving cattle have focused on milk production, such as milk yield (Bagnato et al., 2008; Jiang et al., 2010; Pryce et al., 2010). In spite of this, a number of studies in cattle have identified associations between growth traits and regions of the bovine genome. QTL associated with growth traits have been reported on chromosomes 2, 3, 5, 6, 14, 20 and 29 (Machado et al., 2003; Gutierrez-Gil et al., 2009; McClure et al., 2010; Snelling et al., 2010; Bolormaa et al., 2011; Kim et al., 2011).

1.6 Thesis objectives

Bovine skeletal muscle is a tissue of significant economic importance to the global economy. In spite of this, relatively little is known about the mechanisms regulating different aspects of bovine muscle growth and development. In particular, much of the underlying genomic architecture involved in (i) several growth traits for selection (ii) the acute responses of muscle tissue to nutrient restriction and re-alimentation and (iii) the long term evolutionary changes differentiating dairy and beef breeds, has yet to be elucidated.

Technological advances in recent years, such as in sequencing, have facilitated the generation of vast amounts of information. This has enabled a remarkable rise in research pertaining to increasingly complex questions. In conjunction with this rise, new methods and approaches have been continually developed to aid the analysis of such large data. Systems level approaches are concerned not only with the analysis of increasingly large datasets, but also the development of novel approaches to aid these analyses. With this in mind, the aim of this thesis is to develop tools and approaches to assist bovine research, and to examine the mechanisms underlying bovine muscle growth.

In recent years, SNP discovery has become highly automated and relatively cheap. As such, vast amounts of SNPs are routinely identified in both model and non-model organisms. However, annotating such large amounts of genetic variants is often challenging, particularly for SNPs identified in non-model organisms. There are currently few tools available that are species non-specific or support non-model organism data. In chapter 2, the development and implementation of SNPdat, a software tool for the annotation of SNPs found in any organisms with a reference sequence and annotation (including those in draft status) is described. Using a bovine SNP dataset, the results and annotations provided by SNPdat are demonstrated.

In Ireland, four traits related to carcass performance have been identified as economically important; carcass weight, carcass fat, carcass conformation of progeny and cull cow carcass weight. In chapter 3, a genome-wide association study was undertaken to identify regions of the bovine genome associated with each trait. Using genotypic and phenotypic data from 1,061 Holstein-Friesian animals, two separate statistical approaches, a frequentist and a Bayesian, were used to estimate marker associations. Regions surrounding significant associations where then examined to identify key genes and pathways involved in each trait.

Bayesian approaches in genetic association studies have grown in popularity in recent years. Assessing convergence is an important aspect of Bayesian inference. However, the use of diagnostics to assess convergence in Bayesian models used for genetic prediction of complex traits in domestic livestock remains unexplored. Chapter 4 pertains to the development of methods to assess convergence in a Bayesian model used for genetic prediction of complex traits. A number of metrics are put forward that could be used to assess convergence in studies that use such models.

Compensatory growth is a complex response to increased energy availability following prolonged periods of energy restriction. This phenomenon is commonly observed across many species, however very little is known about the genes or biological pathways controlling this response. In chapter 5, the muscle transcriptome of cattle undergoing nutritional restriction and subsequent compensatory growth is examined with the objective of identifying key genes and biological pathways involved in restricted and rapid muscle growth.

Since the domestication of cattle and the subsequent emergence of the breed concept, cattle have been subjected to intense selection for milk and meat production. Consequently, key genes involved in bovine growth may be evolving at different rates in beef animals compared to animals used primarily for dairy production. The objective of the study discussed in chapter 6 was to identify genes, from a candidate list of 200, evolving at different rates, and which may be under evolutionary selection pressure, in beef compared to dairy animals.

Chapter 2: SNPdat: Easy and rapid annotation of results from *de novo* SNP discovery projects for model and non-model organisms

2.1 Introduction

Single nucleotide polymorphisms (SNPs) are the most common genetic variant found in vertebrates and invertebrates (Brookes, 1999; Black et al., 2001; Cohuet et al., 2008). At present there are over 60 million reported SNPs within the human genome which account for approximately 90 percent of all variants detected (Kruglyak and Nickerson, 2001; Sherry et al., 2001). In fact, there are over twenty organisms in the Ensembl (Flicek et al., 2012) database with at least one known SNP in the dbSNP (Sherry et al., 2001) database (Table 2.1). Consequently, SNPs are regularly utilised as the favoured molecular marker in association studies (WTCCC, 2007), genetic mapping (Hoskins et al., 2001) and population genetics (Tishkoff and Verrelli, 2003).

In recent years, next-generation sequencing technologies have dramatically increased the output of sequencing information that can be garnered from a single sequencing project. As well as increasing the raw information output, the costs associated with sequencing have fallen considerably. One of the main uses of next-generation sequencing is to discover variation among populations of related samples (Danecek et al., 2011). Consequently, increased throughput and reduced costs have allowed researchers to identify thousands of mutations, including rare variants, with potential influence on phenotypic variation (Altshuler et al., 2000; Mullen et al., 2012). More frequently non-bioinformatics researchers are required to perform analysis of increasingly large datasets.

Disease susceptibility, agriculture and evolution are among the areas concerned with understanding the influence SNPs have on biological function and phenotypic variation of complex traits (Allan and Smith, 2008; Corona et al., 2010; Cutter and Choi, 2010). However, annotating large numbers of SNPs with this type of information can prove daunting and impractical to perform manually.

A number of bioinformatics tools for SNP annotation already exist (e.g. SNPit (Shen et al., 2009), SNPnexus (Dayem Ullah et al., 2012), Snap (Li et al., 2007), Snat (Jiang et al., 2011), SNP Function Portal (Wang et al., 2006), SNPper (Riva and Kohane, 2002), Fans (Liu et al., 2008), FunctSNP (Goodswen et al., 2010), Annovar (Wang et al., 2010)). Although there are over 50 reference sequences for eukaryotic species available from Ensembl (release 65) (Flicek et al., 2012), there are currently only a small number of tools that enable analysis of non-human SNP data (e.g. Snat, Fans, FunctSNP, Annovar). Many tools that are more general can only analyse species with SNP information in dbSNP and some require that the SNPs being annotated already exist in dbSNP. For example, FunctSNP allows the analysis of any organism contained in dbSNP. However, only SNPs that have been provided with dbSNP reference identifiers (rs#) and exact position can be processed. This limitation means that *de novo* SNP analysis is not possible in these cases. Several tools try to circumvent this problem by returning information for known SNPs surrounding the unknown. This works well for densely sampled species like humans but is not a viable option for almost all other species as SNP annotations are currently poor compared to humans (Table 2.1).

To address the above limitations and to facilitate analysis of many species not contained in SNP databases, we have developed SNPdat (**SNP D**ata **A**nalysis **T**ool) (Doran and Creevey, 2013) (Electronic Appendix 2.1). SNPdat is specifically for use with organisms which are not supported by other tools and may have a small number of annotated SNPs available, but can equally be used to analyse datasets from organisms which are densely sampled for SNPs. SNPdat is freely available on the web at http://code.google.com/p/snpdat/. This software was designed with the goal that it must be efficient, flexible and easily incorporated into users' analyses.

Table 2.1 The number of SNP annotations (ss#) in dbSNP for species with a reference sequence available from Ensembl and at least one SNP annotation in dbSNP (build 137).

Species	**Annotations in dbSNP**
Homo sapiens (Human)	60480978
Mus musculus (Mouse)	15721131
Pongo abelii (Orangutan)	10016093
Bos taurus (Cow)	9587248
Rattus norvegicus (Rat)	5227114
Canis familiaris (Dog)	3328578
Gallus gallus (Chicken)	3295452
Macaca mulatta (Macaque)	3041918
Taeniopygia guttata (Zebra Finch)	1751345
Pan troglodytes (Chimpanzee)	1660250
Danio rerio (Zebrafish)	1441888
Ornithorhynchus anatinus (Platypus)	1319269
Monodelphis domestica (Opossum)	1194131
Equus caballus (Horse)	1163580
Tetraodon nigroviridis (Tetraodon)	903110
Sus scrofa (Pig)	566003
Felis catus (Cat)	327037
Caenorhabditis elegans (C. elegans)	331438
Meleagris gallopavo (Turkey)	9256
Gadus morhua (Cod)	2140
Gasterosteus aculeatus (Stickleback)	1644
Callithrix jacchus (Marmoset)	10
Gorilla gorilla (Gorilla)	5

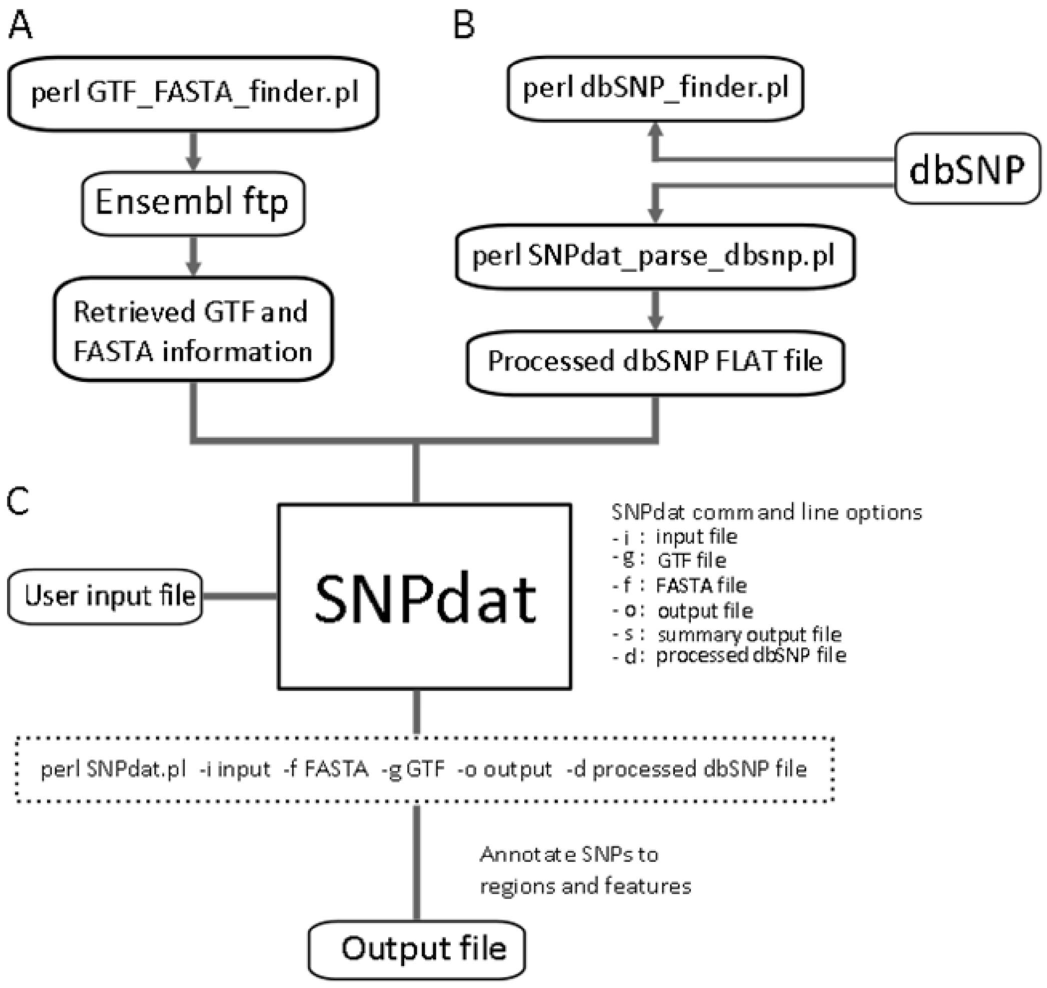
2.2 Software Implementation

SNPdat is a cross-platform command line tool written in Perl, allowing easy incorporation into existing SNP discovery or annotation pipelines or even run by a user on a standard desktop machine. SNPdat runs on all operating systems that support recent versions of Perl, including Linux, MacOSX and Windows. Scripting languages such as Perl are particularly powerful at text handling, which is an important consideration when dealing with large data files such as those generated from sequencing. Because of this, SNPdat can provide a rapid and comprehensive annotation of both novel and known SNPs for any organism with a draft sequence and annotation.

Many available tools require the user to create a local database before SNP annotation can be performed (e.g. FunctSNP, Snat, Annovar, SNPper). However, this process is not practical in all cases or straightforward enough for inexperienced users. For example, to perform SNP annotation using FunctSNP, users must first supply a list of Uniform Resource Locators (URLs) linked with online resource data files and then download them. They must then decompress any of these files matching specific suffixes, convert the data to SQL format to be imported to a SQLite database. This is time consuming and difficult for users inexperienced in bioinformatics to annotate even one SNP.

Additionally, some tools (Annovar, Snat) involve a number of pre-processing steps to parse and reformat either sequence or annotation files. This can be a difficult and confusing step for novice users, especially when dealing with non-model organisms. SNPdat does not require the creation of any local relational databases or pre-processing of any mandatory input files. Figure 2.1

contains an overview of how to use SNPdat and any additional scripts that are included in the SNPdat package.


A
perl GTF_FASTA_finder.pl
Ensembl ftp
Retrieved GTF and FASTA information
B
perl dbSNP_finder.pl
dbSNP
perl SNPdat_parse_dbsnp.pl
Processed dbSNP FLAT file
C
User input file
SNPdat
SNPdat command line options
- i : input file
- g : GTF file
- f : FASTA file
- o : output file
- s : summary output file
- d : processed dbSNP file
Annotate SNPs to regions and features
Output file

```
perl SNPdat.pl  -i input  -f FASTA  -g GTF  -o output  -d processed dbSNP file
```

Figure 2.1 Overview for using SNPdat and additional scripts available. (**A**) Retrieval of GTF and FASTA information using GTF_FASTA_finder.pl. (**B**) Retrieval and processing of data from dbSNP using dbSNP_finder.pl and SNPdat_parse_dbSNP.pl. (**C**) Command line options used to specify input/output files for SNPdat.

2.2.1 File Formats

SNPdat requires only three input files; a variant call formatted (VCF) file or a simple tab delimited text file (containing chromosome ID, genomic location and the mutation for each SNP to be analysed) as the SNP input file, a reference FASTA formatted sequence file for the species of interest, and a gene annotation file in GFF/GTF format (Table 2.2). The variant call format is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants (Danecek et al., 2011). Many SNP discovery tools, such as the Genome Analysis Toolkit (GATK) (McKenna et al., 2010), now report SNPs in VCF format. FASTA format is the standard format for representing nucleotide sequences in bioinformatics. The format is a simple text-based format in which nucleotides are represented using a single letter code (A, T, C, G). GTF files are a standard format for storing information on gene structure. GTF files define genomic structures as features. Typical features include coding sequences (CDS), exons, start and stop codons. Additional features may include untranslated regions (UTRs), introns and microRNAs. Both FASTA and GTF files are available from Ensembl for over 50 eukaryotic species.

Optional files include a processed file of SNP information from other databases such as dbSNP. SNPdat uses the extra information provided by this file to cross reference *de novo* SNPs against known annotations. Separate scripts are provided to automate the retrieval and format the data for any organisms with SNP information in dbSNP.

Table 2.2 GTF file format specification.

Column Title	**Description**	**Example**
seqname	The name of the sequence. Must be a chromosome or scaffold	Chr25
source	The source of the annotation. Typically the program that generated this feature	curated
feature	The name of this type of feature. Some examples of standard features types are "CDS", "exon"	exon
start	The starting position of the feature in the sequence. The first base is numbered 1	286859
end	The ending position of the feature (inclusive)	287050
score	For annotations that are associated with a numeric score (for example, a sequence similarity), this field describes the score. The score units are completely unspecified, but for sequence similarities, it is typically percent identity. Annotations that do not have a score can use "."	.
strand	For those annotations which are strand-specific, this field is the strand on which the annotation resides. It is "+" for the forward strand, "-" for the reverse strand, or "." for annotations that are not stranded.	+
frame	For annotations that are linked to proteins, this field describes the reading frame of the annotation on the codons. It is a number from 0 to 2, or "." for features that have no phase.	2
Attributes	An attribute list that must begin with the two mandatory attributes gene_id and transcript_id. All attributes are separated by semi colon and a single space	gene_id "ENSBTAG00000016571"; transcript_id "ENSBTAT00000022045"; protein_id "ENSBTAP00000022045";

Additional scripts which automate the retrieval of GTF, FASTA and dbSNP information are described in the following sections and are available from the SNPdat webpage (http://code.google.com/p/snpdat/).

2.2.2 Retrieval of GTF and FASTA information

For many researchers that are not familiar with bioinformatics pipelines, knowing how, and where, to get sequence and annotation information can prove difficult. An additional script (GTF_FASTA_finder.pl) (Electronic Appendix 2.2) is provided to retrieve FASTA and GTF information for any of the organisms in Ensembl (Figure 2.1A). This is written in Perl but uses the system call cURL to retrieve the information from Ensembl. This script requires and internet connection. cURL is a part of most Linux distributions and Mac OSX and can also be provided for windows through cygwin, which is a collection of tools that provide a Linux-like environment for windows. This script is interactive; when run it prompts the user to select a release of Ensembl followed by an organism in that release. The GTF and FASTA files for that organism will be downloaded to the directory from which the script is run. Alternatively, GTF and FASTA information can be retrieved manually via the Ensembl website. SNPdat also works with genomic annotations from sources other than Ensembl as long as they are provided in GTF format. This includes the results of computationally derived annotations of *de novo* genomic assemblies or transcriptomes.

2.2.3 Retrieval of information from external databases

Another script (dbSNP_finder.pl) (Electronic Appendix 2.3) was also developed to aid retrieval of information from the dbSNP database. This script retrieves SNP information for any organism contained in dbSNP (Figure 2.1B). This script also uses the cURL system call and requires a connection to the internet. Once run, the user is prompted to select an organism from all of those currently with SNP information in dbSNP. The SNP information is then retrieved for that organism. SNP information from dbSNP can also be downloaded manually from the dbSNP ftp site. When dbSNP information has been retrieved, an additional script (SNPdat_parse_dbsnp.pl) can be used to convert the dbSNP file into a format suitable for use with SNPdat.

2.3 Running SNPdat

To run SNPdat, the user specifies the input/output files and desired options with a single command (Figure 2.1C). In the case of malformed commands, SNPdat will print an error message to the screen and a short example of how the correct command should look. To ensure ease of use, SNPdat does not require the user to install any additional packages or modules and only uses modules included in the core installation of Perl.

Initially SNPdat reads the annotation information into memory from the GTF file. Each SNP is checked for errors such as non-numeric SNP locations and any warnings are printed to the output. All chromosome names provided by the user are compared against the annotation file. A warning message is printed to the output file for every SNP location or chromosome provided which does not exist in the annotation. Once all SNPs have been parsed, SNPdat will read the FASTA file one chromosome at a time. SNPdat will then perform annotations for all SNPs from the input file that are found on the current chromosome of the FASTA. Figure 2.2 contains an overview of each of the possible locations that a queried SNP can be found in. To save on memory usage and time, any chromosomes that do not appear in the list of queried SNPs are skipped.

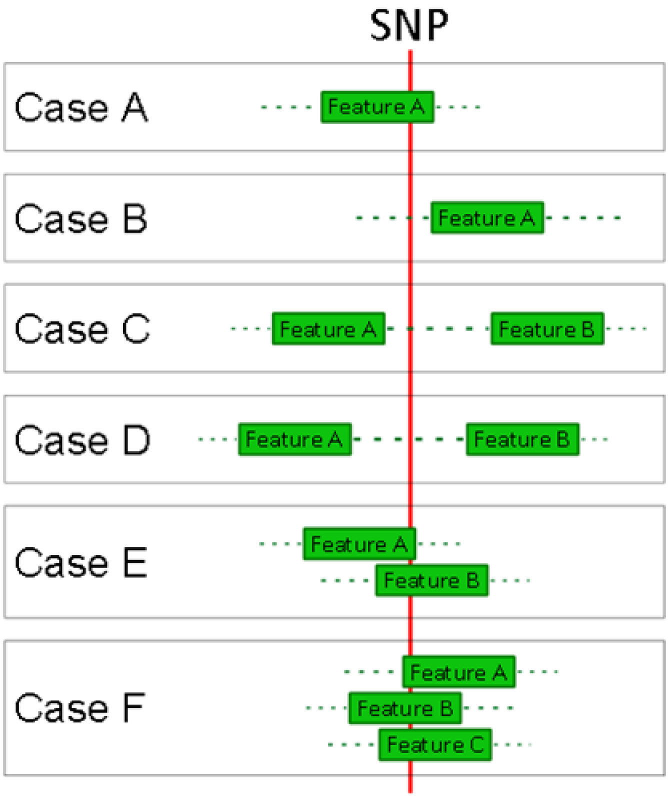


Figure 2.2 Each of the possible annotation cases that SNPdat can handle. A SNP location is shown as the red line. All annotations are relative to the SNP location. (**A**) The SNP is found in a single feature. Any SNPs in case A will be annotated to a single feature. (B) The SNP is intergenic, but only has one feature that is closest. In this case all annotations for the SNP are relative to the closest feature. (C) The SNP is intronic, all annotations for the SNP are relative to the closest feature for the gene containing both features. (D) The SNP is equidistance from 2 or more features; in this case the SNP information is retrieved for each feature and returned on a separate line in the output file. (E) The SNP is found in two different features. SNP information for each of the features is retrieved and reported on separate lines of the output file. These lines of the output file will also report the number of features that a SNP was found in. (F) The SNP is found in 3 or more features. SNP information for each of the features is retrieved and reported on separate lines of the output file.

Output from SNPdat is presented in an easily accessible tab-delimited format containing up to 25 columns of information on each SNP queried. SNPdat returns information on the genomic location of each SNP queried, including information on the distance to the nearest coding regions and other annotated sequence features, what those features are and where they start and finish (see Table 2.3 for more details). SNPdat contains algorithms for estimating information when not provided either the genome file or the annotation file such as the total number of exons for each transcript containing a SNP; the estimated reading frame, using the number of stop codons in each reading frame as a proxy (Figure 2.3), whether or not the region containing the SNP is exonic, intronic or intergenic and distances to coding regions for intronic and intergenic SNPs.

Any queried SNP that does not have sequence information in the FASTA file but has information in the GTF are still annotated by SNPdat. However, the returned information is limited to the first 17 columns and columns 23, 24 and 25 of the output file (Table 2.3).

Table 2.3 Summary description of the annotations provided by SNPdat.

Column Number	Description	Example
1	The queried SNPs chromosome ID	CHR25
2	The queried SNPs genomic location	286966
3	Whether or not the SNP was within a feature	Y
4	Region containing the SNP; either exonic, intronic, or intergenic	Exonic
5	Distance to nearest feature	NA
6	Either the closest feature to the SNP or the feature containing the SNP	CDS
7	The number of different features that the SNP is annotated to	2
8	The number of annotations of the current feature	[1/1]
9	Start of feature (bp)	286859
10	End of feature (bp)	287050
11	The gene ID for the current feature	ENSBTAG00000016571
12	The gene name for the current feature	ITFG3_BOVIN
13	The transcript ID for the current feature	ENSBTAT00000022045
14	The transcript name for the current feature	ITFG3_BOVIN
15	The exon that contains the current feature and the total number of annotated exons for the gene containing the feature	[3/11]
16	The strand sense of the feature	+
17	The annotated reading frame (when contained in GTF)	2
18	The reading frame estimated by SNPdat	NA
19	The estimated number of stop codons in the estimated reading frame	0
20	The codon containing the SNP, position in the codon and reference base and mutation	C[C/G]T
21	The amino acid for the reference codon and new amino acid with mutation in place	[P/R]
22	Whether or not the mutation is synonymous	N
23	The protein ID for the current feature	ENSBTAP00000022045
24	The RS identifier for queries that map to known SNPs	rs134558771
25	Error messages, warnings etc	NA

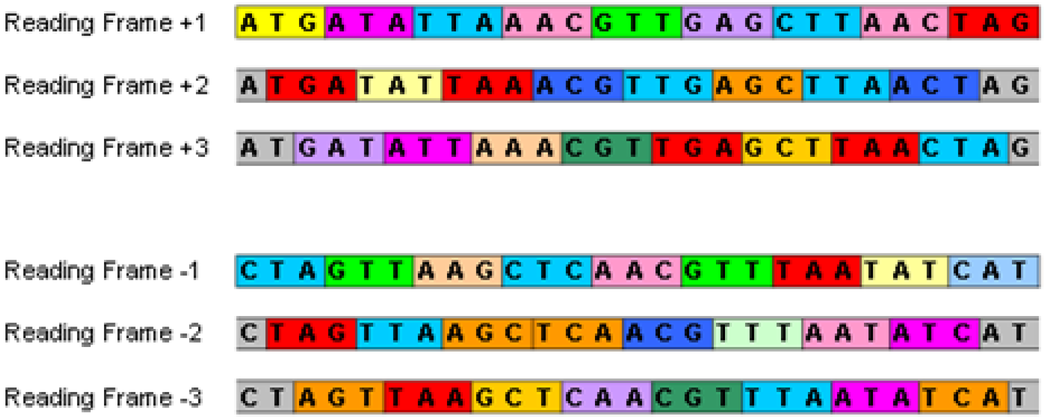

Reading Frame +1
A T G A T A T T A A A C G T T G A G C T T A A C T A G
Reading Frame +2
A T G A T A T T A A A C G T T G A G C T T A A C T A G
Reading Frame +3
A T G A T A T T A A A C G T T G A G C T T A A C T A G
Reading Frame -1
C T A G T T A A G C T C A A C G T T T A A T A T C A T
Reading Frame -2
C T A G T T A A G C T C A A C G T T T A A T A T C A T
Reading Frame -3
C T A G T T A A G C T C A A C G T T T A A T A T C A T

Figure 2.3 Each of the six possible reading frames for a given sequence. SNPdat counts the number of stop codons (coloured in red) in each reading frame and identifies the reading frame with the smallest number of stop codons as the correct reading frame for that feature. The last codon will be ignored as a stop is expected here. In the above, reading frame +1 will be chosen by SNPdat.

2.3.1 Non-coding SNPs

Next, all intronic and intergenic SNPs are identified and processed. The nearest feature to a non-coding SNP is identified and relevant data, such as the distance to the feature, feature IDs, strand sense, start and end position, is retrieved. If the SNP is equidistance from more than one feature, a separate line for each feature will be reported. Column seven of the output file contains the number of features reported for a SNP (Table 2.3).

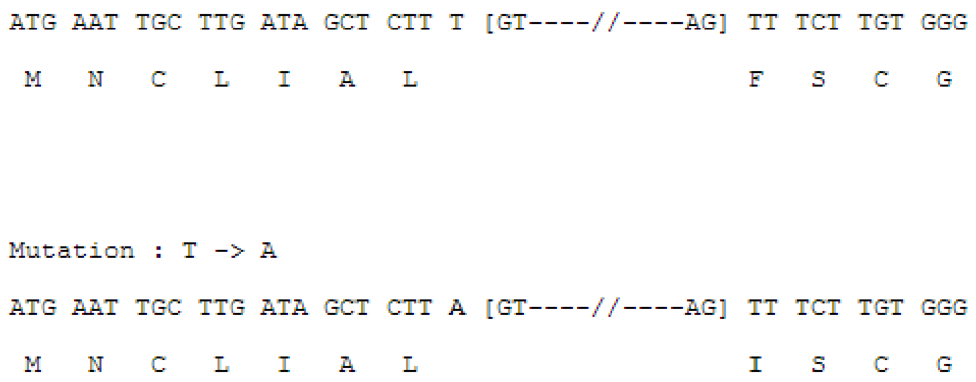
2.3.2 Coding SNPs

All features that a SNP occurs in are identified and printed to separate lines. Information calculated and retrieved for a feature containing a SNP is contained in columns 9-17 of the output file (Table 2.3). Columns 18-22 contain information estimated from the sequence of the feature such as the reading frame (Figure 2.3), the position in the codon, reference and mutant amino acid and whether or not the SNP is synonymous. The estimated reading frame is relative to the strand sense of the feature. If no strand sense is available from the GTF, SNPdat assumes that the strand sense is positive. To estimate the reading frame, when none is provided in the GTF file, SNPdat will extract the sequence and count the number of stop codons (excluding the last codon position) in each of the 6 possible reading frames. The reading frame with the smallest number of stop codons is chosen as the correct reading frame for the feature. In the event of a tie between two or more reading frames, SNPdat will chose the reading frame closest to the first reading frame of the feature strand sense.

Finally, all SNPs are cross referenced against information retrieved from external databases such as dbSNP. SNPs that do not have sequence information in the FASTA file but have information in the GTF are still annotated by SNPdat. However, the returned information is limited to information which can be returned without reference to the DNA sequence (columns 1 to 17 and 23 to 25). This includes information from external databases. See Table 2.3 for more details.

2.3.3 Advanced options

There are also a number of advanced features available for use with SNPdat. There is also an advanced option available for SNPdat users; “feature boundary crossing”. When annotating SNPs to features, should a SNP occur in the last position of the feature but the first position of the codon for that reading frame, SNPdat will by default return the base and question marks for the bases outside of the feature region. However, sometimes users will want to extract this extra information from the next feature in the transcript (Figure 2.4). SNPdat, can do this with the ‘-x’ option.



```
ATG AAT TGC TTG ATA GCT CTT T [GT----//----AG] TT TCT TGT GGG
 M   N   C   L   I   A   L                     F   S   C   G


Mutation : T -> A
ATG AAT TGC TTG ATA GCT CTT A [GT----//----AG] TT TCT TGT GGG
 M   N   C   L   I   A   L                     I   S   C   G
```

Figure 2.4 Annotation of codon spanning two features using the –x option. A mutation from T to A at the first position of the eighth codon, causing an amino acid change from phenylalanine to isoleucine. By default SNPdat would not return the amino acid change as the codon spans two features. With the –x option enabled, information will be retrieved from the both features to complete the SNP annotation to the correct codon.

2.4 Annotation of SNPs discovered using high-throughput DNA sequencing

To demonstrate SNPdat’s ease of use, *de novo* SNPs discovered by Mullen *et al.*(2012) were annotated using SNPdat. As a comparison, Annovar was also used to analyse this dataset.

2.4.1 Materials and Methods

This dataset consisted of 4,566 SNPs discovered using high-throughput DNA sequencing of target enriched pooled DNA samples of 83 genomic regions from groups of dairy cattle. The SNPs included both novel and putative variants from 28 chromosomes including the X chromosome.

2.4.1.1 Retrieval of FASTA and GTF information

For SNPdat: EnsGene annotation and FASTA sequence files for *Bos Taurus* were retrieved from the University of California Santa Cruz (UCSC) ftp site (Meyer et al., 2013). A GTF version of the ensGene annotation file was supplied to SNPdat along with the FASTA file. SNPdat does not require any pre-processing steps and so both of these files were used as input for the software.

For Annovar: The same annotation and FASTA files were retrieved for use with Annovar. The FASTA file was pre-processed to create a sequence file using information from both the FASTA file and ensGene annotation file. The new sequence file and original ensGene file were then supplied as input for Annovar.

2.4.2 Results

Both tools annotate SNPs to coding regions (CDS), 3 prime untranslated regions (UTR), 5 prime UTR, intronic and intergenic regions (Table 2.4). SNPdat annotated SNPs to a larger number of features and transcripts (11,987 known features). Both tools identified mutations leading to stop gains, stop losses and other non-synonymous changes. However, from the SNPdat output file it was possible to estimate information about the annotations that was not possible from the Annovar output (Figure 2.5). Additionally, the output from SNPdat is a simple tab-delimited format with the same number of columns in every row. This meant that the results from SNPdat could be easily imported into other software, such as R (R Development Core Team, 2011), for further analysis.

Both Annovar and SNPdat annotated 299 SNPs in coding regions to a total of 382 transcripts. Of these, 231 SNPs were non-synonymous and 151 SNPs were synonymous mutations (Figure 2.5A). From the SNPdat output file it was possible to determine upstream and downstream distances for SNPs to coding regions (Figure 2.5B). Also, from the SNPdat output file it was determined that 96, 103 and 32 non-synonymous SNPs occurred in the first, second and third codon position, respectively (Figure 2.5C). SNPdat and Annovar both found a large proportion (77%) of intergenic SNPs were within 2,000 base pairs of coding regions. Additionally, from the SNPdat output file it was determined that 39% of intronic SNPs were within a 1,000 base pair region surrounding exons (Figure 2.5D).

Table 2.4 The number of SNPs annotated to different regions by SNPdat and Annovar. Misc features include non-coding RNA and splicing. These features were not included in the GTF version of the ensGene annotation file and so SNPdat was unable to identify them as such.

Region	**SNPdat**	**Annovar**
Coding	299	299
3 prime UTR	108	105
5 prime UTR	29	28
Intronic	3285	3284
Intergenic	845	845
Misc.	0	5
Total	4566	4566

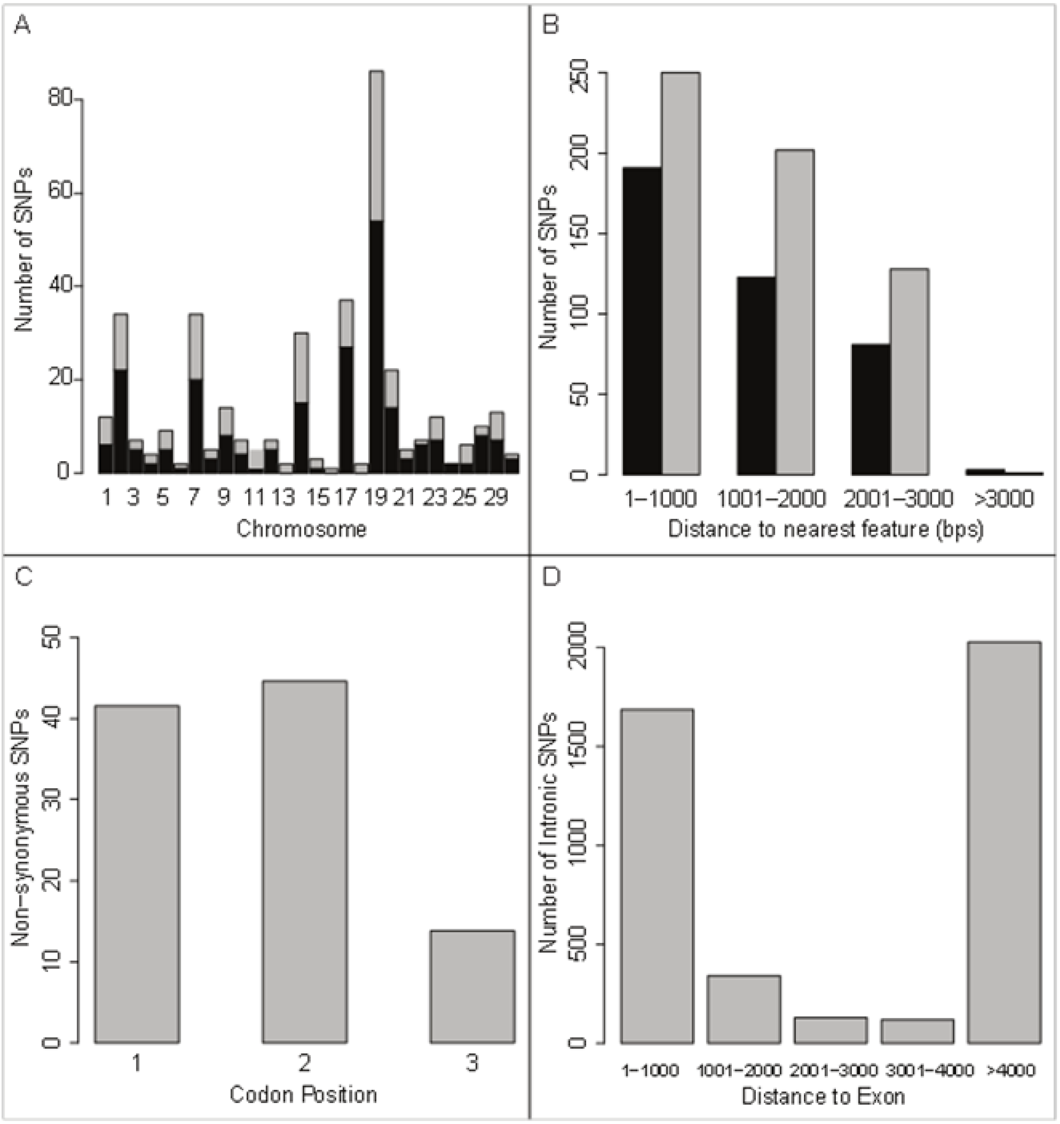


Figure 2.5 Sample of plots obtained using the results of SNPdat. (A) The number of non-synonymous (black) and total number of exonic SNPs (grey) found on each chromosome. **(B)** Distances of intergenic SNPs, upstream (black) and downstream (grey) to the nearest transcripts. **(C)** Synonymous versus non-synonymous SNPs: 231 exonic SNPs were non-synonymous. 96 (41.56%) in the first codon position, 103 (44.59%) in the second codon position and 32 (13.85%) in the third codon position. **(D)** Distances of Intronic SNPs to the nearest exon.

2.5 Discussion

Next-generation sequencing is a growing research area, generating huge amounts of information. As datasets continue to grow and more species have their genomes sequenced, tools needed for rapid annotation of newly discovered SNPs will become increasing important. The rationale behind SNPdat is to provide a simple to use tool for researchers annotating the results of *de novo* SNP discovery projects. It is especially intended for use by researchers with limited bioinformatic experience.

Future updates of SNPdat will most likely include additional options for handling large datasets. For example, SNPdat would be a very good candidate for a parallelised version. Because SNPdat annotates each SNP individually, input files could be split on-the-fly and each portion analysed separately by a single processor before the results are collated at the end. Currently, this can be done manually by the user.

SNPdat can provide a valuable insight into the functional roles associated with discovered SNPs and cross reference information with external sources. As a command line tool it can easily be incorporated into existing SNP discovery pipelines and fills a niche for analyses involving non-model organisms that are not supported by many available SNP annotation tools.

Chapter 3: Whole genome association study identifies regions of the bovine genome and biological pathways involved in carcass trait performance in Holstein-Friesian cattle

3.1 Introduction

Animal growth is an economically important trait for livestock raised for meat production. Carcass traits, related to animal growth, are critical to the biological and economical efficiency of cattle production and, as such, there is great interest in understanding the underlying genomic architecture influencing these traits. Four traits related to carcass performance have been identified as economically important in Irish beef production: carcass weight, carcass fat, carcass conformation of progeny and cull cow carcass weight. Quantitative trait loci (QTL) associated with a particular trait can be used to predict disease risk or genetic merit of an animal (Wray et al., 2007; de Roos et al., 2011). This information may also be used to investigate the molecular mechanisms and biological pathways involved in phenotypic variation between animals. Investigating complex traits in domestic animals may also provide insights into mechanisms underlying similar traits, such as growth and fat deposition, in humans.

Holstein-Friesian cattle are a popular breed of cow primarily used for their ability to produce large amounts of milk. However, Holstein-Friesian cattle are also an important source of meat for beef production and exports. A number of studies in cattle have identified associations between carcass traits and regions of the bovine genome. Carcass trait QTL have been reported on chromosomes 2, 3, 6, 14, 20 and 29 (MacNeil and Grosz, 2002; Kim et al., 2003; McClure et al., 2010). However, most studies reporting carcass QTL have been performed using beef breeds such as Aberdeen Angus (McClure et al., 2010). Although many studies have reported carcass QTL in regions containing genes with a known role in animal growth such as the myostatin gene on bovine chromosome 2 (Grobet et al., 1997; McPherron and Lee, 1997), little is known about the mechanisms or underlying biological pathways involved in growth or carcass traits. Moreover, many of the reported QTL have been identified using raw phenotypic data which is subject to environmental influences such as high levels of nutrition. Unlike raw phenotypic data, measures of genetic merit account for the additive genetic effects that are responsible for performance of an animal (Vanraden et al., 1990). Few studies have investigated the association of SNP genotypes with carcass performance utilizing measures of genetic merit estimated in dairy breeds.

The objective of this study was to identify regions of the bovine genome associated with carcass performance traits using two statistical approaches. A single marker regression and multi-locus Bayesian approach were used to estimate the marker associations. Regions found associated with a trait were then further investigated to identify the potential causal pathways and biological processes underlying each trait.

3.2 Materials and Methods

3.2.1 Ethics statement

Semen samples for genotyping were collected by the Irish Cattle Breeding Federation (ICBF) and partner artificial insemination organisations. All animal procedures were carried out according to the provisions of the Irish Cruelty to Animals Act (licenses issued by the Department of Health and Children).

3.2.2 Genotypic data

Genotypes of 54,001 biallelic single nucleotide polymorphism (SNP) markers from 5,706 Holstein-Friesian sires were available for use in this study. All genotyping was carried out using the Illumina Bovine SNP50 version 1 Beadchip (Illumina Inc., San Diego, CA; (Matukumalli et al., 2009)). SNP positions were based on the Btau 4.0 assembly of the bovine genome. All SNPs on the X-chromosome or with an unknown position in the genome were removed from the dataset. Quality filtering was then undertaken to remove SNPs with inconsistent Mendelian inheritance patterns from sire to progeny. All SNPs that had a minor allele frequency of less than 5% were also discarded. If a SNP had greater than 5% of calls missing, it was excluded from further analysis. Also, SNPs that failed to distinctly cluster into homozygous and heterozygous calls were removed. A total of 42,477 SNPs remained for analysis after quality filtering.

3.2.3 Phenotypic data

Phenotypes for four economically important carcass traits were used in this study; carcass weight, carcass fat, carcass conformation of progeny and cull cow carcass weight. Carcass weight refers to the cold weight of the carcass taken within 2 hours of slaughter after being bled and eviscerated, and after removal of skin, external genitalia, the limbs at the carpus and tarsus, head, tail, kidneys and kidney fats and the udder. Progeny carcass weight (CWT) is the carcass weight of a sire's offspring/progeny measured on males from 300-1200 days and females from 300-875 days of age (females which have not produced a calf). Progeny carcass fat (CFAT) is the quantity of fat on the carcass of the slaughtered animal. Progeny carcass conformation (CONF) is the thickness of muscle on the carcass of the slaughtered animal. Cull cow carcass weight (CULL) refers to the carcass weight of a dairy or beef cow slaughtered for meat at the end of her productive life. Cows are aged between 875 and 4000 days of age. Phenotypes for each of these traits are published as predicted transmitting abilities (PTAs), which are sire genetic merit based not on the sires themselves but on the performance of their progeny across multiple generations. The Irish Cattle Breeding Federation calculated PTAs and their respective reliabilities were available for all animals used in this study. Genotypic and phenotypic data for all animals utilized in this study can be requested from the Irish Cattle Breeding Federation (ICBF). The Irish Cattle Breeding Federation database identifiers for all animals used in this study are contained in Electronic Appendix 3.1. These animals were representative of the Holstein-Friesian population in Ireland. Phenotypic edits were then applied to the animals. An adjusted reliability was estimated for each animal by removing the parental contribution to reliability as described by Harris and Johnson (1998). For each trait separately, animals with an adjusted reliability (reliability less parental contribution) of <70% were removed. Following removal of animals with a low adjusted reliability, 1061 animals remained for further analysis. Summary statistics for each of the phenotypes, following removal of animals with an adjusted reliability of <70%, are in Table 3.1.

Table 3.1 Summary statistics for the phenotypic data. Summary statistics include the total number of phenotype records (N), minimum value, maximum value, mean and standard deviation (σ) for each trait. Phenotypes are expressed as predicted transmitting abilities.

Trait	**N**	**Min**	**Max**	**Mean**	**σ**
Carcass Weight	941	-26.38	13.88	-4.07	6.39
Carcass Fat	768	-0.72	0.62	-0.11	0.23
Carcass conformation	936	-1.62	0.46	-0.67	0.31
Cull cow carcass weight	763	-29.44	29.65	0.33	8.28

3.2.4 Statistical analyses

Two statistical approaches, a frequentist and Bayesian approach, were used to estimate associations between SNPs and each trait separately.

3.2.4.1 Single SNP regression

The single SNP regression (SSR) model included each SNP separately as a continuous variable in a linear animal mixed model using ASReml (Gilmour et al., 2009b). The individual animal was included as a random effect. Relationships between animals were accounted for using the additive genetic relationship matrix. Pedigree information consisted of 6,854 animals. The dependent variable was de-regressed PTA. Marker effects and associated P-values for each SNP were obtained from the analysis. P-values were adjusted to correct for errors arising from multiple testing using a false discovery rate (FDR) approach (FDR < 0.05) described by Storey and Tibshirani (2003). This procedure was carried out using the q-value package in R. Resultant q-values <0.05 were defined as significant. Genomic co-ordinates, identifier information and q-values for all SNPs in the analysis are contained in Electronic Appendix 3.2.

3.2.4.2 Bayesian approach

The second statistical approach utilized the Bayesian mixture model "BayesB" as described by Meuwissen *et al* (2001). This model allows the incorporation of prior knowledge about the distribution of SNPs effects. An inverse chi-squared distribution ($v = 4.234$, $S = 0.0429$) was included in the model as the prior distribution of the mean and genetic variation of each SNP included in the model.

A prior value was assigned to π which quantifies a prior probability of association ($1 - \pi$) for each SNP. As this prior probability is assigned to all SNPs in the analysis, it reflects the prior proportion of SNPs assumed to be associated with a particular trait. Analyses were run with alternative prior proportions assumed to be associated with a particular trait ($1- \pi$) ranging from 0.05 to 6.25×10^{-5}.

Additional analyses were also performed using the proportion of non-significant ($q \geq 0.05$) SNPs that were estimated from the SSR analysis (pSSR), and half and double this value, to determine π. This was then used to quantify a prior proportion of SNPs assumed to be associated with each trait ($1 - \pi$). A total of eleven analyses were run for each trait. Markov Chain Monte Carlo (MCMC) chains were used to sample every 500th iteration from the posterior distribution of SNP effects. Total iterations for each analysis are contained in Table 3.2.

Table 3.2 Maximum iterations each Bayesian analysis was run for.

($1 - \pi$) = prior proportion of SNPs assumed to be associated with a trait; pSSR = the proportion of SNPs not significant from single SNP regression analysis. One minus this value is the prior proportion of SNPs assumed to be associated with each trait; CWT = carcass weight; CFAT = carcass fat; CONF = carcass conformation; CULL = cull cow carcass weight

1 - π	**CWT**	**CFAT**	**CONF**	**CULL**
1 - pSSR/2	800,000	950,000	850,000	700,000
1 - pSSR	800,000	950,000	850,000	700,000
1 - pSSR*2	800,000	950,000	850,000	700,000
6.25×10^{-5}	600,000	700,000	600,000	700,000
1.25×10^{-4}	600,000	700,000	600,000	700,000
2.5×10^{-4}	550,000	650,000	600,000	700,000
5.0×10^{-4}	600,000	700,000	600,000	700,000
1.0×10^{-3}	400,000	500,000	400,000	450,000
2.45×10^{-3}	450,000	500,000	400,000	500,000
1.0×10^{-2}	400,000	500,000	400,000	550,000
5.0×10^{-2}	400,000	500,000	400,000	500,000

3.2.4.3 Convergence testing and confirmation

In any MCMC analysis, convergence of the model must be confirmed before making any posterior inferences. The convergence diagnostics used in this chapter will be discussed in greater detail in Chapter 4. For each of the analyses in this chapter, convergence of the model was confirmed by visual inspection of summed absolute log-likelihood values. All sampled iterations before convergence were discarded as burn-in. The number of iterations discarded as burn-in for each analysis is contained in Table 3.3. From the remaining sampled iterations, posterior probabilities (PPs) of association were calculated for each SNP. A PP is the number of sampled iterations after burn-in that a SNP had a non-zero effect divided by the total number of sampled iterations after burn-in. The PP is indicative of the probability that a SNP is associated with a phenotype. A PP of zero indicates a low probability of association whereas a PP of 1 indicates a high probability of association. SNPs with a PP>0.5 were defined as high PP SNPs.

To ensure that convergence was successfully achieved, a secondary check for model convergence was performed by quantifying and plotting the total number of SNPs that had a PP>0.5 at each iteration. The resultant trace plot was visually inspected to determine if the MCMC chains had run long enough to have confidence that all high PP SNPs had been identified.

A third check for model convergence was carried out. This was to ensure that not only had the posterior probabilities converged, but also the estimated marker effects for each SNP. The combined difference between the estimated SNP effect of those SNPs with a PP>0.5 from the Bayesian approach and the SNP effect for the same set of SNPs as estimated using the SSR approach was calculated using a Euclidean distance. Visual inspection of the trace plot produced by plotting a Euclidean distance at each iteration confirmed convergence of this model parameter.

Table 3.3 Initial iterations discarded as burn-in from each Bayesian analysis. (1 – π) = prior proportion of SNPs assumed to be associated with a trait; pSSR = the proportion of SNPs not significant from single SNP regression analysis. One minus this value is the prior proportion of SNPs assumed to be associated with each trait; CWT = carcass weight; CFAT = carcass fat; CONF = carcass conformation; CULL = cull cow carcass weight

1 - π	**CWT**	**CFAT**	**CONF**	**CULL**
1 - pSSR/2	107,000	126,500	102,500	192,000
1 - pSSR	122,000	102,000	163,500	109,500
1 - pSSR*2	277,000	77,000	211,500	54,000
6.25×10^{-5}	132,000	63,500	58,000	65,000
1.25×10^{-4}	135,500	74,500	58,500	136,500
2.5×10^{-4}	213,000	99,500	102,000	182,500
5.0×10^{-4}	175,000	128,500	60,000	83,500
1.0×10^{-3}	52,500	130,500	90,000	131,500
2.45×10^{-3}	57,500	94,500	120,500	69,500
1.0×10^{-2}	106,000	121,000	73,000	50,500
5.0×10^{-2}	51,500	150,500	195,000	156,000

3.2.4.4 Identifying significant associations using the Bayesian approach

For each analysis, once convergence had been confirmed and the burn-in discarded, a PP was calculated for each SNP. For each trait separately, high PP

(PP>0.5) SNPs for each of the eleven analyses (1 - π = 1- pSSR/2, 1 - pSSR, 1 - pSSR × 2 and 0.05-6.25×10^{-5}) were identified. The number of analyses that a SNP had a PP>0.5 across the 11 analyses was calculated and assigned to a SNP as its occurrence rate.

For each of the eleven analyses within a trait, an average occurrence rate was calculated by summing the individual SNP occurrence rates of SNPs with a PP>0.5 and dividing this value by the total number of SNPs with a PP>0.5. The analysis with the highest average occurrence rate was then identified (Table 3.4). All SNPs with a PP > 0.5 within the analysis with the highest average occurrence rate were then considered significantly associated with the respective trait. This was done for each trait separately, resulting in 4 datasets of significantly associated SNPs corresponding to each trait under investigation (Electronic Appendix 3.2). Each dataset represented the analysis with the largest number of frequently occurring high PP SNPs for each trait.

Table 3.4 Average occurrence rate of high PP SNPs for each Bayesian analysis. (1 – π) = prior proportion of SNPs assumed to be associated with a trait; pSSR = the proportion of SNPs not significant from single SNP regression analysis. One minus this value is the prior proportion of SNPs assumed to be associated with each trait; CWT = carcass weight; CFAT = carcass fat; CONF = carcass conformation; CULL = cull cow carcass weight

1 - π	**CWT**	**CFAT**	**CONF**	**CULL**
1 - pSSR/2	4.23	2.38	1.37	3.94
1 - pSSR	5.5	1.79	1.34	5.13
1 - pSSR*2	5.07	2	1.18	3.56
6.25×10^{-5}	5.75	2.67	3.33	3
1.25×10^{-4}	4.42	1.33	2.43	3.62
2.5×10^{-4}	4.89	2.42	2.14	4.25
5.0×10^{-4}	3	2.2	2.06	4.5
1.0×10^{-3}	3.28	1.52	1.94	3.25
2.45×10^{-3}	2.79	1.49	1.63	3.07
1.0×10^{-2}	2.27	1.45	1.27	4.75
5.0×10^{-2}	1.33	1	1.05	1

3.2.5 Identification of significant SNPs in known QTL

Significant SNP positions from both methods were compared to known QTL regions from cattle QTLdb (Hu et al., 2007). Search terms used to retrieve QTL regions for each trait were “carcass weight”, “carcass fat percentage” and “carcass muscle percentage”. Boundaries for known QTLs were defined by using the “QTL Span” values from cattle QTLdb. The retrieved file was modified to imitate a GTF file format using in-house Perl scripts. SNPs were assigned to a QTL using SNPdat (Doran and Creevey, 2013) if they occurred within the boundary of that QTL.

3.2.6 Pathway analysis

Four datasets, corresponding to each trait, were created by identifying all bovine genes within a 500kb region up and downstream of SNPs found significantly associated with a trait using the Bayesian method. To investigate the combined role that some pathways may have on each of these traits, a combined trait dataset containing all genes from each of the individual trait datasets was also created. The genes in each of these five datasets (each individual trait and the combined trait datasets) were then mapped to their human orthologs using the mapping available from version hg19 of the human genome. For each dataset, the R (The R Project) package GOSeq (Young et al., 2010), without the correction for gene length bias, was used to identify the KEGG (Kanehisa et al., 2012) pathways which were significantly over-represented by the set of genes ($p < 0.05$) compared against a background of all genes in the human genome.

3.3 Results

3.3.1 Significant Associations

3.3.1.1 Carcass weight

Using the SSR method, two SNPs were associated (q<0.05) with carcass weight. These SNPs were on chromosomes 3 and 19 (Figure 3.1).

In the Bayesian analysis, eleven SNPs were associated with CWT including two SNPs on chromosome 3, within 2.5 Mb of each other (Table 3.5). Only one of the 11 SNPs was associated with carcass weight and at least one of the other three carcass traits. This SNP, on chromosome 6 (~85Mb), was associated with both carcass weight and carcass conformation using the Bayesian method. None of the SNPs identified as associated with carcass weight were common to both statistical approaches.

Table 3.5 The number of SNPs that were significantly associated with each trait from the single SNP regression (SSR) or Bayesian analysis. SNPs with a q-value <0.05 from the SSR analysis were considered significantly associated with a trait.

Trait	SSR	Bayesian
Carcass Weight	2	11
Carcass Fat	25	6
Carcass conformation	483	12
Cull cow carcass weight	48	15

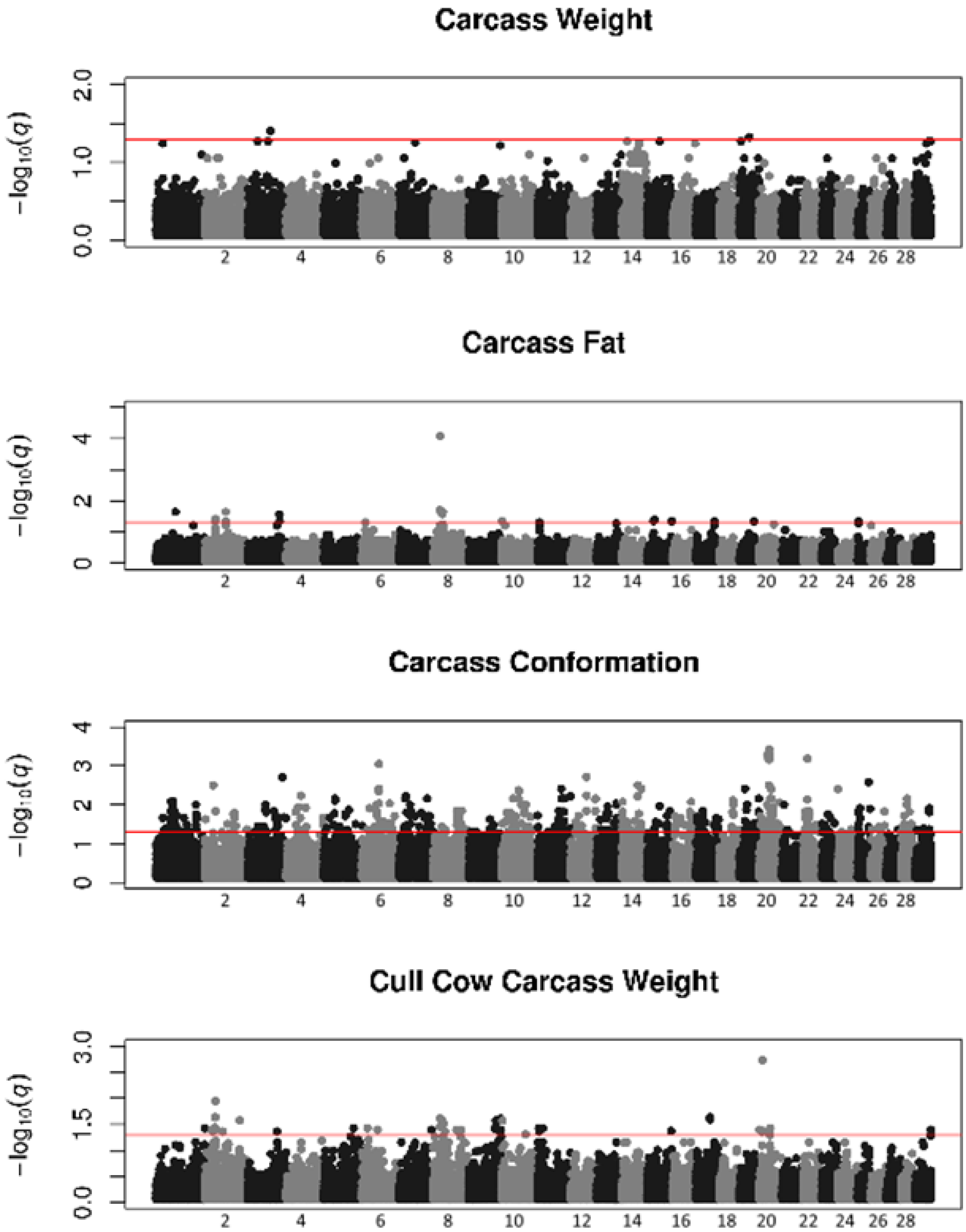


Figure 3.1 Genome-wide association results from the single SNP regression are plotted for each trait. Results for SNPs on all autosomal chromosomes are plotted as negative log transformed q-values. The red continuous line indicates a significance threshold of 1.3 ($q<0.05$). Odd numbered chromosomes are plotted in black and even numbered in grey.

3.3.1.2 Carcass fat

Using the SSR approach, 25 SNPs were associated ($q<0.05$) with carcass fat (Table 3.5). The most significantly associated SNP from this analysis ($q = 8.45 \times 10^{-5}$), rs109514593, was located on chromosome 8 at ~22 Mb (Figure 3.1), while another SNP (rs41607785) approximately 1 Mb away from rs109514593, was also associated with carcass fat. Five SNPs were associated with both carcass fat and cull cow carcass weight. One SNP, rs109776183, was associated with both carcass fat and carcass conformation.

Using the Bayesian method 6 SNPs were associated with carcass fat. Each of these SNPs were on different chromosomes of the genome. One SNP (rs29011003) on chromosome 3 was associated with carcass fat using both the Bayesian and SSR methods. This SNP was located approximately 600 kb away from rs43359171, which was also associated with carcass fat using the SSR approach.

3.3.1.3 Carcass conformation

A total of 483 SNPs were associated ($q < 0.05$) with carcass conformation in the SSR analysis (Table 3.5). Significant SNPs for carcass conformation were located on all chromosomes (Figure 3.1). There were 27 SNPs that showed a strong association with this trait ($q<0.005$), the most significant ($q = 3.787 \times 10^{-4}$) of which was on chromosome 20. This SNP, rs41580285, resided within a cluster of 5 strongly associated SNPs ($q<0.005$), all of which were less than 1 Mb away from the growth hormone receptor (GHR) gene.

Twelve SNPs were associated with carcass conformation in the Bayesian analysis. Four of these SNPs were also associated with carcass conformation using the SSR approach. One of these SNPs was strongly associated with carcass conformation ($q<0.005$) using the SSR method. One SNP, located on chromosome 6, was also associated with carcass weight using the Bayesian method.

3.3.1.4 Cull cow carcass weight

A total of 61 SNPs were associated with cull cow carcass weight using either the Bayesian or SSR method (Table 3.5). Of these, 48 SNPs were associated ($q<0.05$) with cull cow carcass weight using the SSR method (Figure 3.1). One SNP, rs41935177, was detected as being associated ($q<0.005$) in both the SSR and Bayesian method. Seven SNPs from this analysis were associated with cull cow carcass weight and another trait (5 SNPs were associated with carcass fat and 2 with carcass conformation) using the SSR approach.

Fifteen SNPs were associated with cull cow carcass weight in the Bayesian analysis. Two of these SNPs, rs109184437 and rs41935177, were also significantly associated with cull cow carcass weight using the SSR approach. In fact, rs41935177, was the most significantly associated SNP with CULL from the SSR analysis ($q = 1.813 \times 10^{-3}$).

3.3.2 Overlap with known QTL

A total of 254 significantly associated SNPs were found in known QTL for either carcass weight, carcass fat percentage or carcass muscle percentage. From the SSR analysis, 237 SNPs that were significantly associated with a trait were also found in a known QTL for the same trait. Nearly half (20/43) of SNPs found significantly associated with a trait using the Bayesian approach were also found in a known QTL.

3.3.3 Over-represented KEGG pathways

In total, 428 unique bovine genes were within 500kbs of a SNP associated with a trait using the Bayesian approach. Of these, 343 mapped to 333 human orthologs. The most significantly over-represented KEGG pathway detected using these genes was the peroxisome proliferator-activated receptor (PPAR) signalling pathway ($p=9.58 \times 10^{-4}$) (Figure 3.2). This pathway was significantly over-represented in both carcass fat and the combined trait analyses. In fact, six of the seven pathways significantly over-represented in the combined trait analysis were also significantly over-represented for a single trait when only orthologs from that trait were used in the analysis. Tyrosine metabolism (Figure 3.3) was the only pathway that was not significant for an individual trait but was found significantly over-represented using the combined trait dataset. Twenty-five different pathways were significantly over-represented across all analyses and are contained in Table 3.6.

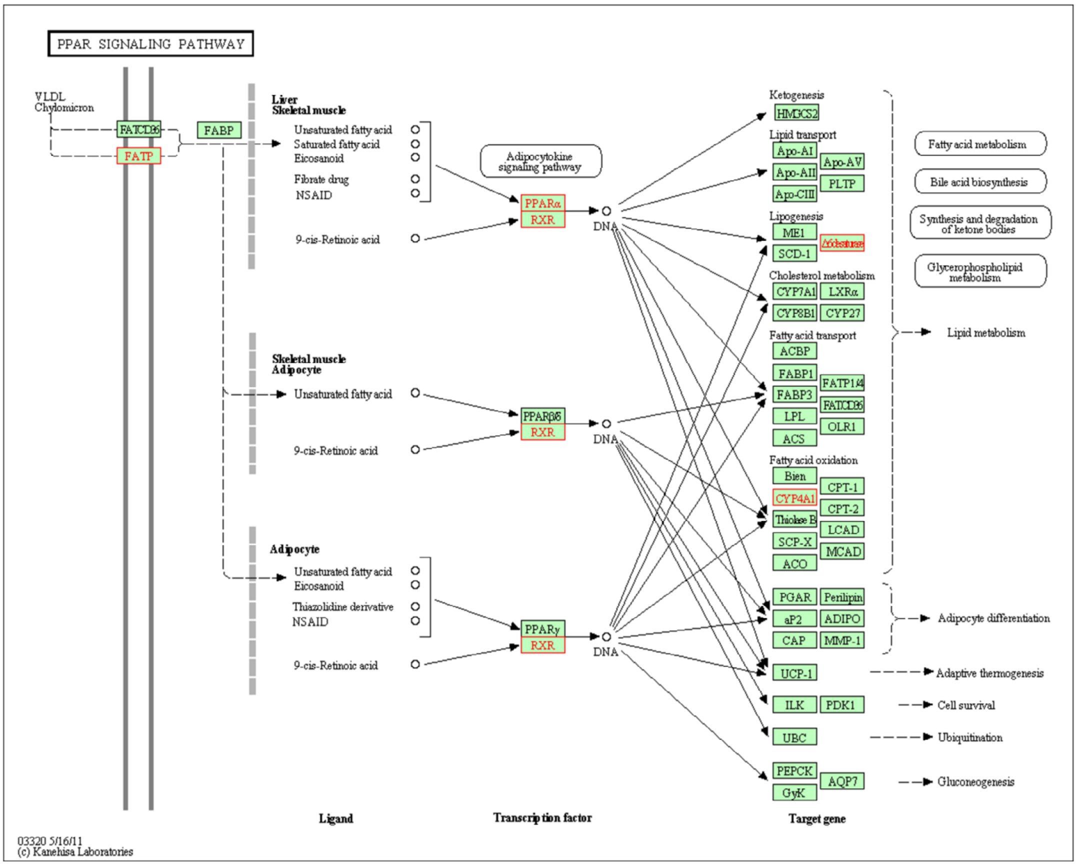


Figure 3.2 The peroxisome proliferator-activated receptor signalling pathway. PPAR was the most significantly over-represented KEGG pathway in the combined trait analysis. Genes in this pathway were in regions surrounding SNPs associated to three different traits using the Bayesian approach (coloured in red).

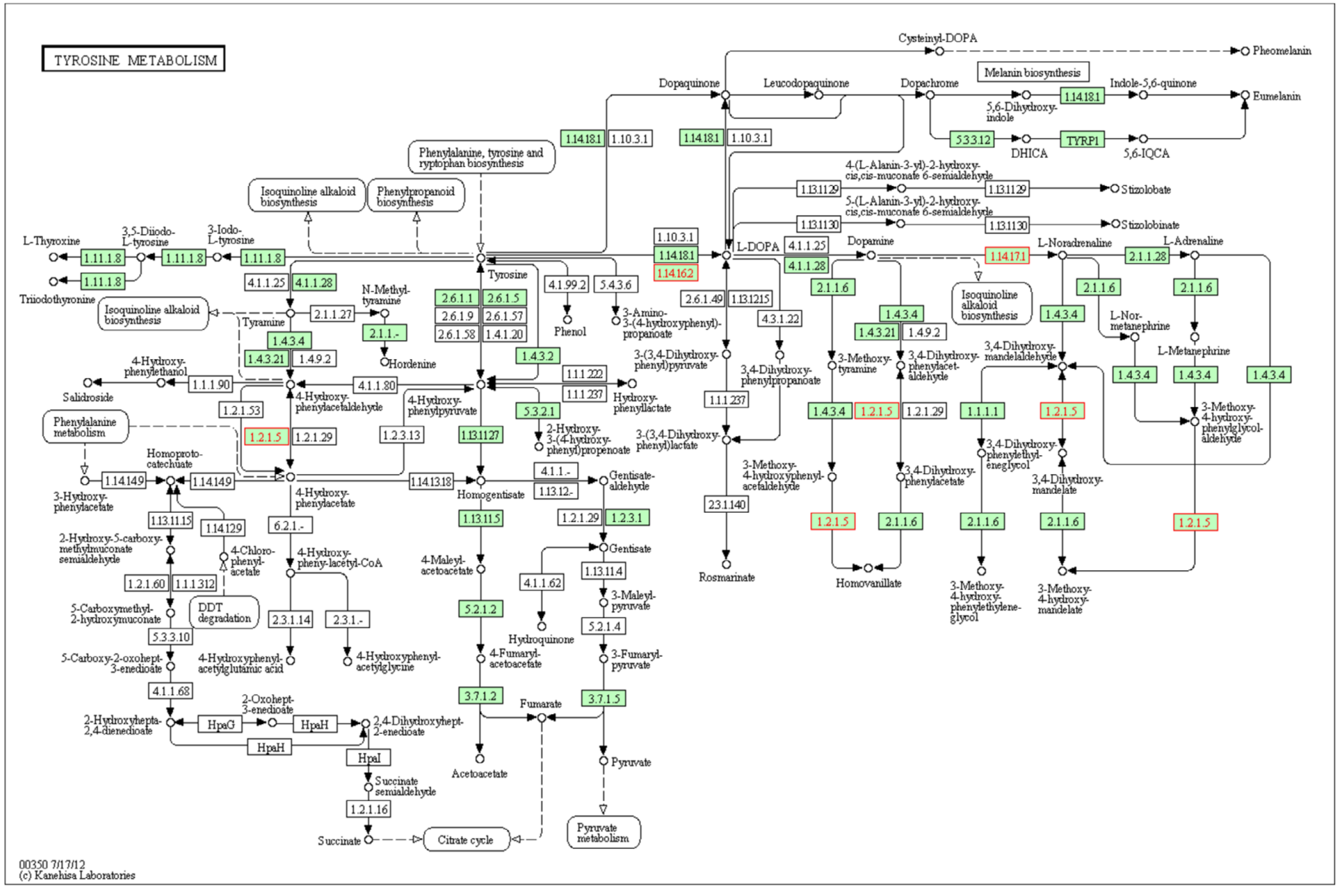


Figure 3.3 The tyrosine metabolism signalling pathway. This pathway was the only significantly over-represented pathway from the combined trait analysis that was not significantly over-represented for an individual trait. Genes in regions surrounding significant SNPs from the Bayesian analysis are highlighted in red.

Table 3.6 Significantly over-represented KEGG pathways and candidate genes. Candidate genes are genes that occurred in the over-represented pathway and were within 500kbs of a SNP significantly associated with the trait using the Bayesian approach. CWT = carcass weight; CFAT = carcass fat; CONF = carcass conformation; CULL = cull cow carcass weight; ALL = significantly over-represented KEGG pathways using combined trait dataset.

Trait	Pathway Name	p-value	Candidate Genes
CWT	Jak-STAT signalling pathway	0.00281	IL12RB2, IL23R, JAK1, LEPR
CWT	P53 signalling pathway	0.02504	GADD45A, GTSE1
CWT	Adipocytokine signalling pathway	0.02972	LEPR, PPARA
CWT	Pathways in cancer	0.03540	FGF18, JAK1, RET, WNT7B
CWT	Phosphatidylinositol signalling system	0.03820	DGKD, INPP5D
CWT	Sulfur relay system	0.03888	TRMU
CFAT	PPAR signalling pathway	0.00096	CYP4A11, CYP4A22, FADS2
CFAT	Protein digestion and absorption	0.00109	PGA3, PGA4
CFAT	Biosynthesis of unsaturated fatty acids	0.00146	FADS1, FADS2
CFAT	Vascular smooth muscle contraction	0.00389	CYP4A11, CYP4A22
CFAT	Fatty acid metabolism	0.00638	CYP4A11, CYP4A22
CFAT	Retinol metabolism	0.00761	CYP4A11, CYP4A22
CFAT	Arachidonic acid metabolism	0.00964	CYP4A11, CYP4A22
CFAT	Non-homologous end-joining	0.03645	FEN1
CONF	Inositol phosphate metabolism	0.00522	INPP5B, PI4KB, PIP5K1A
CONF	Phosphatidylinositol signalling system	0.01180	INPP5B, PI4KB, PIP5K1A
CONF	Biotin metabolism	0.01207	HLCS
CONF	Antigen processing and presentation	0.03286	CTSS, RFX5
CONF	Lysosome	0.03410	CTSK, CTSS, LAMP3
CONF	Hedgehog signalling pathway	0.03953	WNT3, WNT9B
CONF	Basal cell carcinoma	0.04233	WNT3, WNT9B
CULL	Glutathione metabolism	0.03212	ODC1, RRM2
CULL	Pathogenic Escherichia coli infection	0.03947	OCLN, ROCK2
CULL	Cell adhesion molecules (CAMs)	0.04021	CNTN2, OCLN,

			PTPRM
CULL	Vibrio cholerae infection	0.04576	ATP6V1C2, KCNQ1
ALL	PPAR signalling pathway	0.00202	CYP4A11, CYP4A22, FADS2, PPARA, RXRA, SLC27A6
ALL	Phosphatidylinositol signalling system	0.00403	DGKD, INPP5B, INPP5D, PI4KB, PIP5K1A
ALL	Protein digestion and absorption	0.01217	KCNJ13, KCNQ1, PGA3, PGA4
ALL	Non-homologous end-joining	0.02536	FEN1, LIG4
ALL	P53 signalling pathway	0.03208	CCNB1, GADD45A, GTSE1, RRM2
ALL	Tyrosine metabolism	0.03448	ALDH1A3, DBH, TH
ALL	Biotin metabolism	0.03838	HLCS

3.4 Discussion

The aim of this study was to identify regions of the bovine genome associated with carcass performance using phenotypes of four economically important carcass traits in Holstein-Friesian cattle: Carcass weight, carcass fat, carcass conformation of progeny as well as cow carcass weight. This information was then used to identify candidate genes and biological processes that may be involved in each of the traits under investigation. Two statistical approaches, a Bayesian and frequentist, were used to detect associations between SNPs and each of the traits separately. SNPs found associated using either approach were distributed across all autosomal chromosomes.

3.4.1 Predicted Transmitting Abilities as a phenotype

The phenotype of an animal is a combination of genetic and environmental effects. Genetic effects are the results of the genes inherited from parents. Environmental effects are the result of conditions the animal experienced such as nutrition and temperature. Many GWAS studies utilise raw phenotypic records (e.g. weight of an animal at the time of slaughter) to quantify associations between genotypes and phenotypes (e.g. Bolormaa et al., 2011; Kim et al., 2011; Nishimura et al., 2012). However, in this study we have used predicted transmitting abilities (PTAs), which are a measure of genetic merit as our phenotype. Unlike raw phenotypic records, that may be subject to environmental influence such as high levels of nutrition, PTAs account for cumulative genetic effects that are responsible for performance of an animal and its progeny (Vanraden et al., 1990). PTAs are predicted based on the performance of an animal and its relations (i.e. progeny) and have an associated reliability (confidence in the estimated PTA). By using only animals with a calculated reliability higher than 70%, we have greater confidence in the estimated phenotype for an animal. In this study PTAs from 1,061 animals were used as the phenotype. These were calculated for each animal utilising information from a much larger number of animals that was representative of the germplasm of Irish herds.

3.4.2 The Bayesian approach

Both the Bayesian and SSR methods differ fundamentally in their approaches. The single SNP regression method tests each SNP individually, whereas Bayesian approaches test all SNPs in the model simultaneously. This was particularly evident by the Bayesian approach identifying a single marker whereas the SSR approach sometimes identified a cluster of adjacent significant associations for the same location (e.g. chromosome 20 at ~10MB for CULL). Also, the Bayesian approach is advantageous as there is no need to correct for Type I errors arising from many thousands of tests. This allowed us to detect associations that might have been removed as false positives by the correction method applied to the SSR approach. Thus Bayesian approaches may then have greater power to detect SNPs with moderate effects on a trait of interest. Additionally, the ability to incorporate information *a priori* into the model would appear to be advantageous in complex traits which are influenced by many variants. Although inclusion of a prior may bias results to fit that prior (Gianola et al., 2009), it is likely that SNPs with the strongest association will be identified irrespective of the prior proportion of SNPs assumed to have an effect.

Our choice of prior would appear to be robust, as it represents the SNPs that most frequently occurred across different prior specifications.

The rational behind this approach is as follows: It has previously been reported that Bayesian analysis can be biased by the selection of a prior proportion of SNPs assumed to have an effect on a trait (Gianola et al., 2009; Knurr et al., 2013). Results from SSR can be used to select a prior proportion of SNPs believed to have an effect, although this may over-estimate the number of independent SNPs associated with a trait as many significant SNPs may be in LD with a single causative variant or gene. However, we propose that regardless of the proportion of SNPs used as a prior in the Bayesian analysis, the SNPs with the strongest LD signal should be represented in most (if not all) of the results. Our approach then is to select the analysis with the largest number of frequently occurring high PP SNPs. This approach identified that the analysis with the best prior proportion for one of our traits was the same as the proportion of significant SNPs from the SSR analysis (CULL). Two traits had a prior proportion estimated from the Bayesian approach that was smaller than the proportion of significant SNPs estimated using the SSR approach (CFAT and CONF). Although the prior proportion with the largest occurrence rate for CWT was slightly larger than the prior proportion estimated by the SSR (6.25×10^{-5} and 4.52×10^{-5}), the average occurrence rates for each prior were extremely similar (5.75 and 5.5, respectively (Table 3.4)).

3.4.3 Significant Associations

A large number of associations (584 SNPs) were detected across all traits using both statistical approaches. However, most of these were detected for

CONF (483) using the SSR approach (q<0.05). At a significance of q<0.005, a total of 27 SNPs were associated with CONF using the SSR approach. This figure was much more similar to the results from the other three traits. Using this significance threshold for CONF and a significance threshold of q<0.05 for the other three traits, 95 SNPs were associated with at least one trait using the SSR approach. This meant that 134 SNPs were associated with at least one of the traits using both the SSR and Bayesian approaches.

3.4.4 Overlap with known QTL

To date most association studies involving cattle have focused primarily on milk production traits (e.g. (Pryce et al., 2010; Meredith et al., 2012)). For example, from cattleQTLdb (Hu et al., 2007) there are 349, 223, 358 and 247 separately reported QTL for milk yield, somatic cell count, protein content and fat content, respectively. However, for carcass weight, carcass fat percentage and carcass muscle percentage there are only 135, 14, and 11 separately reported QTL. In spite of this, many of the associated SNPs from the present study were detected in known QTL for these traits (48/134). These SNPs can help to further refine QTL that have been detected using microsatellite markers (e.g. (Gutierrez-Gil et al., 2009)). The detection of a number of known QTL in our study, would suggest that our methodology was effective.

3.4.5 Candidate genes

Using both statistical approaches, a number of associations detected for each trait were in close proximity to genes with a known role in animal growth (e.g. GHR, Insulin and IGF2). As well as this, a number of novel candidate genes were identified. For example, significant SNPs on chromosome 20 were detected within 1 MB of FGF11 and on chromosome 6 approximately 500kbs away from Gonadotropin-releasing hormone receptor.

3.4.5.1 Glucagon gene

Three novel associations with carcass fat were detected on chromosome 2, all of which were within a 3.5Mb region upstream of the glucagon gene. In the same region, 5 SNPs that were associated with cull cow carcass weight were all within a 2.9Mb region of the glucagon gene. The glucagon gene plays an important role in a number of biological processes related to metabolism and energy homeostasis (Tan et al., 2009). Glucagon is known to regulate fat metabolism via cAMP-dependent mechanisms in animals (Tan et al., 2009).

3.4.5.2 Leptin gene

A number of associations detected from the Bayesian approach, that were not detected from the SSR approach, occurred in regions containing genes previously reported to be associated to growth in Holstein cows (e.g. leptin gene (Clempson et al., 2011)). Interestingly, associations from the Bayesian method that were not detected using the SSR approach, also occurred in close proximity to the leptin receptor (approx. 300kb upstream). A mutation in the leptin receptor has previously been reported to cause obesity in humans (Clement et al., 1998). Leptin is involved in the hypothalamic control of energy homeostasis, an indicator of body fat reserves and regulator of energy expenditure (Delavaud et al., 2002). In ruminants, such as cattle, a positive correlation has been demonstrated between circulating concentrations of leptin and fat accumulation (Geary et al., 2003).

3.4.6 Over-represented KEGG pathways

Carcass traits are governed by many complex biological systems, reflecting the combined influence of many genetic factors. However, there may be central biological processes that link together the genetic regulation of all of these traits. The combined trait analysis thus gave us greater power to detect pathways that control different aspects of each trait in combination but were not detected using the individual trait dataset (e.g. tyrosine metabolism (Figure 3.3)).

3.4.6.1 Peroxisome proliferator-activated receptor signalling

Peroxisome proliferator-activated receptor (PPAR) signalling pathway was the most significantly over-represented pathway (p = 0.00096) in both the analysis involving CFAT and the combined trait analysis (p = 0.00202). PPARs are a group of transcription factors that play an essential physiological role in the regulation of cellular differentiation, development and lipid metabolism (Berger and Moller, 2002). The PPAR signalling pathway is one of the most important mechanisms of adipocyte tissue development and lipogenesis (Canovas et al., 2010). Interestingly, genes from the CWT, CFAT and CONF gene datasets were also in this pathway suggesting that PPAR may also play a role in each of these traits (Figure 3.4). This was not unexpected given the known genetic associations among these traits (Pabiou et al., 2012).

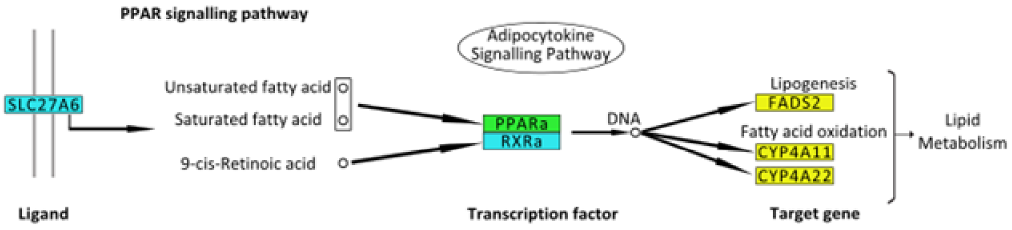


Figure 3.4 Genes from the PPAR signalling pathway that were in regions surrounding SNPs associated with at least one trait using the Bayesian approach. Genes that are coloured in blue, green and yellow were within 500kbs of a SNP associated with carcass conformation, carcass weight and carcass fat, respectively. The complete figure of the PPAR signalling pathway, showing all genes in this pathway, is contained in Figure 3.2.

3.4.6.2 Phosphatidylinositol signalling system

Another interesting pathway was the phosphatidylinositol signalling system (Figure 3.5). This pathway was significantly over-represented for both the CONF ($p = 0.0118$) and CWT ($p = 0.0382$) datasets but was also significantly over-represented ($p = 0.004$) in the combined trait analysis. The phosphatidylinositol signalling system has been found to be enriched for genes that were differentially expressed for growth and fatness traits in pigs (Canovas et al., 2010). This pathway, along with pathways significantly over-represented from the combined trait dataset, may contain core biological processes linked to phenotypic variation observed in each of the traits under investigation.

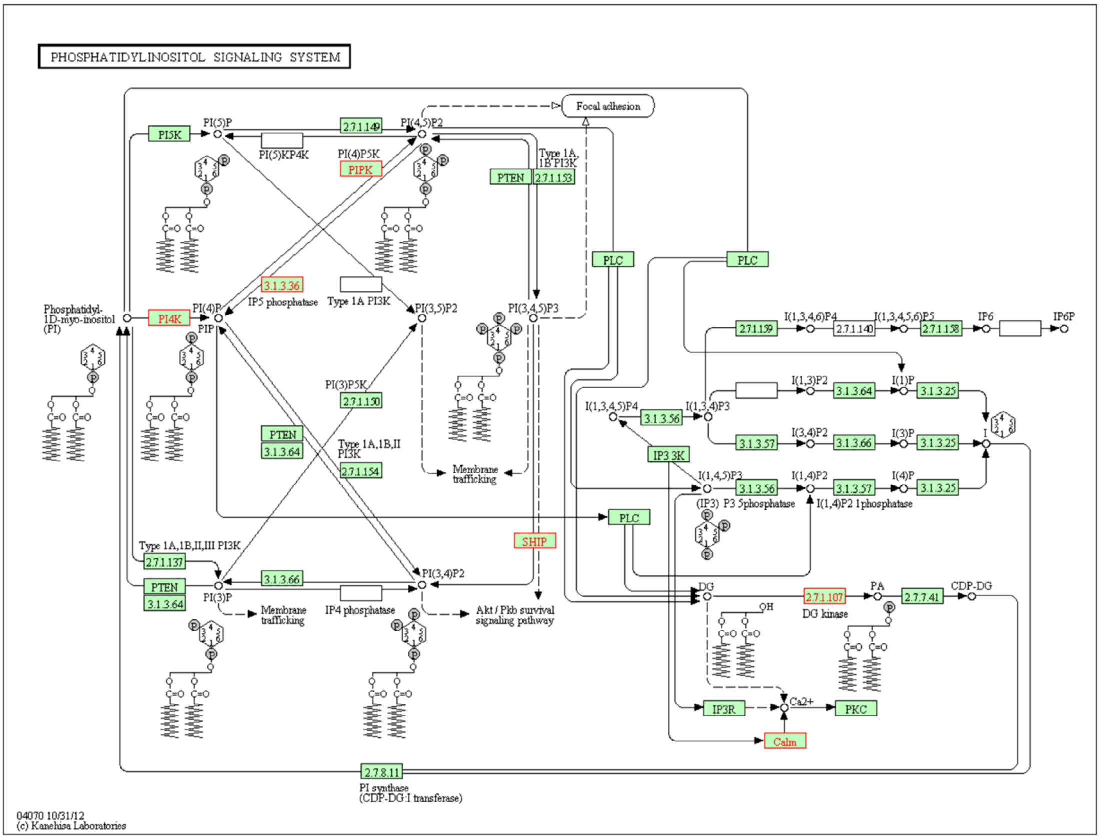


Figure 3.5 The phosphatidylinositol signalling system. This pathway was significantly over-represented in the CONF, CWT and the combined trait analyses. Genes in regions surrounding significant SNPs from the Bayesian analysis are highlighted in red.

3.4.7 Conserved biological function

There are numerous examples of single genes (or mutations in a gene) influencing similar phenotypes in different species. Some well known examples include the control of hair colour by the melanocortin receptor gene (Mc1r) in humans (Yamaguchi and Hearing, 2009), with similar effects on coat colour in species such as cattle (Klungland et al., 1995), pigs (Kijas et al., 1998) and horses (Marklund et al., 1996). For complex traits, there is little known on the conservation of genes with low to moderate effects on a phenotype across species. However, there are number of examples that suggest a degree of conservation of gene classes between mammalian species (*e.g.* stature (Pryce et al., 2011) and milk proteins (Lemay et al., 2009)) exists (Pryce et al., 2011). From our study, we have identified some well known biological processes that influence similar traits in humans such as PPAR signalling and its influence in fat deposition and metabolism. A number of pathways with a novel association in cattle, but with known effects in other organisms have also been identified (e.g. Hedgehog signalling pathway). The Hedgehog signalling pathway has been implicated in the determination of human height (Weedon et al., 2008) and mammalian adipogenesis (Rosen, 2006). This would suggest that a number of the biological processes influencing growth characteristics that are conserved in organisms such as humans are also conserved in cattle and are likely to be candidates for further study in cattle.

Chapter 4: Convergence diagnostics for a Bayesian model used in genetic prediction

4.1 Introduction

In recent years, improved computational resources and analysis techniques, such as Markov chain Monte Carlo (MCMC), have enabled the application of Bayesian approaches in many areas of science (Dodds and Vicini, 2004; Stephens and Balding, 2009). Bayesian approaches have become established in a wide variety of scientific fields including pathway modelling (Ko et al., 2009), phylogenetics (Nylander et al., 2008), and genetics (Marjoram and Tavare, 2006; O'Hara et al., 2008; Stephens and Balding, 2009). In particular, Bayesian approaches have gained popularity in genome-wide association studies (GWAS) and genomic prediction of complex traits in animal and plant breeding (known as genomic selection) (de Los Campos et al., 2013; Gianola, 2013; Knurr et al., 2013).

To date, the majority of GWAS have used single marker (usually a single nucleotide polymorphism (SNP)) regression models to identify important genomic regions associated with a particular phenotype (e.g. Jiang et al., 2010; Snelling et al., 2010; Kim et al., 2011). Single marker regression models test each SNP individually for an association with a phenotype (McCarthy et al., 2008), however as complex traits are likely to be influenced by a large number of genes, models that analyse all markers simultaneously should provide more accurate results than models that analyse one or a few markers at a time (van den Berg et al., 2013). Many Bayesian regression models have been developed for genomic prediction, collectively referred to as the “Bayesian alphabet” (Gianola et al., 2009). Originally, these models included BayesA and BayesB (Meuwissen et al., 2001), however several variations of these, that differ in their prior specifications, have been proposed for use in genomic selection (see de Los Campos et al., 2013; Gianola, 2013).

Despite the extension of the Bayesian alphabet and the growing popularity of MCMC methods in genetic association studies and genomic prediction, the use of convergence diagnostics is still relatively uncommon and not well established in the field. Convergence diagnostics relate to the idea that a Markov chain, after a sufficient number of iterations, will eventually converge to a stationary phase (target distribution) (Oszkiewicz et al., 2012). Starting from any point in the parameter space, once convergence has been achieved, all further samples will be concentrated around the mode of the stationary phase.

One of the central issues associated with MCMC inference is the length of the burn-in period, that is the number of iterations before the chain reaches a stationary phase (Brooks and Roberts, 1998). Samples taken from the burn-in do not reflect samples taken from the stationary phase, and as such, to reduce potential bias caused by the effect of starting values (seeds), iterations within the burn-in period are usually discarded before making posterior inferences (Cowles and Carlin, 1996; Brooks and Roberts, 1998; Dodds and Vicini, 2004).

In general, there are two approaches to identifying convergence: monitoring of a single chain for a long time or monitoring many chains for shorter periods of time (Oszkiewicz et al., 2012). The latter usually involves running the Markov chain repeatedly from different initial states (seeds) and ensuring that all of the chains converge to approximately the same location in the parameter space. It is however, not always computationally feasible to run multiple chains concurrently. For the purpose of this chapter, I will focus on the former, which involves running a single chain for a long time and checking if it has reached a stationary phase. Graphical methods to assess convergence are commonly used as they are easily applied to different Bayesian approaches and usually straightforward to interpret. Graphical methods usually entail plotting a sampled parameter against the number of iterations, and are referred to as a trace plots (also called history plots). If the parameter has converged to a stationary phase, the trace plot will move around the mode of the stationary phase. There is no one conclusive diagnostic that can provide assurance of convergence, in fact convergence of a number of parameters, and not just those of interest should be checked before making any inferences (Cowles and Carlin, 1996; Gill, 2008).

In this chapter, we implement a number of metrics that can be used to assess convergence, and identify the burn-in phase graphically. Additionally, we investigate the influence of prior hyper-parameters on convergence and model optimisation in the Bayesian framework before providing recommendations on addressing each of the aforementioned issues (assessing convergence, length of the burn-in phase, and influence of priors). Although, for the purpose of this study, we focuses on BayesB (Meuwissen et al., 2001), the methods outlined here could easily be implemented in any analysis utilising a Bayesian approach for genomic prediction.

4.2 Materials and Methods

4.2.1 Raw data

In chapter 3, a genome-wide association study was performed using two separate statistical approaches; a linear regression and Bayesian approach. The raw genotypic and phenotypic data used in chapter 3 is the same as the raw data used in this chapter. Briefly, genotypes of 54,001 SNPs from 5,706 Holstein-Friesian sires were available for analysis. Four economically important phenotypes, in the form of predicted transmitting abilities (PTAs), were available for each of the sires; carcass weight (CWT), carcass fat (CFAT), carcass conformation (CONF) and cull cow carcass weight (CULL). Following a series of filters to remove poor quality data, genotypes of 42,477 SNPs from 1,061 animals were available (section 3.2.2 and section 3.2.3; Table 3.1).

4.2.2 Estimated SNP effects

Two statistical approaches, a frequentist linear regression (SSR) and Bayesian approach (BayesB) were used to quantify SNP effects and marker associations for each trait separately. The two approaches are outlined in greater detail in sections 3.2.4.1 (SSR) and 3.2.4.2 (BayesB). SNP effects can be negative, positive or zero, depending on whether the SNP is estimated to be negatively associated, positively associated or not associated with the trait respectively.

4.2.2.1 SSR

In short, the dependent variable (PTA) was regressed on each SNP separately in a linear mixed model accounting for relationships between animals using the additive genetic relationship matrix in ASReml (Gilmour et al., 2009a). The individual animal was included as a random effect. The estimated SNP effect for each SNP was retained. Estimated p-values were corrected for errors arising from multiple testing using the FDR approach described by Storey and Tibshirani. Corrected p-values (termed q-values) < 0.05 were defined as significant.

4.2.2.2 BayesB

The second statistical approach was BayesB (Meuwissen et al., 2001), which included all SNPs simultaneously in a Bayesian mixture model. An inverse chi-squared distribution (v = 4.234, S = 0.0429) was included in the model as the prior distribution of mean and genetic variation of each SNP included in the model. A prior value (1-π) was included in the Bayesian analysis as the prior probability of association for each SNP. As this value is assigned to all SNPs in the analysis, it represents a prior proportion of SNPs believed to be associated with the trait under investigation.

4.2.2.3 Specifying a value of 1-π *a priori*

To investigate the impact of the prior value (1-π) on convergence, analyses were run with incrementally decreasing values of (1-π) ranging from 5.0×10^{-2} to 6.25×10^{-5}, representing a decreasing proportion of SNPs assumed to be associated with the traits being examined (Table 4.1). To assess the usefulness of including information from separate analyses such as a single SNP regression, additional analyses were run using the proportion of SNPs found to be significantly (q < 0.05) associated with each trait from the SSR analysis (pSSR), and also half and double these values (pSSR/2 and pSSR*2, respectively), to explain (1-π). In total, eleven analyses were performed for each trait. For convenience, these analyses have been labelled pA-pK representing an increasing value of (1- π). From this point onward, these prior values will be referred to using the label assigned in Table 4.1.

Table 4.1 Values assigned to (1- *π*) for each Bayesian analysis. CWT = carcass weight; CFAT = carcass fat; CONF = carcass conformation; CULL = cull cow carcass weight; * = pSSR/2; § = pSSR; † = pSSR*2. From this point onward, these prior values will be referred to using the label assigned in the first column.

Label	CWT	CFAT	CONF	CULL
pA	2.35×10^{-5}*	6.25×10^{-5}	6.25×10^{-5}	6.25×10^{-5}
pB	4.71×10^{-5}§	1.25×10^{-4}	1.25×10^{-4}	1.25×10^{-4}
pC	6.25×10^{-5}	2.5×10^{-4}	2.5×10^{-4}	2.5×10^{-4}
pD	9.42×10^{-5}†	2.94×10^{-4}*	5.0×10^{-4}	5.0×10^{-4}
pE	1.25×10^{-4}	5.0×10^{-4}	1.0×10^{-3}	5.65×10^{-4}*
pF	2.5×10^{-4}	5.89×10^{-4}§	2.45×10^{-3}	1.0×10^{-3}
pG	5.0×10^{-4}	1.0×10^{-3}	5.69×10^{-3}*	1.13×10^{-3}§
pH	1.0×10^{-3}	1.18×10^{-3}†	1.0×10^{-2}	2.26×10^{-3}†
pI	2.45×10^{-3}	2.45×10^{-3}	1.14×10^{-2}§	2.45×10^{-3}
pJ	1.0×10^{-2}	1.0×10^{-2}	2.27×10^{-2}†	1.0×10^{-2}
pK	5.0×10^{-2}	5.0×10^{-2}	5.0×10^{-2}	5.0×10^{-2}

4.2.2.4 Posterior probabilities

A posterior probability (PP) is indicative of the strength of evidence from the posterior distribution that a SNP is associated with a particular phenotype (Meredith et al., 2012; Yang and Tempelman, 2012). A PP ranges from 0 (no association) to 1 (strong association). The PP for a SNP at iteration *t* is calculated as follows: The number of times that a SNP had a non-zero effect (i.e. was calculated to have some association with the trait) in all sampled iterations (in our analyses every 500th iteration was sampled) up to iteration *t* is counted. This value was then divided by the total number of iterations sampled for each SNP (up to iteration *t*), resulting in a value which was the posterior probability of the SNP having an effect (Newcombe et al., 2012).

During the MCMC portion of the Bayesian analysis, the effect (on the phenotype) calculated for each SNP can vary widely. At different iterations the effect calculated for a SNP may be zero or non-zero, converging to a stationary probability (proportional to the background distribution) as the chain reaches the stationary phase. For example, after 100,000 iterations (sampled every 500th) a SNP, *x*, might have a non-zero effect 75% of the time ($PP_x = 0.75$) but after 200,000 iterations the same SNP might only have a non-zero effect in 50% of the sampled iterations ($PP_x = 0.5$). Therefore, ensuring the posterior estimates of individual SNPs have reached stationarity is important for the calculation of (1-π) for the entire dataset. Generally, the burn-in phase once identified would be discarded and posterior probabilities for each SNP calculated from the remaining sampled iterations.

4.2.3 Assessing convergence and effect of different prior (1-π) specifications

For each sampled parameter, graphical assessment of convergence (stationarity) was assessed by plotting the parameter against the iteration number. If the parameter has converged, the resulting plot should reach a value in the parameter space (i.e. all of the possible values that the parameter can reach) and remain close to this value as the number of iterations increases. Convergence of the model was identified by a plateau in the estimated parameter as the number of iterations increased.

In addition to assessing convergence, four metrics were also used to investigate the influence of prior specifications and convergence; summed absolute log likelihood, the number of SNPs with a PP>0.5; Euclidean distance and Coefficient of determination. For each MCMC chain, each of these metrics was calculated at every sampled iteration. A surface plot was created of the results for each sampled iteration plotted against the prior specification of (1-π) used for that analysis. The axis with the prior specification of (1-π) was ordered from pA (smallest) to pK (largest) denoting increasing prior proportions of SNPs assumed to have an effect on the trait of interest. This allowed the effect of different prior (1-π) specifications to be observed across increasing iterations of the MCMC chains. Detailed descriptions of each of the metrics are as follows:

4.2.3.1 Summed SNP effects

As previously described, SNPs have positive, negative or zero effects based on their estimated association with the trait. For the purposes of genomic selection for particular traits, the sum of all effects (multiplied by the genotype) for the alleles possessed by a single animal is used as an estimate of the genetic potential (gEBV) of the animal for that trait (Equation 4.1). The gEBV of an animal, *k*, at iteration, *t*, is given as:

$$gEBV_{kt} = \sum_{i=1}^{n} z_{ki} x_{it} \qquad [4.1]$$

Where Z_{ki} is the genotype (expressed as 0, 1 or 2) for animal *k* at SNP *i*. x_{it} is the effect of SNP *i* at iteration *t*, and n is the total number of SNPs included in the analysis.

The gEBV has previously been used to assess convergence in Bayesian MCMC analyses(Daetwyler et al., 2010). However, the rate of convergence might vary between animals, meaning that a separate plot of the gEBV would need to be produced for each animal included in the analysis. This is not practical as most analyses require data from hundreds of animals. Additionally, assessment of a single parameter may not be representative of the model as a whole. To investigate this, the sum of all SNP effects (SSE) (Equation 4.2) and separately the sum of all absolute SNP effects (SAE) (Equation 4.3) was calculated at each iteration sampled. The SNP effects were calculated using BayesB (Meuwissen et al., 2001). The summed values were plotted for all the iterations samples to identify convergence to stationarity.

$$SSE_t = \sum_{i=1}^{n} x_{it} \qquad [4.2]$$

$$SAE_t = \sum_{i=1}^{n} |x_{it}| \qquad [4.3]$$

Where, x_{it} and $|x_{it}|$ is the effect and absolute effect, respectively, of SNP *i* at iteration *t*, and n is the total number of SNPs included in the analysis.

4.2.3.2 Summed absolute log-likelihood

For every sampled iteration, a log-likelihood score is calculated for each individual SNP effect. This value is the log of the likelihood of the SNP having the effect estimated, given parameters such as the phenotypic data, genotypic data and other variables (called “the model” from this point on). A log-likelihood value for the entire model at a single iteration was calculated by summing the individual absolute log-likelihoods for all SNPs. To investigate convergence from a likelihood perspective, the summed absolute log-likelihood for the model at each sampled iteration was plotted against the iteration number to produce a trace plot. The resulting plot was visually inspected to assess convergence (stationarity), evident as a plateau in the summed absolute log-likelihood as the number of iterations increased (See Figure 4.1 for examples). This approach was then repeated for each analysis involving BayesB.

For each trait, these values were combined from each analysis using a different prior (1-π) specification to create a surface plot. These surface plots were used to assess the appropriateness of each of the prior values of (1-π) for each of the respective traits.

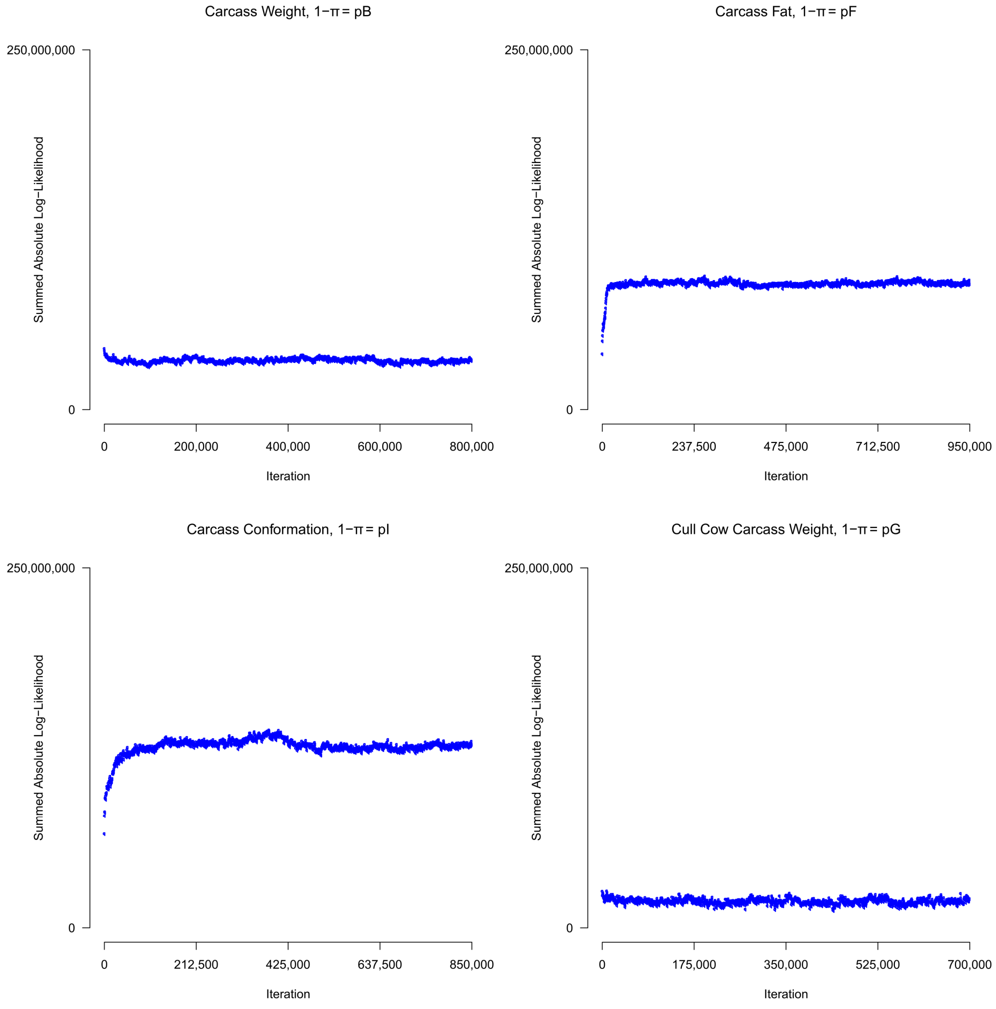


Figure 4.1 Summed absolute log-likelihood plots for all analyses with prior value π equal to pSSR. For each figure, the model log-likelihood (y-axis) is plotted by the iteration number (x-axis).

4.2.3.3 The number of SNPs with PP>0.5 metric

The number of SNPs with a posterior probability (PP) greater than 0.5 was calculated at each sampled iteration. For each sampled iteration, this was calculated across all sampled iterations up to this point. These values were plotted to produce a trace plot for each MCMC chain. For each trait separately, these values from each analysis were combined to create a surface plot. These surface plots were used to assess the influence and appropriateness of each of the prior values of (1-π) for each of the respective traits. The most appropriate prior is identified as an elevation in the plot surface across prior values, while also converging to stationarity as the number of iterations increases

Additionally, to highlight the influence of failing to remove iterations from the burn-in, the mean and mode number of SNPs with a PP>0.5 was calculated using the estimated value from all sampled iterations and separately the final 100,000 sampled iterations from each analysis. Sample means close to the mode illustrate mixing around the mode which is indicative of convergence.

4.2.3.4 Euclidean distance metric

A Euclidean distance (ED) (Equation 4.4) calculation was used to assess the similarity of the SNP effects from the SSR analysis and each of the iterations from the Bayesian analysis. This approach assumes that the SNP effects of individual SNPs calculated from the SSR approach are a reasonable estimation of the real effect, given the trait being examined. The problem with SNP effects calculated from the SSR analyses however, is that they are non-independent of neighbouring SNPs, producing the characteristic distributions of effect across neighbouring SNPs. To address this issue, we only compared the estimated SNP effects for those SNPs with a calculate PP>0.5 from the Bayesian analysis to the effect calculated by the SSR approach for the same SNP. The estimated effects for these SNPs from the Bayesian approach ($x_1, x_2, \ldots, x_n$) were compared to those calculated from the SSR approach ($y_1, y_1, \ldots, y_n$) using a Euclidean distance. This approach assumes that as the Bayesian MCMC chain approaches convergence, the calculated SNP effect for all SNPs with a PP>0.5 should approach those calculated by the SSR approach.

The ED was calculated as follows: Given points $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_1, \ldots, y_n)$ in n-dimensional space, the ED is a measure of the proximity (or similarity) of x and y in the space. The position of x in n-dimensional space is given by the Euclidean vector ($x_1, x_2, \ldots, x_n$). Similarly, the position of y is given by the Euclidean vector ($y_1, y_1, \ldots, y_n$). The ED between points x and y in n-dimensional space can be defined mathematically as:

$$d(x,y) = d(y,x) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad [4.4]$$

Where x is a Euclidean vector of SNP effects estimated from the Bayesian approach and y is Euclidean vector SNP effects estimated from the SSR for the same set of SNPs. A Euclidean distance of 0 indicates that the set of values, x and y, occupy the same location in the n-dimensional space and are exactly the same. This was calculated for each sampled iteration and combined for each prior (1-π) specification to produce a surface plot which was used to assess convergence and the impact of (1-π) on estimates of the ED.

4.2.3.5 Coefficient of determination metric

The Coefficient of Determination (the well known R^2; Equation 4.5) was also used to assess convergence. Similar to the approach used to calculate the ED metric, at every iteration, the effects estimated from the Bayesian approach of all SNPs with a PP>0.5 ($x_1, x_2, \ldots, x_n$) were compared to the effects estimated from the SSR for the same set of SNPs ($y_1, y_1, \ldots, y_n$). The R^2 between these data sets was then calculated as follows:

$$R^2 = \left\{ \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} \sqrt{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}} \right\}^2 . \qquad [4.5]$$

Where, x_i and y_i are the SNP effects of SNP i from the Bayesian and SSR approach respectively and where $\overline{x}$ is the mean SNP effect estimated from the Bayesian approach and $\overline{y}$ is the mean SNP effect from the SSR approach. The total number of SNPs being analysed (the number with a PP>0.5 from the Bayesian analysis) in x and y is denoted by n. As with other metrics analysed the R^2 of the SNP effect from two approaches was calculated at each sampled iteration. This was carried out for each prior (1-π) specification analysed and visualised as a surface plot for each trait. The surface plot was then used to assess convergence of the metric and separately which prior values of (1-π) appeared optimised for this metric.

4.2.4 Alternate starting conditions

Another approach to assess convergence in a MCMC analysis involves running multiple Markov chains with the same model prior specifications which differ only in the starting seed for the random number generator used in the algorithm. Evidence of the multiple chains converging to approximately the same sample space is used to assess convergence and there are several statistics that are generally used for this purpose (For a review, see Cowles and Carlin, 1996). However, a problem was identified with the BayesB software which required correction of the code before this could be carried out. The original code provided by the developers contained a "hard coded" seed of "123456789", this meant that every analysis carried out using the same dataset and same priors were always going to be identical (a problem which may not have been identified in other published studies using this software). To address this we changed the random number generator seed to a different value for multiple runs for the same dataset and priors. Summed absolute log-likelihoods were estimated at each iteration and plotted. These additional analyses were carried out for each trait using the prior specification (1-π) of; 1 - pSSR/2, 2.5×10^{-4}, 1 – pSSR and 1.25×10^{-4}. Convergence was assessed visually and compared to the estimated summed absolute log-likelihoods from the initial chains.

4.3 Results

4.3.1 Summed absolute log-likelihoods

For each analysis using the Bayesian approach, the summed absolute log-likelihood plot was visually inspected to identify the burn-in phase (Table 3.3). The number of iterations required to reach a stationary distribution varied between analyses ranging from approximately 60,000 iterations (for CONF trait and prior specification for (1-π) of pA) to approximately 200,000 iterations (for CONF trait and prior specification for (1-π) of pH). As an example, the trace plots for summed absolute log-likelihoods calculated using the Bayesian approach with prior value (1-π) of 1-pSSR (i.e. 1 minus the proportion of SNPs calculated by the SSR approach to have an effect) for all 4 traits are shown in Figure 4.1. The surface plots calculated with this metric for all prior specifications tested for CWT, CFAT, CONF and CULL traits are shown in figures 4.2, 4.3, 4.4 and 4.5 respectively.

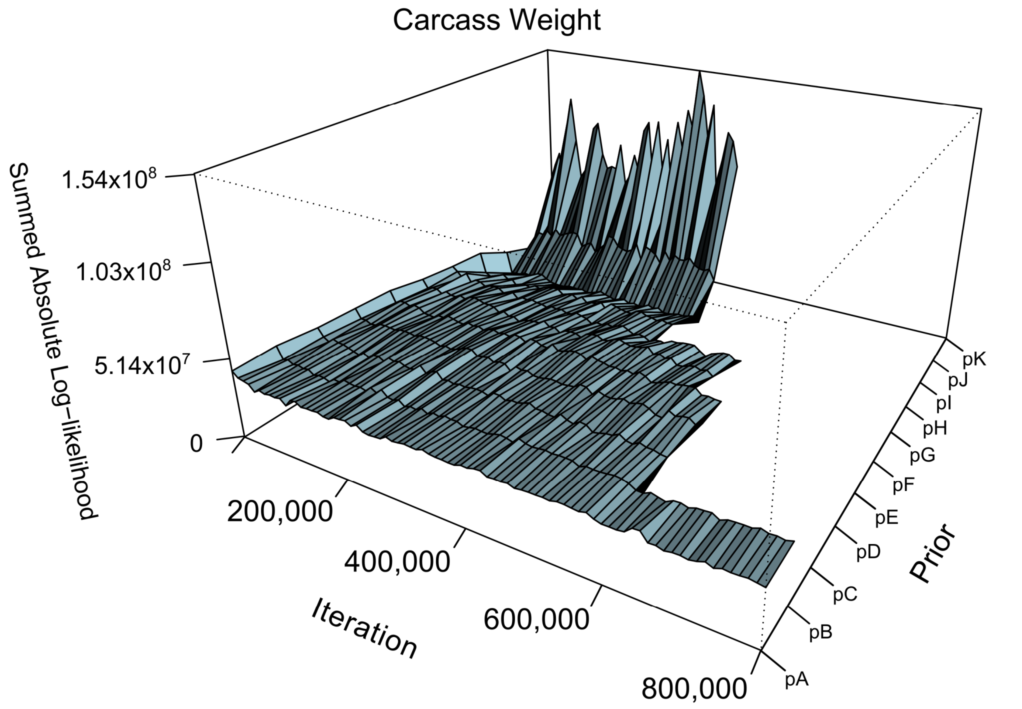


Figure 4.2 Surface plot of the summed absolute log-likelihoods for the carcass weight trait. For each prior (pA-pK), the summed absolute log-likelihood at each sampled iteration is plotted.

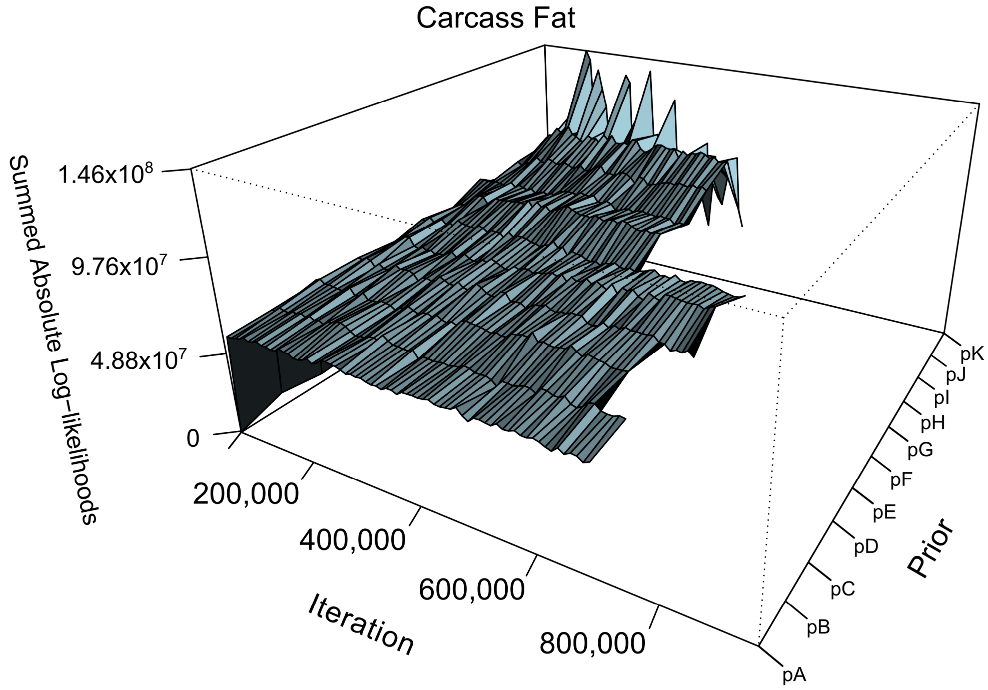


Figure 4.3 Surface plot of the summed absolute log-likelihoods for the carcass fat trait. For each prior (pA-pK), the summed absolute log-likelihood at each sampled iteration is plotted.

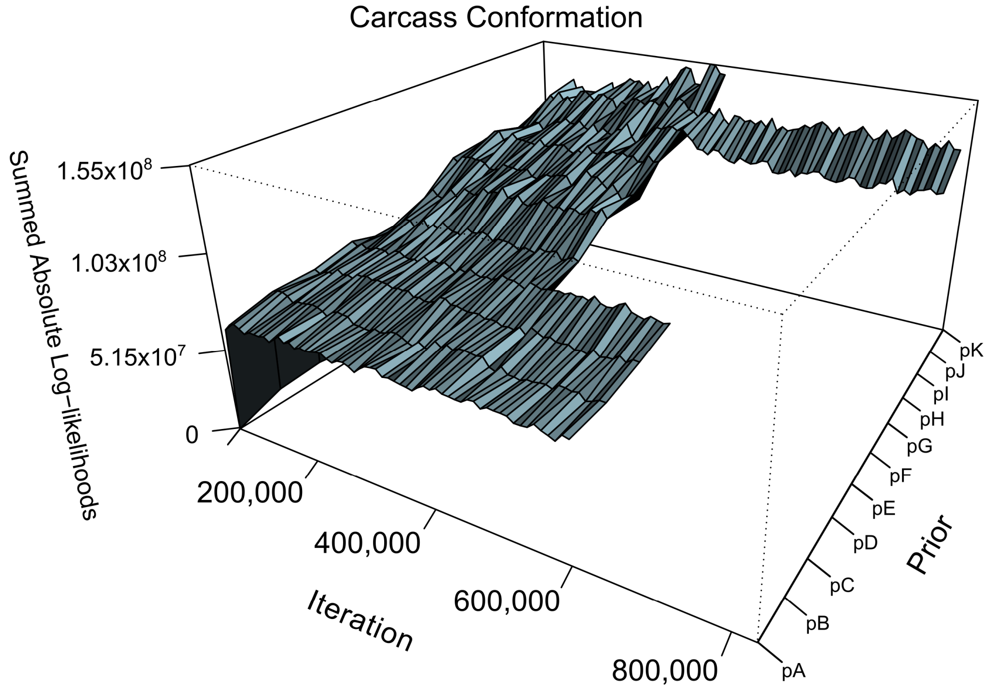


Figure 4.4 Surface plot of the summed absolute log-likelihoods for the carcass conformation trait. For each prior (pA-pK), the summed absolute log-likelihood at each sampled iteration is plotted.

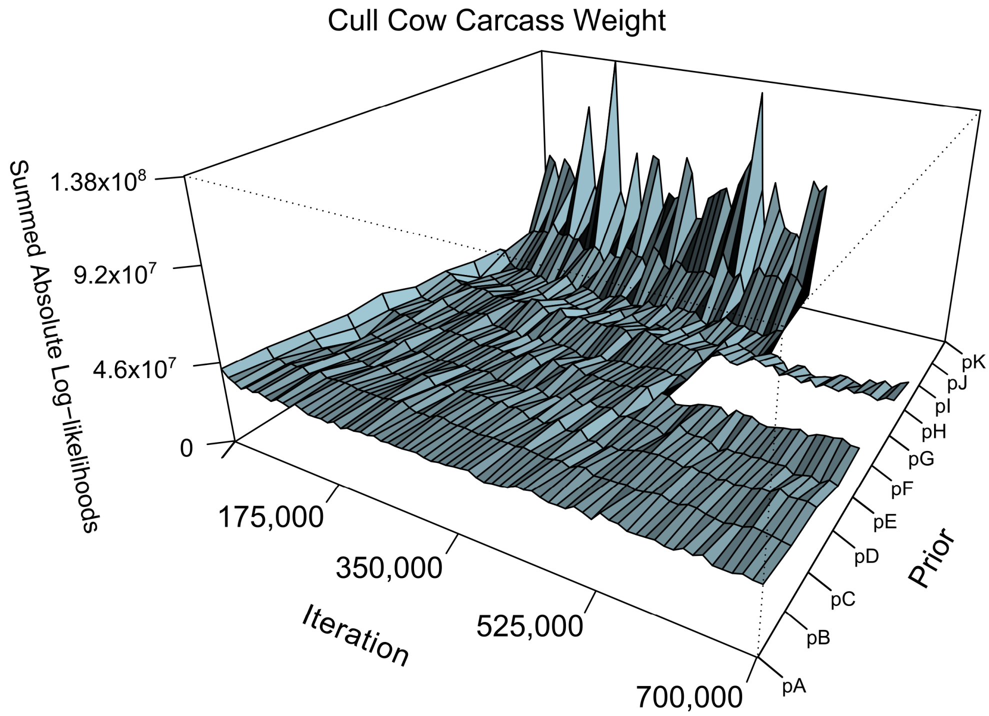


Figure 4.5 Surface plot of the summed absolute log-likelihoods for the cull cow carcass weight trait. For each prior (pA-pK), the summed absolute log-likelihood at each sampled iteration is plotted.

4.3.2 Summed SNP effects

The plots from the summed SNP effect and the summed absolute SNP effect metrics exhibited similar convergence behaviour. The stationary phase was achieved after only a few thousand iterations for most of the analyses. For example, in the analysis of carcass conformation when using a prior specification of twice that estimated by the SSR approach ((1-π) = pJ), both metrics reached a stationary phase after approximately 10,000 iterations (Figure 4.6).

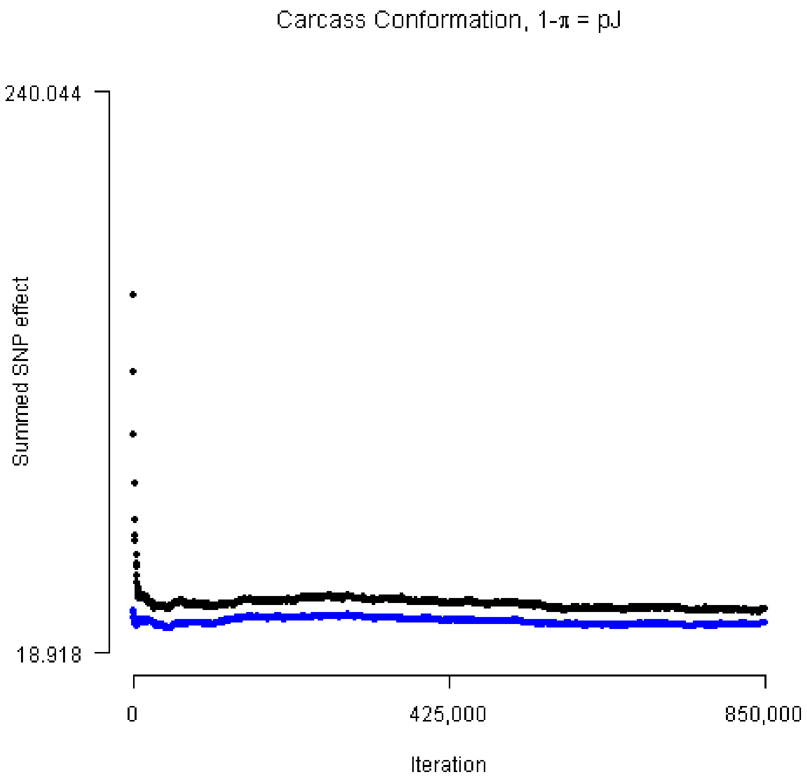


Figure 4.6 Summed absolute SNP effects and the summed SNP effects at each iteration. The summed absolute SNP effect at each iteration for the analysis with prior (1-π) = pJ is plotted in black. The summed SNP effect at each iteration for the analysis with prior (1-π) = pJ is plotted in blue. The iteration number is plotted on the x-axis and the corresponding effect on the y-axis.

4.3.3 The number of SNPs with a PP > 0.5

Early in the iteration space, the number of SNPs with a PP>0.5 at each iteration varied considerably between analyses (specifications of 1-π) of a single trait (Figure 4.7; Appendix 1). This was evident in the surface plot created for each trait. In general, the differences between the values at the same iteration of each prior (1-π) specifications were greatest at the early iterations. By iteration 400,000 the estimates between analyses using different priors for the same trait were similar, although still varied. This was the case for all analyses using this metric.

For example, at iteration 25,000 of the CFAT analyses there were 5, 19 and 313 SNPs that had a PP>0.5 using the prior values (1-π) of pA, pG and pK, respectively (Figure 4.7). However, by iteration 400,000, the number of SNPs with a PP>0.5 for the same analyses was 10, 23 and 18. This indicates that the burn-in period for the CFAT analysis with (1-π) = pK may be much longer than the analyses of (1-π) = pA or pG, and that all of the priors would have a burn-in phase longer than 25,000 iterations.

Additionally, the mean and mode estimates for the last 100,000 sampled iterations were, in general, much closer than the estimates of the mean and mode using all sampled iterations (Table 4.2; Table 4.3). This was particularly evident for all analyses using the prior value (1-π) = pK, where the difference in the mean and mode ranged from 46.51 to 146.05 using all sampled iterations compared to the last 100,000 sampled iterations which had a range from 0 to 0.77. The number of SNPs with a PP>0.5 at each iteration throughout the final 100,000 iterations were in general the same as the mode reflecting convergence to a stationary phase. This is not unexpected as the last 100,000 iterations excluded the burn-in period and convergence had been observed.

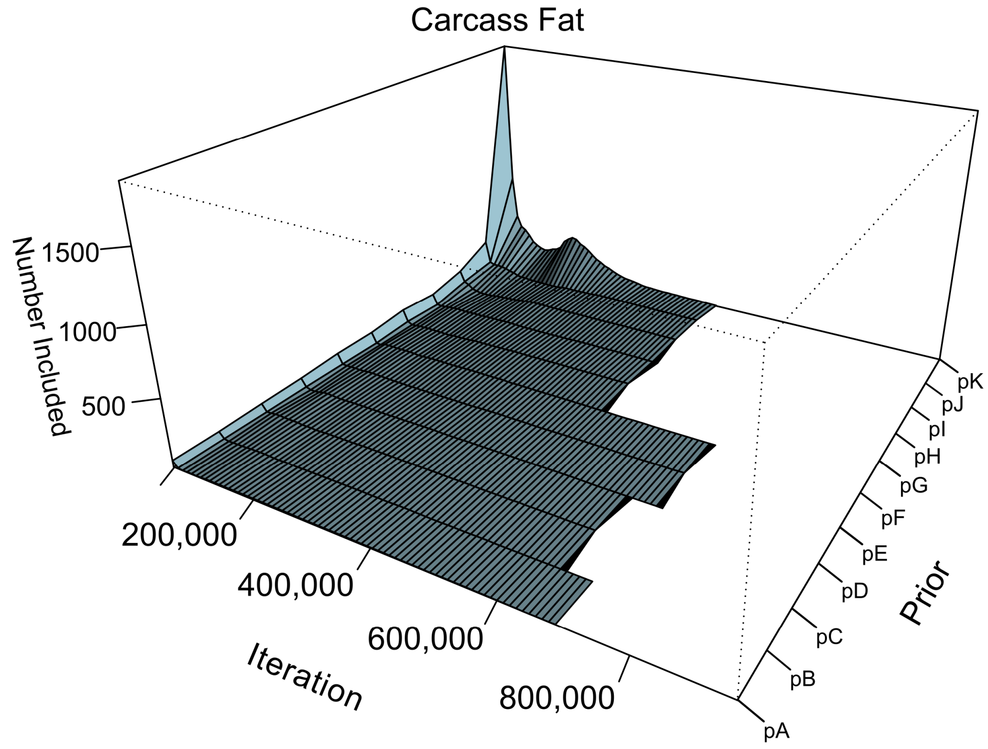


Figure 4.7 Surface plot of the number of SNPs with a PP>0.5 for the carcass fat phenotype. For each prior (pA-pK), the number of SNPs with a PP>0.5 at each sampled iteration is plotted. Surface plots of the number of SNPs with a PP>0.5 for the remaining traits are included in Appendix 1.

Table 4.2 The mean and mode number of SNPs with a PP>0.5 in each analysis. The mean and mode number of SNPs with a PP>0.5 was calculated using all sampled iterations for each analysis. (1 – π) = prior proportion (in order or size from smallest (pA) to largest (pK)) of SNPs assumed to be associated with a trait. CWT = carcass weight; CFAT = carcass fat; CONF = carcass conformation; CULL = cull cow carcass weight.

	CWT		**CFAT**		**CONF**		**CULL**	
1-π	**mean**	**mode**	**mean**	**mode**	**mean**	**mode**	**mean**	**mode**
pA	13.57	13	9.49	11	15.97	12	14.65	18
pB	11.46	11	9.19	6	24.38	24	9.64	8
pC	13.05	12	13.9	14	28.18	29	14.15	9
pD	13.51	11	13.16	10	31.25	29	18.07	17
pE	12.42	12	14.65	15	24.38	24	23.2	24
pF	18.69	19	22.02	20	28.18	29	26.41	29
pG	22.47	22	31.63	37	53.06	44	19.05	16
pH	27.19	22	30.56	33	74.08	67	24.11	24
pI	43.47	47	38.71	43	53.06	44	29.37	29
pJ	26.87	22	48.86	42	68.38	69	22.22	9
pK	49.03	0	151.88	14	227.05	81	46.51	0

Table 4.3 The mean and mode number of SNPs with a PP>0.5 for the last 100,000 iterations in each analysis. The mean and mode number of SNPs with a PP>0.5 was calculated using only the sampled iterations from the last 100,000 iterations. (1 – π) = prior proportion (in order or size from smallest (pA) to largest (pK)) of SNPs assumed to be associated with a trait. CWT = carcass weight; CFAT = carcass fat; CONF = carcass conformation; CULL = cull cow carcass weight.

	CWT		**CFAT**		**CONF**		**CULL**	
1-π	**mean**	**mode**	**mean**	**mode**	**mean**	**mode**	**mean**	**mode**
pA	13	13	5.33	5	11.92	12	10.76	10
pB	10.36	10	5.63	6	23.75	24	11.84	12
pC	11.83	12	10.66	11	27.18	27	9.25	9
pD	10.98	11	10.42	10	29.32	29	16.86	17
pE	10.02	10	14.66	15	24.33	24	21.62	20
pF	17.48	17	20	20	46.33	46	27.93	29
pG	31.59	32	38.15	38	46.51	47	16.45	16
pH	21.76	22	26.21	27	67.12	67	24	25
pI	34.25	36	43.36	43	40.61	37	26.57	29
pJ	22.78	23	42.2	42	72.09	72	9.38	9
pK	2.29	3	13.89	14	80.23	81	0	0

4.3.4 Euclidean distance

For this metric, values close to zero indicated high similarity between the SNP effects estimated from the Bayesian and SSR analyses. Similar to the number of SNPs with a PP>0.5 metric, the estimates of the Euclidean distance (ED) were largest at early iterations and varied considerably between prior (1–π) specifications (Figure 4.8; Appendix 2). However, in general, by iteration 200,000 the estimates of the ED between prior values of (1–π) were similar and each displayed convergence to a stationary phase. For the CFAT analysis (Appendix 2), the estimates of the ED varied between prior specifications even after 200,000 iterations suggesting that the estimated SNP effects might be sensitive to the values assigned to the prior (1–π).

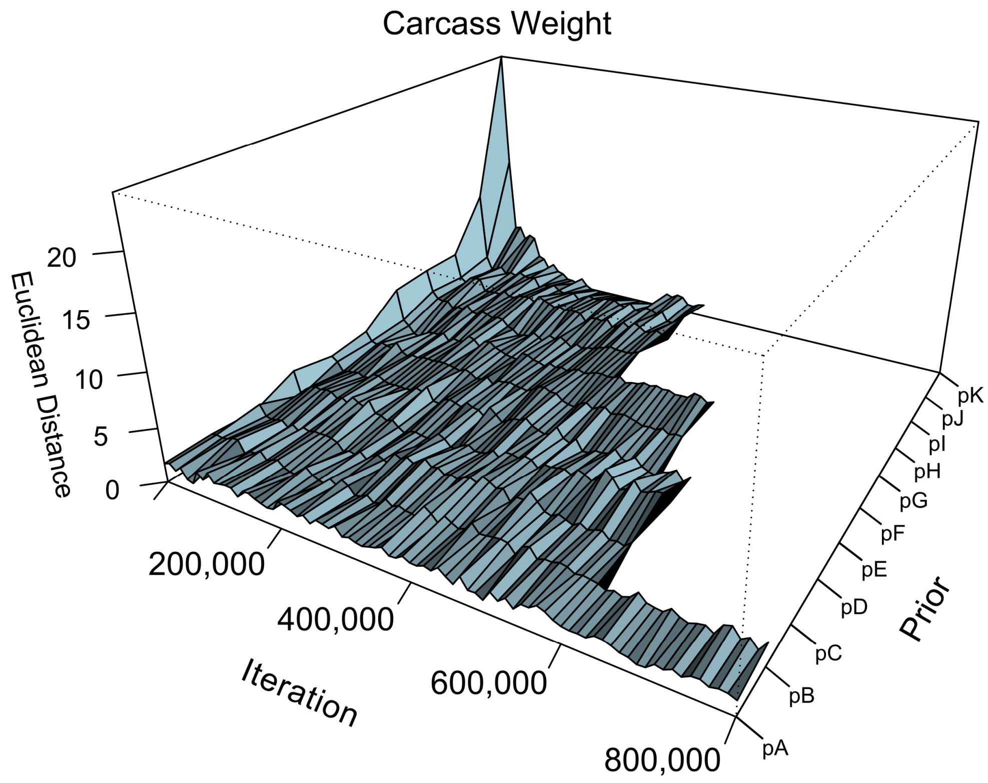


Figure 4.8 Surface plot of the Euclidean distance for the carcass weight phenotype. For each prior (pA-pK), the Euclidean distance at each sampled iteration is plotted. Surface plots of the Euclidean distance for the remaining traits are included in Appendix 2.

4.3.5 Coefficient of determination

All analyses reached convergence using the R^2 metric. The estimated values of the R^2 varied between prior (1–π) specifications throughout the iteration space for the analysis of CFAT and CONF (Appendix 3). Similar to the Euclidean distance, this indicates that the estimated SNP effects might be sensitive to the values assigned to the prior (1–π). In spite of this, estimated values across priors (1–π) were similar for the CWT analysis (Figure 4.9). This was also the case for the analysis of CULL (Appendix 3). R^2 values close to one indicate that the relationship between the estimated SNP effects from the Bayesian and SSR approach is similar. If an individual prior (1–π) was better suited than any of the other prior specifications, the R^2 value would be higher in all iterations for that prior. This would appear as a slope in the surface plot, peaking at the prior specification with values closest to 1. This was the case for only CONF analysis (Figure 4.10).

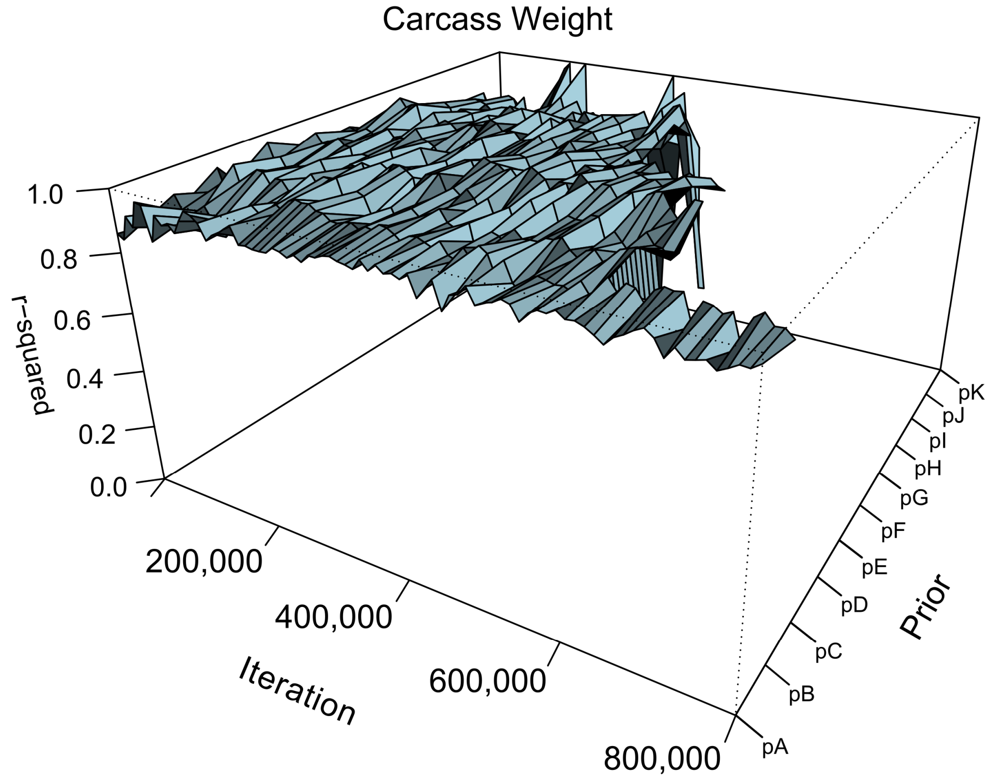


Figure 4.9 Surface plot of the R^2 for the carcass weight phenotype. For each prior (pA-pK), the R^2 at each sampled iteration is plotted.

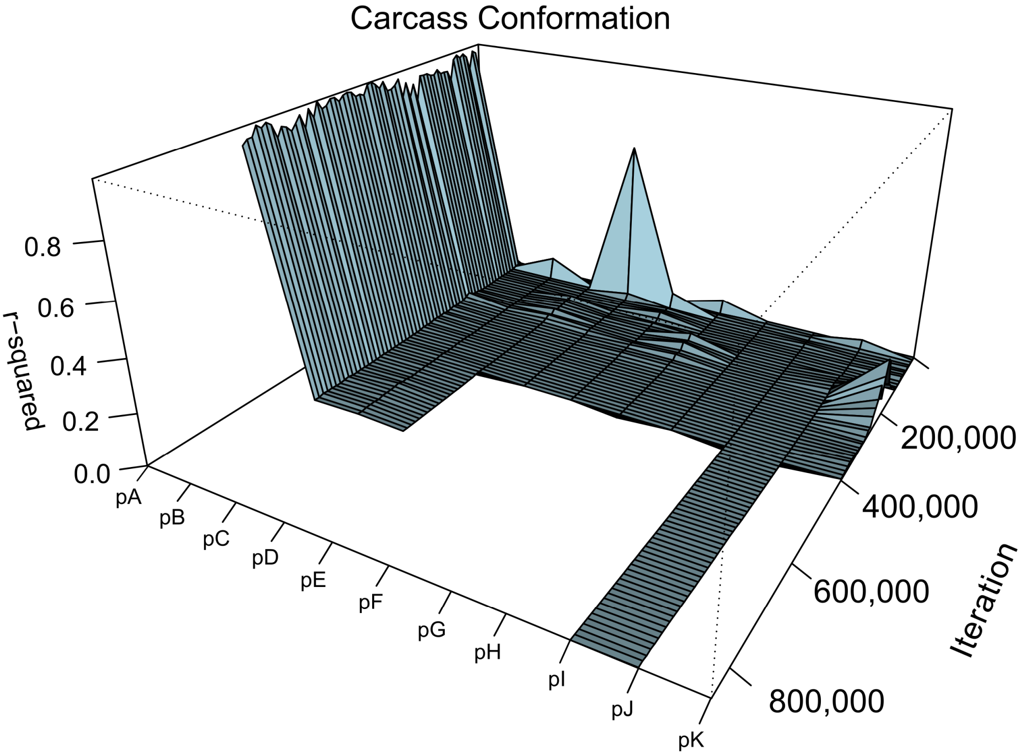


Figure 4.10 Surface plot of the R^2 for the carcass conformation phenotype. For each prior (pA-pK), the R^2 at each sampled iteration is plotted. Surface plots of the R^2 for the remaining traits are included in Appendix 3.

4.3.6 Alternate starting conditions

Convergence of both chains to approximately the same stationary phase in the parameter space was achieved for all additional analyses (Figure 4.11). This was indicated by the summed absolute log-likelihood estimates from both chains reaching approximately the same values and continuing to mix around those values. Convergence was slowest for the carcass conformation phenotype but was reached by iteration 200,000 for both chains.

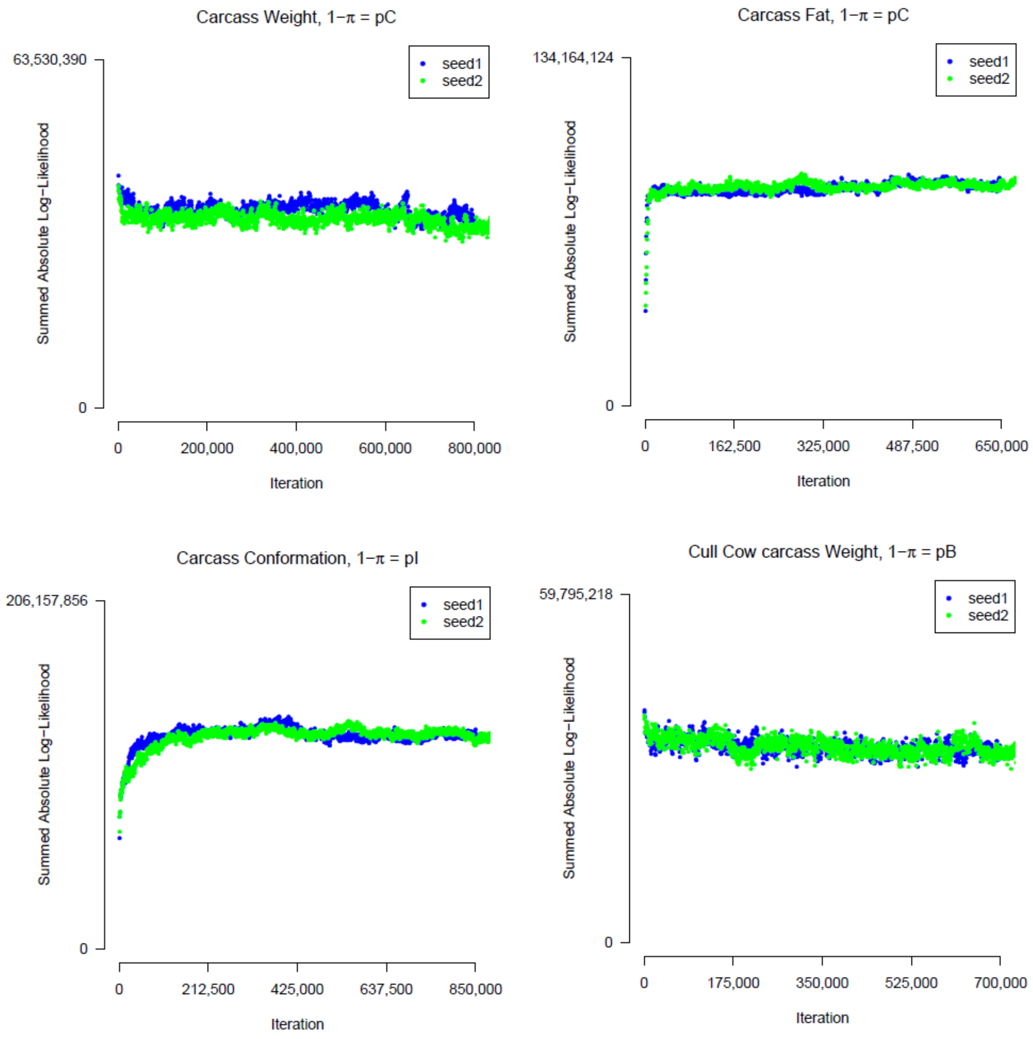


Figure 4.11 Summed absolute log-likelihood plots for the additional analyses with alternate starting conditions. For each figure, the model log-likelihood (y-axis) is plotted against the iteration number (x-axis). Seed 1 and seed 2 represent the initial states that each chain started with.

4.4 Discussion

In a relatively short time MCMC approaches have gained considerable momentum in many fields including genetic association studies. Convergence diagnostics are an important implementation of Bayesian MCMC approaches. In many fields, where MCMC approaches have been in use for a long period of time, approaches and diagnostics to assess convergence have become established (e.g. Nylander et al., 2008). The aim of this study was to develop approaches for assessing convergence that are easily implemented in models from the Bayesian alphabet. The approaches outlined in this chapter, are graphical in nature and, thus relatively easy to interpret and implement. As well as this, they provide a basis for which convergence can be assessed and a period of burn-in identified. Methods outlined in this chapter can be used to assess convergence and identify a stationary phase in the MCMC chain from which posterior inferences should be made.

4.4.1 Including information from separate analyses

Despite the statistical drawbacks of linear regression approaches, single SNP regression is the most common approach used in genome-wide association studies. This is often because computationally it is much more feasible to analyse each SNP individually as opposed to the Bayesian approach of including all SNPs simultaneously. Additionally, most researchers are more familiar with linear regression approaches and interpreting the p-values of association obtained from this type of analysis. Because of this, we used information from the SSR to estimate a number of the metrics used to assess convergence. As well this, the proportion of significant associations from the SSR approach was also used to inform the Bayesian model *a priori* (as a specification of 1-π). Using results from the SSR analysis in this way may be useful in the future as it allows researchers to incorporate statistically significant information in a Bayesian framework. This also allowed the comparison of estimates from two separate approaches that differ fundamentally in their statistical setup.

4.4.2 Summed absolute log-likelihoods

The log-likelihood is the log of the likelihood of an estimated SNP effect, given parameters such as the phenotypic data and the genotypic data. The model log-likelihood was calculated as the sum of the absolute log-likelihoods for all SNP estimates. As such, the model log-likelihood incorporates information from all parameters and not just those of interest. In this way, the model log-likelihood is a good way to assess convergence of the model.

Trace plots based on log-likelihoods have been used to monitor convergence during MCMC analysis in fields such as phylogenetics (Nylander et al., 2008; Satija et al., 2009; Qi et al., 2012) and physics (Oszkiewicz et al., 2012). The number of iterations needed to reach a stationary phase varied between models, although a greater number of iterations were required for larger values of (1-π). This indicated that the model may be sensitive to values assigned to (1-π) *a priori*. Because of this, it would not be possible to choose *a priori* the number of iterations required to reach the stationary phase. Instead the required number of iterations needed to reach convergence would need to be chosen based on the behaviour of the trace plot for each analysis separately.

4.4.3 Summed SNP effects

Metrics based on the SNP effect have been used as evidence of model convergence in previous GWAS studies e.g. (Daetwyler et al., 2010). This however does not take the entire model into account. Convergence of all parameters, not just those of interest should be considered before making any posterior inferences (Nylander et al., 2004). At first sight, the summed absolute SNP effect and the summed SNP effect seemed to provide a good estimate of convergence as both seemed to reach a stationary phase at approximately the same iteration. Additionally, for almost all analyses convergence was achieved quickly using either metric. However, convergence was not achieved in all analyses using this metric. Furthermore when convergence was achieved using these metrics, it was misleading. During early iterations of the parameter space, a large amount of SNP effects are set to zero. As the model traverses the parameter space, SNPs maintain a zero effect for long periods before being included with a non-zero effect. Similarly SNPs that had an effect in early iterations may eventually have a zero effect. The net sum of SNP effects may remain the same throughout the iteration space while these changes are occurring.

A good example of this was found using the CONF analysis with prior (1-π) = pJ. Using the SAE or SSE, convergence was indicated to have been reached by only a few thousand iterations (Figure 4.6). However, 3 SNPs, rs110195460, rs109648728 and rs41636123, had an estimated effect of zero until iteration 105,000, 63,500 and 121,500, respectively but had a non-zero effect in all iterations after 240,000 (Figure 4.12). The estimated SNP effect could be used to monitor convergence, but this would require producing a trace plot for all SNPs included in the analysis which would not be feasible for almost all analyses as thousands of SNPs are analysed simultaneously.

The genomic estimated breeding value (gEBV) has been used to assess convergence (Daetwyler et al., 2010). The gEBV for an animal is the sum of all of the estimated SNP effects times the genotype (coded as 0, 1 or 2) for the animal at each SNP location. However, the rate of convergence may vary between animals, meaning that a separate plot of the gEBV would need to be produced for each animal included in the analysis. However, this is not practical as most analyses require data from hundreds or thousands of animals. Also, if the gEBV behaves in the same way as the SAE or SSE, the estimated SNP effects may continue changing, while the net change in gEBV may be negligible. This may indicate convergence prematurely and subsequently lead to poor posterior inferences.

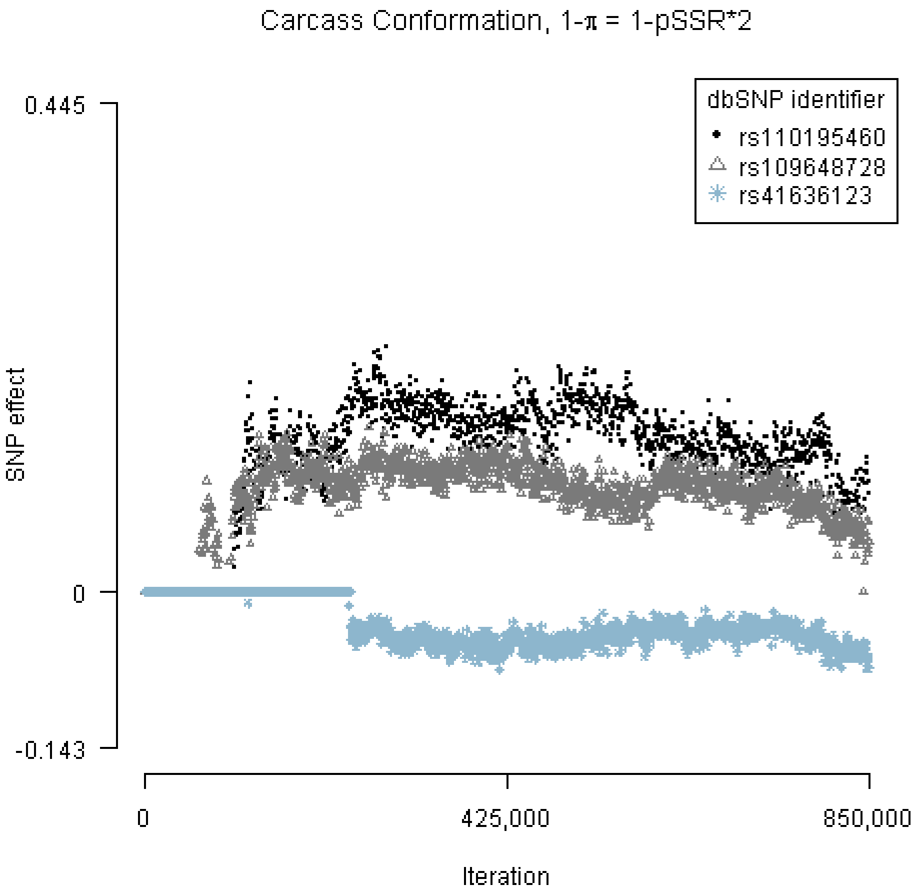


Figure 4.12 Example of 3 SNPs which had estimated non-zero effect much later than the SSE or SAE. From the analysis for carcass conformation (1-π = pJ = 2.27 × 10^{-2}), three SNPs; rs110195460, rs109648728 and rs41636123 had an estimated effect of zero until iteration 105,000, 63,500 and 121,500 respectively, but had a non-zero effect in all sampled iterations after iteration 240,000.

4.4.4 The number of SNPs with a PP > 0.5

Convergence diagnostics relate to the idea that a chain after a sufficient amount of time will eventually reach a stationary phase in the parameter space and continue to mix around the mode of that stationary phase. This was the case for the number of SNPs that had a PP>0.5. The mean and mode of the number of SNPs with a PP>0.5 for all analyses were quite similar for the last 100,000 iterations (Table 4.3), indicating that the samples from the last 100,000 iterations were taken from around the mode of the same distribution and that convergence had been achieved. In fact, the mean and the mode for SNPs with a PP>0.5 for the last 100,000 iterations was exactly the same for a number of the analyses e.g. CWT, 1-π = pA and CFAT, 1-π = pF. As expected, the number of SNPs with a PP>0.5 was greatest at early iterations for all analyses but decreased rapidly so that by approximately iteration 200,000 all models had reached a stationary phase. This was particularly evident for each of the analyses with a prior value of 1-π = pK. Additionally, similar to the model log-likelihood plots, the burn-in phase indicated from this metric varied between analyses. This meant that the burn-in for each analysis would need to be assessed for each analysis separately, and could not be chosen *a priori*.

At early iterations, the impact of prior specifications of (1-π) on estimates of this metric was particularly evident from inspection of the surface plot for each trait. Although the number of SNPs with a PP>0.5 was much closer at the final iteration for all analyses of a trait, these values still varied. Additionally, for the analysis of CULL (1-π = pK), the mean and the mode number of SNPs with a PP>0.5 was zero for the final 100,000 iterations (sampled every 500th), however the SAE and SSE estimates were both non-zero at all sampled iterations. This suggests, that the model is still estimating SNPs to have both zero and non-zero effects (switching between iterations), and has not converged to a stationary phase.

4.4.5 Coefficient of determination

The coefficient of determination, R^2, is a well known metric for describing the relationship between two variables. In this case the estimated SNP effects from the Bayesian and the SSR approaches (for SNPs with PP>0.5) were compared using the R^2 at each sampled iteration. The R^2 between SNP effects estimated from both the Bayesian approach and the SSR approach, allowed us to determine how well the estimates from the SSR approach were described by the estimates from the Bayesian approach. As the number of SNPs with a PP>0.5 approached a stationary distribution, a constant R^2 suggested that the estimated SNP effects from the Bayesian model had also converged.

Although estimates did not need to be close to 1 in order to convergence, analyses that had estimates close to one and converged were particularly interesting as they seemed to be estimating values closest to the SSR approach. The R^2 did converge for all models, but both the number of iterations required to reach convergence and the estimated R^2 at convergence varied between analyses of a trait. Also, it must be noted that the R^2 between two datasets may be high when the mean and standard deviation between these datasets is greatly different. This means that although the R^2 may be high, the estimated SNP effects from both models may be quite different. Additionally, at some iterations, the number of SNPs with a PP>0.5 may be zero (or just one) meaning that the R^2 cannot be estimated. Similarly, when the number of SNPs with a

PP>0.5 is low, the R^2 may be high due to the lack of data included in the estimate. To address these limitations, a Euclidean distance (ED) was also calculated for the same set of SNPs.

4.4.6 The Euclidean distance

As mentioned, two data sets can have R^2 close to 1 but be very different. This is not the case with the ED. The ED is a common metric that reflects the proximity of 2 points in a n-dimensional space, such that the distance between *x* and *y* is equal to the distance between *y* and *x* and the distance between *x* and *y* is equal to zero when $x = y$. The ED was used here to gauge the similarity between the estimated SNP effects (of SNPs with a PP>0.5) of the Bayesian and the SSR approaches. For all analyses, the ED was greatest at early iterations. Although convergence was achieved in most analyses, the number of iterations needed to reach convergence was different for each analysis.

The surface plot of this metric allowed us to investigate both the convergence behaviour of each prior for a trait and the similarity of estimates between prior specifications. However, as with the other metrics used, the estimated ED varied between different prior specifications of (1-π). This was particularly evident in the CFAT and CONF surface plots (Appendix 2).

4.4.7 Running additional chains

There is considerable debate in the MCMC community over whether or not it is better to base convergence assessment and inferences on two or more replicate chains, beginning from alternate starting points (seeds), or one long chain (Geyer, 1992; Huelsenbeck et al., 2002; Beiko et al., 2006). One disadvantage of replicate chains is that each chain has its own burn-in phase, which must be removed. When convergence is slow, a large amount of computational effort is spent obtaining samples that must then be removed (Beiko et al., 2006). Additionally, the resources needed to run multiple chains in parallel are not always readily available. Although a strong argument for the single long chain approach was made by Geyer (1992), it is generally accepted that when sufficient computational resources are available, it is advisable to run multiple chains concurrently (Gelman and Rubin, 1992; Beiko et al., 2006; Lee et al., 2008). In spite of this, the use of multiple chains running in parallel is not commonly implemented in GWAS or genetic prediction studies that utilise a model from the Bayesian alphabet (e.g. Purfield et al., 2013). For this reason, we focused on the single longer chain approach, although the metrics put forward in this study could be used to sample across chains to assess convergence. Also, these metrics if used to sample across chains could easily be plotted to provide a graphical assessment of convergence.

4.4.8 Recommendations for future studies

For each analysis, the number of iterations required to reach convergence differed for each analysis. This meant that a burn-in phase could not be chosen *a priori*. This was not unexpected however, as the number of iterations needed to reach convergence will be influenced by many factors including, prior assumptions, starting values (seeds), the number of samples and the number of SNPs. Including estimates from the burn-in phase may affect posterior estimates (see Table 4.2 and Table 4.3). In general, by iteration 100,000 each analysis had converged, although this was not always the case reflecting the difficulty in choosing a burn-in *a* priori. Because of this, convergence of each model should be assessed separately and all iterations before convergence discarded.

We recommend using primarily the summed absolute log-likelihood metric to assess convergence, as this metric incorporates information from all parameters and not just those of interest. However, as pointed out by Cowles and Carlin (1996), there is no one conclusive diagnostic that can provide assurance of convergence. Convergence of a number of parameters, and not just those of interest should be checked before making any inferences. For this purpose, the number of SNPs with a PP>0.5 and the Euclidean distance metrics are good secondary diagnostics.

One of the limitations of a Bayesian approach is potential bias created by including information *a priori*. This could be seen in the surface plots for many of the analyses. Potentially, the impact of prior specifications in a Bayesian model may be reduced by combining estimates from multiple prior (1-π) specifications (Knurr et al., 2013). Then, SNPs that have a PP>0.5 in the majority of analyses could be identified.

Chapter 5: Transcriptional response to reduced energy intake and subsequent compensatory growth of *M. longissimus thoracis et lumborum* in Aberdeen Angus steers

5.1 Introduction

Many growing organisms naturally undergo periods of growth depression in their lifetime, usually due to reduced feed intake, but possess the potential to rapidly recover from such periods when favourable conditions, such as the availability of a higher energy diet, occur (Hornick et al., 2000; Tolla et al., 2003; Jobling, 2010). This phenomenon, commonly referred to as compensatory growth, allows organisms to achieve a genetically pre-determined inherent size following periods of restricted energy intake (Connor et al., 2010). Compensatory growth is a common feature among many species including fish (Ali et al., 2003), cattle (Lehnert et al., 2006) and even humans (Ashworth, 1969). In spite of this, relatively little is known about the genes or biological processes underlying the transcriptional responses regulating compensatory growth in any species.

In beef cattle production systems, animal feed costs represent the single largest variable cost in achieving a marketable steer (Connor et al., 2010; Finneran et al., 2010; Mao et al., 2013). Demand for animal feed is greatest during Winter, and consequently costs are highest during this period. Conversely, feed costs are lowest in Spring and early Summer due to the availability of pasture for grazing. Because of this, there is growing interest in utilising compensatory growth in management strategies and understanding the biological mechanisms involved in the compensatory growth response.

The exploitation of compensatory growth is common practice in beef production systems worldwide. Particularly in pastoral systems, such as in Ireland, there is an opportunity to reduce overall feed costs by rebalancing feed demand away from times when feed is expensive towards times when feed is cheap. A compensatory growth regimen can be incorporated into a production system by reducing the availability of feed for an animal during Winter when it is most expensive, until Spring and early Summer when pasture is cheap and abundant (Fiems et al., 2007). Thus, this strategy could be used to offset high production costs over Winter, while maintaining overall production targets. Furthermore, management strategies that elicit a compensatory growth response present an opportunity to investigate and elucidate the underlying mechanisms involved in restricted growth and subsequent compensatory growth (Picha et al., 2008).

Compensatory growth is recognised by a significantly faster growth rate in animals that have previously experienced growth depression compared to the growth rate of control animals that have not experienced growth depression. It is also accompanied with increased efficiency of energy use thus allowing the animal to quickly recover body mass (Mangel and Munch, 2005). The ability to compensate is influenced by several factors such as the duration of restriction, the severity of restriction and the quality of diet during re-alimentation. Although muscle tissue is influenced considerably by starvation and re-feeding (Hornick et al., 2000; Johansen and Overturf, 2006), elucidating the mechanisms underlying compensatory growth has proven difficult.

The transcriptional responses involved in the compensatory growth response have been examined only on a limited basis (Connor et al., 2010). Many studies have focused on a pre-selected list of candidate genes (e.g. Johansen and Overturf, 2006; Picha et al., 2008), or relied on microarrays (Lehnert et al., 2006) which have a smaller range of fold change detection than next-generation sequencing approaches (Xu et al., 2013). Next-generation sequencing approaches are not limited by the pre-selection of candidate genes, and are more sensitive to the detection of genes compared to microarray analyses (Wang et al., 2009b; Foley et al., 2012; McCabe et al., 2012).

Bovine skeletal muscle is a tissue of major economic importance to the global economy (Keady et al., 2011). In particular, *M. longissimus et lumborum* is of high commercial value in beef production. As such, the objective of this study was to investigate the transcriptional activity regulating the compensatory growth response in *M. longissimus et lumborum* using RNA-seq technology. Significantly differentially expressed genes in animals under going nutritional restriction, and subsequent compensatory growth, compared to a control group, were identified. This information was then used to identify significantly over-represented pathways. This is the first study to assess bovine muscle tissue responses to compensatory growth using next-generation sequencing technology.

5.2 Material and Methods

The initial stage of this study was performed by Dr. Sarah Keady. This included the management of all animals, sample collection, total RNA extraction and preparation of mRNA-seq libraries for sequencing. These procedures are explained in brief in the following sections. For a more detailed description of the procedures see Appendix 4.

5.2.1 Animal ethics statement

All animal procedures were carried out according to the provisions of the Irish Cruelty to Animals Act (licenses issued by the Department of Health and Children) and the European Communities Regulation 2002 and 2005.

5.2.2 Animal model

The animal model consisted of two periods representing Winter (day 1 to 99) and Spring (day 100 to 131). Twelve crossbred Aberdeen Angus × Holstein-Friesian steers were assigned to one of two feeding treatments for the entire study period. The average age of the animals at the commencement of the study was 362 days (standard deviation of 15.5 days). These two treatments replicated feeding conditions experienced by animals in a management system that utilises compensatory growth (experimental group) and a management system that does not i.e. where animal have *ad libitum* access to feed (control group).

5.2.2.1 Differential feeding period (Winter)

During the first period, known as the differential feeding period, animals were assigned to one of two different feeding treatments; a control diet (n = 6) or an energy restricted diet (n = 6). The control diet consisted of high energy concentrates *ad libitum* (Dry matter (DM) 825 g/kg, *in vitro* DM digestibility (DMD) 862 g/kg, crude protein (CP) 120.9 g/kg, ash 43 g/kg, neutral detergent fibre (NDF) 557 g/kg and acid detergent fibre (ADF) 351 g/kg) and 7kg of grass silage (low energy) per animal daily (DM 228 g/kg, *in vitro* DMD 677 g/kg, CP 112 g/kg, ash 80 g/kg, NDF 557 g/kg, ADF 351 g/kg and pH 3.6). The energy restricted diet consisted of 0.5 kg of the same concentrate feed as that offered to the control group plus *ad libitum* access to grass silage. These feeding regimes replicated the same feeding conditions that animals would be subjected to in a either a traditional management system or a system that typically exploited compensatory growth. Muscle biopsies for each animal were collected from the *M. longissimus thoracis et lumborum* (between the 12th and 13th ribs) at the end of the differential feeding period on day 99 of the study.

5.2.2.2 Re-alimentation period (Spring)

The second period, known as the re-alimentation period, consisted of offering both groups of animals *ad libitum* access to the same total mixed ration diet composed of 80% concentrates and 20% grass silage. Muscle biopsies from the *M. longissimus thoracis et lumborum* (between the 12th and 13th ribs) were taken from each animal on day 32 (day 131 of the experiment) of the re-alimentation period when the compensatory growth response is strongest as indicated by the live weight gain (Appendix 5).

5.2.3 Sample collection and total RNA extraction

Sample collection and total RNA extraction was carried out by Dr. Sarah Keady at the Teagasc Animal & Grassland Research and Innovation Centre in Grange. The procedures are described in brief in the following sections. Appendix 4 contains a detailed description of all procedures.

Briefly; a muscle biopsy sample was taken from each animal using a trochar and cannula instrument. Biopsies were washed in Dulbecco's phosphate buffered saline (DPBS), snap frozen in liquid nitrogen and stored at -80 ^{o}C.

RNA was extracted from homogenised muscle tissues using 200 µl of chloroform and a single round of centrifugation (12,000 g for 15 minutes at 4^{o}C). After centrifugation, the upper aqueous layer was transferred to a 1.5 ml tube. From this isopropanol was added and the sample briefly vortexed followed by centrifugation (12,000 g for 10 minutes at 4^{o}C). After the supernatant was removed, ethanol was added, and the sample vortexed and then centrifuged (7,500 g for 5 minutes at 4^{o}C). After this, the supernatant was removed, and the remaining RNA pellet air-dried. Nuclease-free water (20 µl) was added to the tube and the pellet dissolved completely by gently pipetting. RNA quantity was assessed using the Nanodrop spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE, USA). The Agilent Bioanalyser 2100 (Agilent Technologies) was used to assess the quality of the RNA. Samples with an RNA integrity number (RIN) equal to or greater than 8 were deemed acceptable and used as template for the preparation of mRNA-seq libraries.

5.2.4 mRNA-seq library preparation and sequencing

PolyA RNA was isolated from the extracted mRNA as described in (McCabe et al., 2012). In brief, 5-10 ug total RNA was processed with oligo (dT) beads using two rounds of oligo-dT purification. 5-10 ug RNA was fragmented with zinc fragmentase (Applied Biosystems, Warrington, UK), first strand cDNA synthesis was performed using the Invitrogen random hexamer primers and SuperScript II (Invitrogen), and second strand synthesis was performed using Invitrogen DNA Polymerase 1 (Invitrogen). End repair and polyadenylation were performed using NEB Next Tailing Module (New England Biolabs). ® End Repair Module and NEB Next dA- Illumina single read adapters were ligated to blunt ended, polyadenylated fragments with a NEB Quick ligation kit (New England Biolabs). Adapter-ligated cDNA fragment libraries were sequenced on an Illumina Genome Analyzer II (GAII). Sequencing was performed on the Illumina GAII at the Conway Institute (University College Dublin, Ireland) using 40bp paired-end version 4 cluster preparation kits according to manufacturers' instructions (Illumina, United Kingdom).

In total, twenty-four cDNA paired-end read libraries were prepared from total RNA which was extracted from muscle tissue biopsies at two separate time points and sequenced as 40 bp paired-end reads on 28 lanes randomly distributed across 4 flowcells on an Illumina GAII. These libraries represented 12 animals (6 control and 6 experimental) at 2 different time points (end of differential feeding and during the re-alimentation period).

5.2.5 Raw read quality control

An important first step in the analysis of RNA sequences is pre-processing and quality control. In order to generate better mapping results, FASTQ files should be checked for quality before aligning. Additionally, as per-base sequence quality usually decreases towards the end of the read, sequences with low quality scores should be trimmed (Murigneux et al., 2013). The quality of raw reads was determined using the FastQC program (version 0.10.1) (FastQC). FastQC is an easy to use tool, which provides an overview of the quality of raw sequence data. FastQC provides several summary graphs and tables which can be used to determine if there are any problems with the data. The assessments included per base sequence quality, per sequence quality, over-represented sequences and per base N content. Raw reads were then quality trimmed using the fastq-mcf command in the ea-utils package (Aronesty, 2013). Bases with a phred score of less than 28 (99.5% accuracy in the base call) were trimmed. Reads that were less than 15 bps after trimming were removed from further analysis.

5.2.6 Alignment to the Bovine genome

Reads from each of the lanes were aligned separately to the UMD3.1 assembly of the bovine genome using Tophat (version 2.0.6) (Kim et al., 2013). Tophat aligns reads using the high-throughput, ultrafast short read aligner Bowtie (version 2.0.5) (Langmead and Salzberg, 2012). The following options were specified; quality scores are ASCII characters equal to the Phred quality scores plus 64 (−−Solexa1.3-quals), the very sensitive setting for Bowtie (--b2-very-sensitive) and library type is paired-end data (--library-type fr-unstranded). Resulting alignment files were reported in SAM format. SAM files were sorted using Samtools (version 0.1.18) (Li et al., 2009) and filtered for possible PCR and optical duplicates using Picard tools (version 1.60) (The Picard toolkit). This output was then sorted by read name, instead of the default chromosomal location, for use with HTSeq (version 0.5.3) (HTSeq).

5.2.7 Read Counts

The script HTSeq-count provided with the HTSeq program (HTSeq) was used to count the number of reads that mapped to each annotated gene using the union mode (-m union) for reads partially over-lapping exons, and the Ensembl 69 annotation of the bovine genome (Flicek et al., 2012). As our sequencing analysis included technical replicates of individual samples, we summed the counts for all of these lanes, resulting in a single count for each gene for each sample. Counts for all samples were collated into a single file that contained the read counts for all genes in all samples. Genes with fewer than 5 uniquely aligned reads in all samples were excluded from further analysis.

5.2.8 Differentially expressed genes

The bioconductor package DESeq (version 1.12) (Anders and Huber, 2010) was used within R (version 2.15.2) (The R Project) to identify significantly differentially expressed genes in the experimental group versus the control group. Two lists of significantly differentially expressed (SDE) genes were returned from DESeq, corresponding to each time point. DESeq models the count data using a generalisation of the Poisson distribution, the negative binomial distribution. This is used to account for biological and technical variance and test for differential expression between two experimental conditions. As there are over 20,000 genes annotated to the bovine genome, the Benjamini and Hochberg (BH) (Benjamini and Hochberg, 1995) method, implemented in R (The R Project), was used to correct for errors arising from many thousands of individual tests. Genes with a BH corrected p-value <0.1 were considered differentially expressed and retained for further analysis. This meant that two datasets of significantly differentially expressed genes remained, each corresponding to an individual time point, i.e. at the end of the differential feeding period and during the compensatory response.

5.2.9 Over-represented KEGG pathways

Significantly differentially expressed bovine genes (BH corrected p-value <0.1) were mapped to their human orthologs using the mapping available from hg19 of the human genome (Ensembl 69). The bioconductor package GOseq (Young et al., 2010) was then used to identify significantly over-represented KEGG (Kanehisa et al., 2012) pathways (p <0.05). In RNA-seq analyses, longer or more highly expressed genes are more likely to be detected as differentially expressed than shorter and/or lowly expressed genes (Oshlack and Wakefield, 2009; Young et al., 2010). When investigating a gene list for biological function or pathway over-representation, GOseq includes a correction for the higher number of reads that will be assigned to long or highly expressed genes. GOseq estimates a probability weighting function from the data, which is used to quantify the probability that gene will be differentially expressed based solely on its length. In GOseq, the default method for calculating pathway over-representation p-values is an extension of the hypergeometric distribution, the Wallenius non-central hypergeometric distribution. This was used to identify significantly over-represented KEGG pathways (p <0.05).

5.3 Results

5.3.1 mRNA-sequencing and read alignment

The average number of raw reads generated across all sequencing runs was 33.43 million. Following quality trimming, an average of 32.19 million reads remained per lane, and the average read length was 39.1 bps. Of these, the average number of reads that aligned exactly once was 16.27 million reads. This data has been deposited in NCBI's Gene Expression Omnibus (GEO) (Barrett et al., 2013) and are accessible through GEO Series accession number GSE48481. Filenames for the files contained in GEO under this accession number, and the corresponding alignment statistics are contained in Appendix 6.

5.3.2 Differentially expressed genes

At both time points, the lowest limit of detection was set to 5 or more uniquely aligned reads in at least one animal in either the control or the experimental group. At this limit, 14,257 and 13,719 genes were detected as expressed at time point 1 and time point 2 respectively.

5.3.2.1 Differential feeding period

On day 99 of the study, 440 genes were found differentially expressed in muscle tissue of cattle of the experimental group versus the control group (Electronic Appendix 5.1). The most significantly differentially expressed genes were related to metabolism; acyl-coenzyme A synthetase medium-chain family member 1 (ACSM1) (log 2 fold change = -3.58, p = 1.92×10^{-56}) and fatty acid binding protein 4 (FABP4) (log 2 fold change = -2.99, p = 1.55×10^{-38}). In addition, leptin precursor (OB) gene, the adipocyte differentiation gene adipogenin (ADIG) and insulin-induced gene 1 (INSIG1) were all significantly down-regulated. Notably, glucagon receptor precursor (GCGR) was significantly up-regulated in the experimental group, indicating increased mobilisation of stored energy.

5.3.2.2 Re-alimentation period

At the second time point (day 32 of re-alimentation), 163 genes were significantly differentially expressed in the experimental group compared to the control group (Electronic Appendix 5.2). The most significantly differentially expressed gene was myosin heavy chain 4 (MYH4) ($p = 2.99 \times 10^{-39}$), which was up-regulated. Growth hormone releasing hormone (also known as somatoliberin) (GHRH) (log 2 fold change = 5.37, $p = 8.06\times 10^{-17}$) was also up-regulated. In total, 67 genes were significantly up-regulated at this time point.

5.3.3 Over-represented KEGG pathways

5.3.3.1 Differential feeding period

Of the 440 genes that were differentially expressed at the end of the differential feeding period, 406 mapped to 429 human orthologs. Using this gene list, 17 KEGG pathways were significantly over-represented (Table 5.1). The most significantly over-represented KEGG pathway was staphylococcus aureus infection. Other significantly over-represented pathways included Glycolysis/Gluconeogenesis, a number of pathways related to metabolism (e.g. Arachidonic acid metabolism) and the peroxisome proliferator-activated (PPAR) signalling pathway.

Table 5.1 Significantly over-represented KEGG pathways at the end of the differential feeding period. Using the significantly differentially expressed genes at time point 1 as input, 17 KEGG pathways were found significantly over-represented ($p < 0.05$).

KEGG Pathway Name	p-value
Staphylococcus aureus infection	8.25×10^{-5}
Proteasome	2.83×10^{-4}
Protein processing in endoplasmic reticulum	3.39×10^{-4}
Systemic lupus erythematosus	2.49×10^{-3}
Thiamine metabolism	2.67×10^{-3}
PPAR signalling pathway	3.76×10^{-3}
Metabolic pathways	4.51×10^{-3}
Arginine and proline metabolism	7.15×10^{-3}
Prion diseases	8.65×10^{-3}
Amoebiasis	1.63×10^{-2}
Complement and coagulation cascades	1.76×10^{-2}
Ribosome	1.91×10^{-2}
Citrate cycle (TCA cycle)	2.88×10^{-2}
Arachidonic acid metabolism	3.03×10^{-2}
ECM-receptor interaction	3.45×10^{-2}
Biotin metabolism	3.89×10^{-2}
Glycolysis / Gluconeogenesis	4.07×10^{-2}

5.3.3.2 Re-alimentation period

Of the 163 significantly differentially expressed genes at time point 2, 155 mapped to 171 unique human orthologs. From this, 9 KEGG pathways were found to be significantly over-represented (Table 5.2). Similar to time point 1, a number of pathways related to metabolism were found significantly over-represented (e.g. “Carbohydrate digestion and absorption” and “metabolic pathways”). In fact, the most significantly over-represented pathway was “Starch and sucrose metabolism” ($p = 1.87 \times 10^{-6}$). Furthermore, pathways involved in regulating muscle accretion were also significantly over-represented (e.g. cell cycle). Also, the transforming growth factor-β (TGF-β) signalling pathway was also significantly over-represented during this period when animals, and indeed muscle tissue, are growing at a much faster rate than control animals.

Table 5.2 Significantly over-represented KEGG pathways during re-alimentation. Using the significantly differentially expressed genes at time point 2 as input, 9 KEGG pathways were found significantly over-represented ($p < 0.05$).

KEGG Pathway Name	**p-value**
Starch and sucrose metabolism	1.87×10^{-6}
Carbohydrate digestion and absorption	5.03×10^{-5}
Valine, leucine and isoleucine degradation	7.23×10^{-3}
Osteoclast differentiation	2.41×10^{-2}
Lysine biosynthesis	2.75×10^{-2}
Cell cycle	2.94×10^{-2}
Metabolic pathways	4.00×10^{-2}
Salivary secretion	4.07×10^{-2}
TGF-beta signalling pathway	4.55×10^{-2}

5.4 Discussion

The current study investigated the transcriptional responses in bovine muscle tissue to periods of reduced nutritional intake and subsequent compensatory growth following re-alimentation. Although some research on this phenomenon has been done previously (e.g. Park, 2005; Lametsch et al., 2006), many of those studies focused on candidate genes subsets (Luna-Pinto and Cronje, 2000; Picha et al., 2008). As such many of the genes and biological processes involved in compensatory growth have yet to be elucidated. Furthermore, as far as we are aware, this is the first study to use next-generation sequencing to assess transcriptional responses to compensatory growth conditions in bovine skeletal muscle. In this study, next-generation RNA sequencing of muscle tissue was used to identify significantly differentially expressed genes at two time points; at the end of a nutritional restriction period, and 32 days into the re-alimentation period. Significantly differentially expressed genes were then used to identify significantly over-represented KEGG pathways at each time point separately in an attempt to elucidate the molecular mechanisms regulating compensatory growth in muscle of cattle.

5.4.1 Compensatory growth model

Compensatory growth is influenced by many factors, which include the duration and severity of restricted feeding. In this study, the duration and severity of the restricted diet was chosen to represent the incorporation of compensatory growth in an on farm management strategy. This meant that the duration of restriction was approximately 3 months, representing Winter, and the severity of restriction did not lead to long-term adverse or potentially harmful effects on the animals.

In a previous study investigating compensatory growth responses in Belmont Red steers, gene expression was examined much later in the re-alimentation period at day 84 (Lehnert et al., 2006). In that study, authors reported only a single gene (myosin regulatory light chain) as significantly differentially expressed during the re-alimentation period. At 84 days, the experimental animals had entered normal growth trajectories meaning that the genes controlling compensatory growth response were not differentially expressed in the experimental group versus the control group i.e. the animals which were previously restricted were responding similar to those fed the *ad libitum* diet throughout the experiment. The timing of the second biopsy in our study, i.e. 32 days into re-alimentation, marked the initial phase of the compensatory growth response which is accompanied by accelerated muscle growth, when the response would be strongest. This facilitated detection of acute responses in muscle tissue to improved nutrition following a prolonged period of feed restriction.

5.4.2 Differential feeding period

At the end of the differential feeding period a total of 440 genes were found to be significantly differentially expressed. The majority of these genes (248) were significantly down-regulated. Additionally, a large number of KEGG pathways were significantly over-represented. Many of the genes and pathways identified as significant during the differential feeding period are known to be involved in processes related to metabolism (e.g. leptin gene (OB), fatty acid binding protein 4 (FABP4), glucagon receptor (GCGR), peroxisome proliferator-activated receptor (PPAR) signalling and metabolic pathways) and adipocyte differentiation (e.g. adipogenin (ADIG), Glycolysis / Gluconeogenesis and PPAR).

5.4.2.1 Leptin gene

Leptin (produced by the OB gene) was significantly down-regulated ($p = 2.57 \times 10^{-7}$) at the end of the end of the differential feeding period. Leptin plays an important role in regulating adipose tissue mass and lipid metabolism (Sahu, 2004; Sainz et al., 2009). Leptin has also been shown to be involved in the regulation of skeletal muscle growth (Trostler et al., 1979; Ceddia et al., 2001; Sainz et al., 2009). Interestingly, leptin deficiency has shown to be linked to decreased muscle mass in mice (Trostler et al., 1979). Down-regulation of leptin in animals experiencing extended periods of reduced nutritional intake may be regulating depressed muscle growth rate. This is not unexpected given the reduced emphasis on muscle growth during the differential feeding period.

5.4.2.2 Staphylococcus aureus infection

The KEGG pathway Staphylococcus aureus infection was the most significantly over-represented pathway at the end of the differential feeding period. Six genes were significantly differentially expressed in this pathway, 5 of which were significantly down-regulated (complement component 1, q subcomponent, B chain (C1QB), complement component 1, q subcomponent, C chain (C1QC), immunoglobulin gamma Fc receptor II-a (FCGR2A), immunoglobulin gamma Fc receptor II-b (FCGR2B) and platelet-activating factor receptor (PTAFR)). Only manna-binding lectin serine peptidase 1 (MASP1) was significantly differentially up-regulated. Furthermore, 4 of these genes are also involved in Systemic lupus erythematosus, which was also significantly over-represented ($p = 2.49 \times 10^{-3}$). Although KEGG is particularly well annotated for, and thus biased toward, human immune function, this was an unexpected result. Interestingly, similar results have been reported previously. Connor et al. (2010) reported changes to genes regulating immune function and inflammation in black Angus steers following a 10 week period of feed restriction. In addition to this, genes involved in immune response have been shown to be down-regulated following periods of starvation in mice (Dhahbi et al., 2004), rainbow trout (Salem et al., 2007) and even Drosophila melanogaster (Fujikawa et al., 2009). This may indicate reduced immune response capacity due to partitioning of energy away from immune function in animals experiencing long periods of restricted energy intake.

5.4.2.3 Peroxisome proliferator-activated receptor signalling

Another interesting significantly over-represented pathway was the peroxisome proliferator-activated receptor (PPAR) signalling pathway. The PPAR signalling pathway plays an important role in numerous functions related to lipogenesis, adipocyte differentiation and lipid metabolism (Canovas et al., 2010). Six genes in the PPAR signalling pathway were significantly differentially expressed at the end of the differential feeding period, five of which were significantly down-regulated (Figure 5.1). The most significantly differentially expressed gene in the PPAR signalling pathway was fatty acid binding protein 4 (FABP4) ($p = 1.55 \times 10^{-38}$). FABP4 was significantly down-regulated at the end of the differential feeding period. Although muscle biopsies were taken from each animal, FABP4 is primarily expressed in adipocyte tissue (Hertzel and Bernlohr, 2000). Aberdeen Angus cattle contain a high level of intramuscular fat (Keady et al., 2013). It is possible that some of the intramuscular fat contained in skeletal muscle was also taken with the biopsy. This would explain the presence of FABP4. Fatty acid binding proteins are part of a large gene family comprising proteins mediating intracellular uptake and transport of long-chain fatty acids within the cell (Hertzel and Bernlohr, 2000). Fatty acid binding proteins also play an important role in systemic energy homeostasis (Storch and McDermott, 2009). In rainbow trout FABP has also been shown to be down-regulated following periods of starvation (Salem et al., 2007). Additionally, FABP inhibition in chicken adipocytes has been linked with decreased lipid accumulation (Shi et al., 2011). Down-regulation of FABP4 in the current study is not surprising as, following a long period of restricted feeding, energy availability is dramatically reduced thus inhibiting the ability to accumulate body fat.

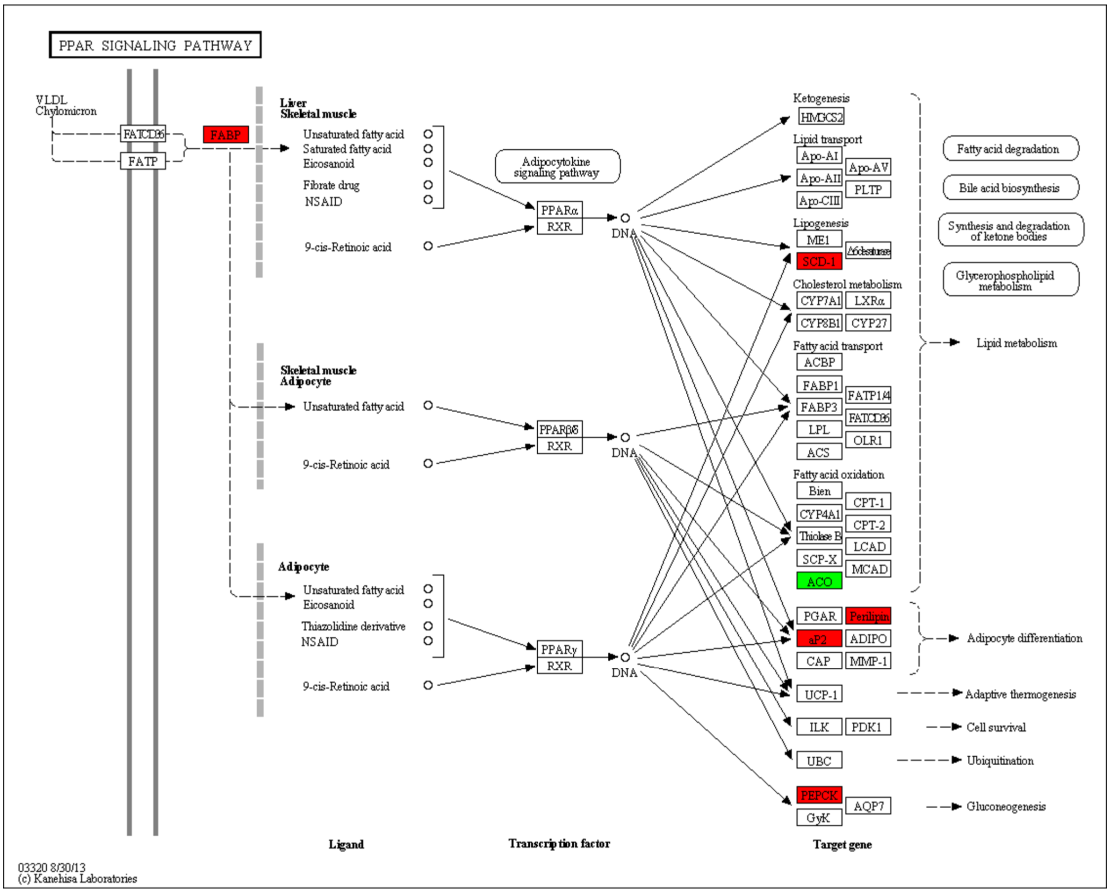


Figure 5.1 Significantly differentially expressed genes in the PPAR signalling pathway at time point 1. Significantly up-regulated genes are shown in green, and significantly down-regulated genes are shown in red.

5.4.3 Re-alimentation period

At the second time-point 163 genes were significantly differentially expressed in the experimental group compared to the control group. From this dataset, 9 KEGG pathways were significantly over-represented. Many of these pathways are related to metabolism and muscle differentiation. This included starch and sucrose metabolism, metabolic pathways and TGF-β signalling. Many of the genes found differentially expressed at this time-point indicated the activation of processes related to muscle growth such as myosin heavy chain 4 (MYH4), inhibin-β subunit A (INHBA), Poly(A) binding protein interacting protein 2B (PAIP2B) and forkhead box M1 (FOXM1).

5.4.3.1 Forkhead box M1 gene

Forkhead box (FOX) proteins are a family of cell growth and proliferation associated transcription factors. FOXM1, which was up-regulated during re-alimentation, stimulates cell proliferation and promotes cell cycle progression (Wierstra and Alves, 2007). FOXM1 levels have been shown to increase at the onset of the S-phase and persist until the end of mitosis (Laoukili et al., 2005). FOXM1 regulates genes that control G1/S-transition, S-phase progression, G2/M-transition and M-phase progression. Interestingly, cell division cycle associated genes 3 (CDCA3) and 8 (CDCA8) were both up-regulated. In the integrated transcription factor platform (ITFP) database (Zheng et al., 2008), both CDCA3 and CDCA8 are reported to be regulated by FOXM1. Additionally, aurora B kinase (AURKB) was also significantly up-regulated at time-point 2. This gene has been shown to interact with CDCA8 (Gassmann et al., 2004). This data suggests that during the compensatory growth response, the expression of genes regulating cell cycle is increased to meet the demands of accelerate muscle growth.

5.4.3.2 The TGF-β signalling pathway

Interestingly, the TGF-β signalling pathway was significantly over-represented during the re-alimentation period of the study. This is not surprising, given that TGF-β signalling is well known to be involved in regulating skeletal muscle growth. In particular, myostatin, a negative regulator of muscle growth has been well studied (Lee and McPherron, 2001; Lee, 2004; Kollias and McDermott, 2008). Myostatin acts on muscle tissue by binding to activin type II receptor (ACVRII) (Burks and Cohn, 2011). Repressing myostatin action has recently garnered interest as a potential therapeutic application in treatment of muscle atrophy (Roth and Walsh, 2004; Zhou et al., 2010; Han et al., 2013). Although, myostatin signalling was not significantly differentially expressed in this study, inhibin-β subunit A (INHBA) was significantly up-regulated (Figure 5.2). Inhibins, act as activin antagonists by forming high affinity complexes with ACTRII and betaglycan (Lewis et al., 2000; Tsuchida et al., 2009). It is hypothesised that the effects of the TGF-β signalling pathway and myostatin are reduced by increased INHBA activation, thereby promoting accelerated cell growth and proliferation in muscle tissue of animals experiencing compensatory growth.

5.4.4 Potential targets for muscle growth regulation

Increased understanding of the mechanisms controlling muscle mass has attracted growing attention in animal and medical science. Compensatory growth is an interesting model to study the effects on starvation on muscle growth and subsequently the acute responses in muscle to nutrient availability. Significantly differentially expressed genes from this study represent a candidate gene list warranting further investigation potentially yielding markers that could be included in selective breeding programmes in cattle. Also, increased understanding of the transcriptional regulation of muscle growth can facilitate improved treatments of various clinical problems (Bonaldo and Sandri, 2013).

In particular the role of INHBA in regulating myostatin warrants further investigation.

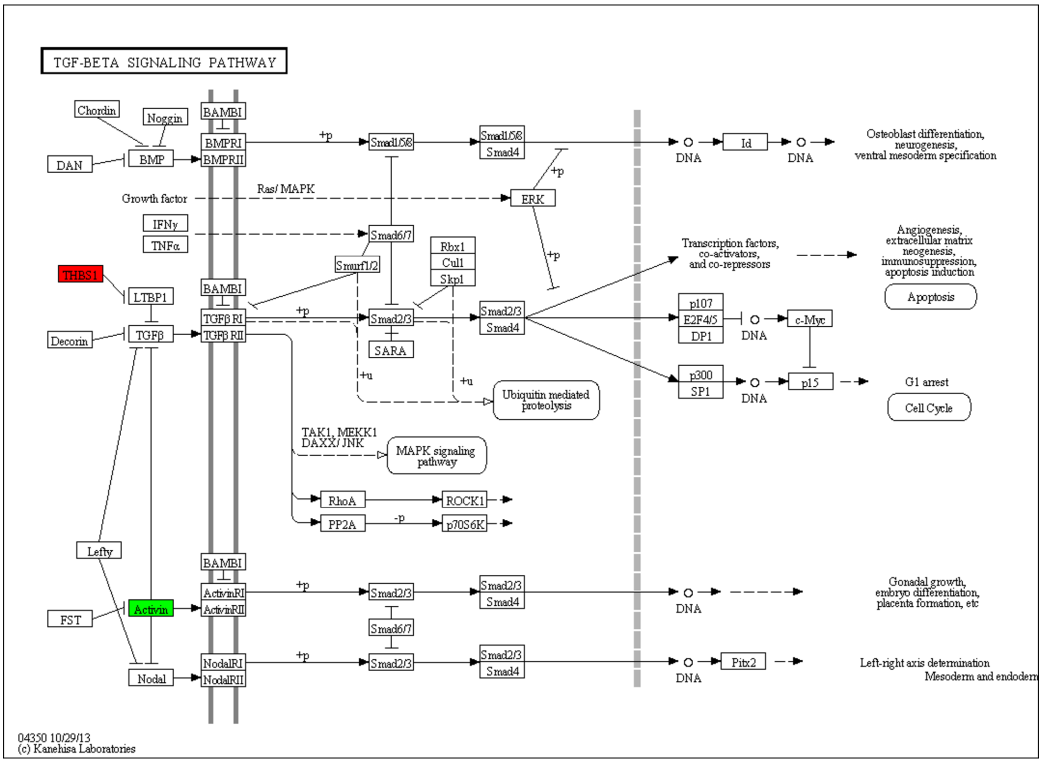


Figure 5.2 Significantly differentially expressed genes in the TGF-β signalling pathway at time point 2. Significantly down-regulated genes are coloured in red and significantly up-regulated genes are in bright green.

Chapter 6: Divergent evolution of somatotropic axis genes in dairy and beef (*Bos taurus*) cattle

6.1 Introduction

Cattle were first domesticated between 8,000 and 10,000 years ago, giving rise to two lineages observed today; *Bos taurus* (taurine) and *Bos indicus* (indicine or zebu) (Willham, 1986; Vigne, 2011). Modern domestic cattle originate from independent domestications of the same, now extinct, ancestral species; the auroch (*Bos primigenius*) (McTavish et al., 2013; Ramey et al., 2013). The taurine lineage was probably first domesticated in the Near East, whereas domestication of aurochsen on the Indian subcontinent gave rise to the indicine lineage (Vigne, 2011). Several phenotypic differences exist between both lineages most notably a prominent hump at the shoulders (Chan et al., 2010). Additionally, indicine cattle are more resistant to heat stress and have lower nutrient requirements (Canavez et al., 2012).

Taurine cattle, predominantly found in Europe, have been subjected to more intense selection for milk and meat production compared to Indicine cattle (McTavish et al., 2013). Intensification of selection in taurine cattle increased following the emergence of the breed concept 200 years ago (Taberlet et al., 2011). Consequently, more than 1,000 cattle breeds have been established many of which are specialised producers of either milk or meat. Intense selection for favourable alleles involved in different aspects of either milk or meat production has inevitably impacted the genetic structure of the bovine genome (Hayes et al., 2009; Bastiaansen et al., 2012).

The somatotropic axis is a complex network of genes, involving growth hormone (GH) and insulin-like growth factor 1 (IGF1), controlling several important traits in cattle (Renaville et al., 2002). Although much attention has been given to characterising the pivotal role played by GH and IGF1 (e.g. Lucy, 2008; Mullen et al., 2010; Mullen et al., 2011b; Waters et al., 2011), these genes form only part of the somatotropic axis. Genes from the axis have been implicated in health, metabolism, growth, fertility and lactation in cattle (Etherton, 2004; Lucy, 2008; Lucy et al., 2009; Mullen et al., 2012; Waters et al., 2012). Given the important role that this axis plays in many economically important traits, it is not surprising that beneficial polymorphisms within these genes may be under selection in traditional breeding programmes. More over, it is possible that this artificial selective pressure may have played a pivotal role in driving physiological changes leading to the establishment of breeds that are specialised producers of either dairy or beef products. Also, given the observed phenotypic and physiological differences between modern taurine and indicine cattle (Chan et al., 2010; O'Neill et al., 2010; Sartori et al., 2010), and the importance of the somatotropic axis on functionally important traits such as health and fertility, the somatotropic axis may have been an important factor that facilitated divergence of these lineages, and the physiological adaptation to differing environments.

Genetic and transcriptional differences between dairy and beef breeds have been reported previously (Gibbs et al., 2009; Hayes et al., 2009; MacEachern et al., 2009a; Sadkowski et al., 2009). Indeed, alleles segregating at different frequencies have been identified in dairy and beef populations (Hayes et al., 2009). Many studies have focused on the identification of recent signatures of selection. Identifying regions of the genome, specifically genes influencing economically important traits, under long term adaptive evolution between dairy and beef breeds can elucidate the mechanisms differentiating breeds, the impact of selection, inform conservation efforts and provide targets for future selection programmes (O'Neill et al., 2010). Thus, somatotropic axis genes should be good candidates for identifying positive selection that may have resulted from domestication.

The objective of this study was to identify genes from the somatotropic axis evolving at different rates, and under evolutionary selection pressure, in beef compared to dairy animals. Next-generation sequencing of 200 candidate genes was undertaken using DNA from beef and dairy taurine animals. To facilitate evolutionary hypotheses, the same regions were also sequenced in an out-group population of *Bos indicus* animals (Brahman).

6.2 Materials and Methods

6.2.1 Animal Selection

Genomic DNA was available for 750 Holstein-Friesian progeny-tested artificial insemination bulls (referred to as Dairy from here on). These animals represented phenotypes divergent for a number of production traits related to Holstein-Friesian selection goals (such as milk yield and carcass weight). An analysis involving a subset of 150 of these animals divergent for genetic merit of calving interval has already been published (Mullen et al., 2012). In addition to this, 300 beef animals composed of 6 beef breeds were also chosen. Fifty Aberdeen Angus, Belgian Blue, Charlois, Hereford, Limousine and Simmental animals were chosen representing the most popular beef breeds (collectively referred to as Beef from this point forward). Both the Dairy and Beef datasets were generated as part of a separate study, but were made available for use in this study.

For this study, an additional 55 unrelated Brahman animals were also chosen for targeted re-sequencing. These animals were derived from an African population that was not under intense selection for either Beef or Dairy production traits (dual-purpose use). Brahma were chosen as the out-group species because they are part of a separate taxonomic group to the Dairy and Beef but have a common ancestral species (the Auroch), i.e. Brahman are sister taxa to the Dairy and Beef breeds.

6.2.2 Gene Selection

In a previous study (Mullen et al., 2012), 83 of these genes were chosen based on there involvement in various aspects of the somatotropic axis such as gluconeogenesis and nutrient partitioning. For this study, the same 83 genes plus an additional 117 were selected for targeted re-sequencing. These genes were chosen based on a comprehensive review of the somatotropic axis, and are believed to be involved in traits such as growth, energy efficiency, metabolism, fertility and milk yield (Electronic Appendix 6.1).

6.2.3 Library preparation, target enrichment and sequencing

All procedures described in this section were carried out by Dr. Matthew McCabe at the Animal & Grassland Research and Innovation Centre, Teagasc. The procedures outlined for the Dairy and Beef datasets were carried out as part of a separate study which has already been published (see Mullen et al., 2012). These procedures are described in the following sections.

6.2.3.1 Dairy and Beef

For the Dairy animals, DNA was pooled in groups of 75 animals using equimolar quantities (100 ng) of DNA from each animal as described in (Mullen et al., 2012). Briefly, the pools were prepared for high-throughput 80 bp paired-end DNA sequencing using the Illumina Genome Analyser II platform. Indexed paired-end sequencing libraries were captured and enriched for the genes of interest using the SureSelect Target Enrichment for Illumina paired-end Sequencing (Agilent Technologies Ltd., Cork, Ireland) according to manufacturer’s instructions. Sequence capture baits were designed to target whole gene (exons and introns) sequences including 3 kb of both the 5’ and 3’ flanking UTR sequence for 83 genes central to the function of the somatotropic axis. To maximise the number of genes included for analysis, the remaining baits were designed to target only the coding sequences and 5’ and 3’ flanking UTR regions and encompassed 117 additional genes (the remaining genes). Targeted captured libraries for each pool contained different indexes located at the 5’ end of both reads, allowing multiple libraries to be sequenced on a single flow cell lane.

This procedure was repeated for the Beef animals using 6 pools of 50 animals (one pool for each breed) resulting in 80 bp paired-end DNA sequences from the Illumina Genome Analyser IIx (cluster kit 4PE and sequencing kit version 5) platform for the same regions.

6.2.3.1 Brahman

Standard paired-end DNA libraries were prepared from 6 captures of 55 samples using cleaned genomic DNA. Two of the 6 captures were pooled in equimolar amounts for high-throughput sequencing on the Illumina Hiseq2000 sequencing platform. The remaining 4 captures (similar in size and yield) were pooled for sequencing on an additional Hiseq2000 flowcell lane. All sequencing of was carried out at the Beijing Genomics Institute (BGI), Shenzhen, China.

6.2.4 Raw read quality control

In order to generate better mapping results, the raw FASTQ files were checked for quality before aligning to the Bovine genome. The program FastQC (version 0.10.1) (FastQC) was used to assess the quality of raw reads. The assessments included per base sequence quality, per sequence quality, per base N content and the sequence length distribution. Raw reads were then quality trimmed using the fastq-mcf command in the ea-utils package (Aronesty, 2013).

From the end of each read, bases with a phred score of less than 30 (99.9% accuracy in the base call) were trimmed. Reads that were less than 20 bps after trimming were discarded from further analysis.

6.2.5 Mapping and variant calling

All DNA sequence data were aligned to the UMD3.1 assembly of bovine genome using the Burrows-Wheeler Aligner (BWA) (version 0.6.1) (Li and Durbin, 2009). BWA is a fast read alignment tool that reports mappings in the standard SAM format. As all DNA sequence data came from the same pool of 55 samples, alignment files were merged. The merged alignment file was then sorted using Samtools (version 0.1.18) (Li et al., 2009) and filtered for possible PCR and optical duplicates using Picard tools (version 1.60) (The Picard toolkit).

The Genome Analysis Toolkit (GATK) (version 2.1) (McKenna et al., 2010) was then used for indel realignment and base recalibration. DNA sequence polymorphisms were then identified using the UnifiedGenotyper. Only variants with a phred-scaled confidence of 30 or higher were called (-stand_call_conf 30). The variant call format (VCF) file containing all identified polymorphisms in the Brahman dataset were then merged with the equivalent files from both the Dairy and Beef datasets, retaining information specific to the variant call in each group.

6.2.6 Variant QC and calling against the ancestral allele

Quality filtering of the merged VCF file was then undertaken to remove SNPs that were not biallelic, had a base quality score less than 30 or less than 7 reads at the SNP position. In addition, a minimum of 4 reads supporting the alternative (non-reference) allele was required across all groups. Any SNP that was only observed in the Brahman dataset were also removed from further analysis. After this, the alternative allele frequency was calculated for all SNPs in each of the three datasets. SNPs were then called against the ancestral allele based on the alternative allele frequency in the out-group dataset (Brahman). This was done by firstly calculating the alternative allele frequency for all SNPs. If the alternative allele at a SNP position was the majority allele in the Brahman dataset, the reference allele and alternative allele were switched and a new alternative allele frequency calculated for that SNP in all groups. SNPdat (version 0.1.5) (Doran and Creevey, 2013) was then used to identify synonymous and non-synonymous SNPs, and annotate all SNPs to genes from the UMD 3.1 assembly of the bovine genome.

For each gene, all SNPs that were annotated to a gene were identified using SNPdat. These SNPs were then concatenated to create a consensus sequence (consisting solely of the variants) for each gene based on the majority alternative allele at each SNP position. This was repeated for each group of animals (Dairy, Beef and Brahman) resulting in three sequences (in FASTA format) for all genes in the analysis.

6.2.7 Relative rate test

Relative rate tests measure genetic distances from an out-group to members of a related sister taxa in-group with which the out-group shares a common ancestor (Bleiweiss, 1998). Differences in genetic distance along these paths are taken to represent rate variation among in-group lineages (Figure 6.1).

Importantly, relative rate tests require that the paths from the out-group must pass through an internal node shared by all in-group taxa (Bleiweiss, 1998).

All genes with a consensus sequence less than 50 bps (i.e. had fewer than 50 SNPs detected across the three groups) were removed from further analysis. With the remaining gene sequences, a relative rate test was carried out using the RRTree software (Robinson-Rechavi and Huchon, 2000). RRTree allows us to calculate and compare substitution rates between sequences, compared to an out-group, using a relative rate test. Dairy and Beef sequence datasets were defined as the in-groups, and the Brahman sequence dataset as the out-group. Resultant p-values were corrected for errors arising from multiple testing using the Benjamini-Hochberg (Benjamini and Hochberg, 1995) method in R (The R Project). A corrected p-value <0.1 was used to identify significantly different rates between sequences of the in-groups.

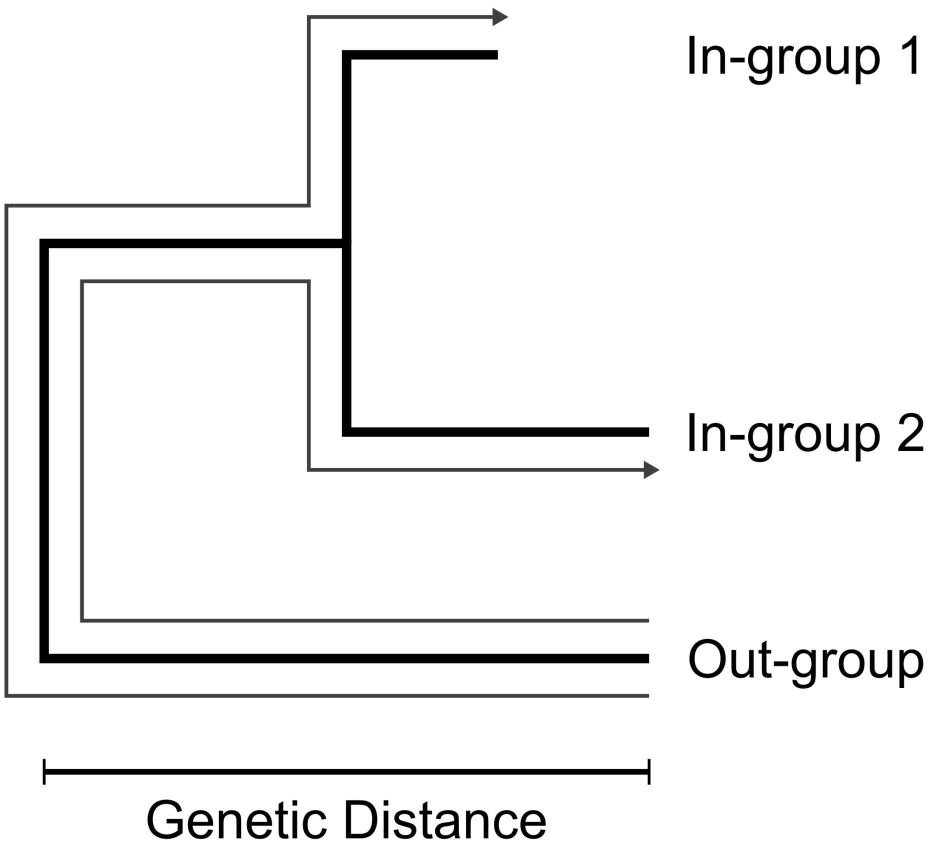


Figure 6.1 Distance-based relative rate tests from an out-group to two in-group taxa. A slower rate is inferred for in-group 1 based on a shorter path length from the out-group to in-group 1 compared to in-group 2 (Adapted from Bleiweiss, 1998).

6.2.8 McDonald-Kreitman test

McDonald and Kreitman (1991) proposed a statistical test to detect the level of adaptive evolution within a species at the molecular level (Parsch et al., 2009). If no selection is occurring, the ratio of synonymous to non-synonymous mutations within a species is expected to be equal to the ratio of synonymous to non-synonymous mutations between species. SNPdat was used to identify all synonymous and non-synonymous SNPs that occurred in coding regions of targeted genes. SNPs with an alternative allele frequency less than 0.95 were defined as polymorphic, whereas SNPs with an alternative allele frequency greater than or equal to 0.95 were defined as fixed in the population. To facilitate comparison of Beef to Dairy breeds, all SNPs that were fixed in both the dairy and beef datasets were removed. The number of fixed synonymous (D_s) and non-synonymous (D_n) SNPs, and the number of polymorphic synonymous (P_s) and non-synonymous (P_n) SNPs were counted for each gene. If no adaptive evolution has occurred then the ratio of synonymous to non-synonymous mutations between species should equal the ratio of synonymous to non-synonymous mutations within species (i.e. $D_n/D_s = P_n/P_s$). The ratio of synonymous to non-synonymous mutations between species (i.e. fixed SNPs) was compared to the ratio of synonymous to non-synonymous mutations within species using Fisher's exact probability test in R (The R Project). Although the G-test of independence was used in the original McDonald-Kreitman (1991) paper, Fisher's exact test is analogous to the G-test but should be used when sample size is low (Winters et al., 2010).

6.2.9 Uncharacterised genes

Genes that were not intended to be captured from the analysis but were identified as significant in either the relative rate test or the McDonald-Kreitman test were examined using the BLASTN v2.2.28 program (Zhang et al., 2000; Morgulis et al., 2008). For any such gene, the nucleotide sequence for the entire gene was extracted from the bovine genome. The program BLASTN v2.2.28 (Zhang et al., 2000; Morgulis et al., 2008) was used to identify orthologous genes based on the alignment of the gene sequence.

6.3 Results

6.3.1 Target enrichment and sequencing

For each animal group from the Dairy and Beef data sets, captured regions were sequenced as 80 bp paired-end reads on 9 lanes across 3 flowcells of an Illumina GAIIx. The average number of raw reads generated for each Dairy group was 21.96 million reads. Of these, the average number of reads that aligned to the bovine genome was 18.28 million reads. For the Beef groups, the average number of reads generated per group was 19.85 million reads with an average of 17.57 million reads aligning to the bovine genome.

For the Brahman datasets, raw reads were generated from 100 bp paired-end DNA sequencing on an Illumina Hiseq2000. On average 196.64 million paired reads were generated for each run. After quality trimming of the raw reads, 5 paired reads were removed and the average read length was 97.3 bps. On average 191.68 million paired reads aligned to the UMD3.1 bovine genome.

6.3.2 SNP discovery and annotation

In total, 16.87 million SNPs were identified in the Dairy, Beef and Brahman datasets. After quality filtering of SNPs (See Table 6.1 for details), 270,982 SNPs remained and were annotated using SNPdat (Doran and Creevey, 2013). SNPs were annotated to coding features (4,012), introns (67,846) and intergenic (199,124) regions. Approximately one third (97,755) of the SNPs annotated using SNPdat, were also reported in dbSNP (Sherry et al., 2001). Of the 4,012 SNPs annotated to exonic coding regions, 1,687 resulted in non-synonymous codon changes including 56 stop losses and 89 stop gains.

Table 6.1 The number of SNPs available after each quality filter was applied to the SNP data. The information in the second column is the total number of SNPs remaining after the quality filter (first column) was applied.

Quality Filter	Total SNPs Remaining
Raw SNPs	16,869,214
After non-biallelic removed	16,841,565
quality score less than 30	14,005,713
less than 7x coverage	511,332
less than 4 reads supporting alternative	498,664
Brahman only removed	298,260
Annotated by SNPdat	270,982

6.3.3 Relative rate test

After correcting for multiple testing, 6 genes were found to be evolving at significantly different rates in either Dairy or Beef (Table 6.2). Four of these were evolving at a significantly faster rate in Beef. Both the intercellular adhesion molecule 2 (ICAM2) and growth hormone receptor (GHR) genes were the most significantly faster in Beef genes (corrected p = 1.38×10^{-5}). The effect of GHR on bovine growth traits in both dairy and beef cattle has been well documented (Curi et al., 2005; Waters et al., 2011; Waters et al., 2012). The two genes that were significantly faster in Dairy were Dynein, axonemal, heavy chain 5 (DNAH5) and Dynein, axonemal, heavy chain 11 (DNAH11).

6.3.4 McDonald-Kreitman test

After removing SNPs that were fixed in both the Beef and Dairy populations, a total of 1,164 SNPs were annotated to coding regions of 173 targeted genes. Of these SNPs, 747 resulted in an amino acid change including

25 stop gains and 40 stop losses. The ratio of non-synonymous to synonymous polymorphic SNPs compared to the ratio of non-synonymous to synonymous fixed SNPs was not significantly different ($p < 0.05$) for any genes under investigation.

Table 6.2 Genes with significantly faster rates in Dairy and Beef (BH corrected p value < 0.1).

Gene ID	Gene Name	P	Faster	Padj
ENSBTAG00000019432	ICAM2	1.0×10^{-7}	Beef	1.38×10^{-5}
ENSBTAG00000001335	GHR	1.0×10^{-7}	Beef	1.38×10^{-5}
ENSBTAG00000021972	DNAH5	1.0×10^{-7}	Dairy	1.38×10^{-5}
ENSBTAG00000013078	DNAH11	1.0×10^{-7}	Dairy	1.38×10^{-5}
ENSBTAG00000018303	PAPPA2	2.17×10^{-5}	Beef	2.39×10^{-3}
ENSBTAG00000030416	NA	1.12×10^{-2}	Beef	1.12×10^{-2}

6.3.5 Uncharacterised genes

Only one gene (ENSBTAG00000030416) that was unintentionally captured in the sequencing was identified as having a significantly faster rate in Beef animals compared to Dairy animals. In Ensembl release 74 (Flicek et al., 2012), this gene is classified as a novel processed pseudogene. Using BLASTN (Zhang et al., 2000; Morgulis et al., 2008), this gene had the best alignment score and shared a high level of identity (92%) with the bovine solute carrier family 2, member 3 (SLC2A3 also known as GLUT3). This gene, GLUT3, was included in the target list of genes from the somatotropic axis.

6.4 Discussion

The aim of this study was to identify bovine somatotropic axis genes involved in evolutionary changes leading to differentiation of taurine dairy and beef breeds. Targeted sequencing of 200 somatotropic axis genes was used to identify synonymous and non-synonymous SNPs in taurine and indicine cattle. The McDonald-Kreitman test was used to investigate adaptive evolution occurring in dairy and beef (taurine) genomes, and a relative rate test was performed to identify genes evolving at significantly faster rates in either Beef or Dairy cattle.

6.4.1 Gene selection

The somatotropic axis is a complex network of genes central to a number of economically important traits in both dairy and beef cattle. A number of studies have investigated genes from the somatotropic axis and reported roles in health, fertility, lactation and growth (Lucy, 2008; Keady et al., 2011; Mullen et al., 2012; Waters et al., 2012). In this study, 200 genes representing the central components of the axis were chosen for investigation. These genes were chosen based on a thorough review of the literature and represent a comprehensive list of genes involved in traits of functional (i.e. health and fertility) and production (i.e. lactation and growth) importance in cattle. Moreover, due to the critical role that the somatotropic axis plays in influencing complex traits and the obvious phenotypic differences that exist between dairy and beef animals, these genes represent a candidate list of genes distinguishing taurine beef and dairy breeds.

6.4.2 Relative rate test

Cattle were first domesticated approximately 10,000 years ago. Since then, cattle have been under selection for several traits related to growth and milk. Selection pressure dramatically increased about 200 years ago, eventually leading to the development of many breeds specialised in either beef or milk production. If changes in the genetic structure of the genes targeted in this study have been involved in the phenotypic differentiation of Beef and Dairy breeds observed today, even before the emergence of the breed concept, it is likely that one breed may have experienced a faster rate of frequency change compared to the other.

6.4.2.1 Growth

Several genes related to growth were found to be evolving at significantly faster rates in Beef animals compared to Dairy animals (Table 6.2). One of these genes, pappalysin 2 (PAPPA2), has been shown to cleave insulin-like growth factor binding protein 5 (IGFBP5) resulting in increased IGF1 bioavailability (Overgaard et al., 2001b; Yan et al., 2010). Interestingly, IGFBP5 has been implicated in fat deposition in pigs and cattle (Fan et al., 2009; Wang et al., 2009a). The function of PPAPA2 is still at an early stage of elucidation (Conover, 2012), although studies have indicated a role in human pregnancy (Christians and Gruslin, 2010), reproduction in cattle (Luna-Nevarez et al., 2011) and growth in mice (Conover et al., 2011). Furthermore, quantitative trait locus (QTL) mapping in mice identified PAPPA2 as a candidate gene associated with postnatal growth and regulation of body size (Christians et al., 2006). In fact, mice homozygous for a PAPPA2 knockout were normal size at birth but suffered from postnatal growth retardation (Conover et al., 2011). In addition, a PAPPA2 knockout has also been shown to affect bone size and shape in growing mice (Christians et al., 2013).

Another interesting result was the identification of a significantly faster rate of evolution in the growth hormone receptor (GHR). Growth hormone (GH) is an anterior pituitary secreted hormone whose actions are mediated by the ubiquitously expressed GHR (Pilecka et al., 2007). The primary role of GH is to promote postnatal growth (Kopchick and Andry, 2000). In fact, administration of recombinant GH has been shown to increase muscle mass of growing steer calves (Vann et al., 2001). GH also plays an important role in metabolic processes, promoting decreases in fat and increases in lean body mass (Herrington and Carter-Su, 2001). Binding of GH to GHR induces transcription of several genes including IGF1 (Jiang et al., 2007). IGF1 is an important circulating growth factor controlling tissue growth (Schiaffino and Mammucari, 2011; Bonaldo and Sandri, 2013). Most IGF1 in circulation is found as part of a complex with one of six IGFBPs, which increases the half-life of IGF1 in blood. IGF1 null mice exhibit severe growth retardation (Baker et al., 1993). Similarly, knockout of GH or GHR results in mice which are 50% smaller than their wild-type littermates (Jiang and Ge, 2013). Furthermore, several studies have identified associations between GH and GHR, and growth traits in cattle (Mullen et al., 2010; Mullen et al., 2011a; Waters et al., 2012).

A faster rate of evolution in both of these genes supports the hypothesis that Beef breeds, more than Dairy breeds, have experienced positive selection of alleles for growth characteristics. These changes may also have helped drive the changes that lead to the establishment of breeds specialised in meat production.

6.4.2.2 Angiogenesis and inhibition of apoptosis

Angiogenesis is an important process in normal postnatal skeletal muscle growth (Takahashi et al., 2002). Angiogenesis in muscle tissue is the process in which new blood vessels form from existing vessels promoting muscle growth. Intercellular adhesion molecule 2 (ICAM2), which was identified as evolving at a faster rate in Beef animals compared to Dairy ($p = 1.38 \times 10^{-5}$), has been suggested as a potential regulator of angiogenesis (Huang et al., 2005). In both *in vitro* and *in vivo* studies involving mice, lack of ICAM2 expression resulted in impaired angiogenesis and increased apoptosis (Huang et al., 2005). Furthermore, ICAM2 mediates a cell survival signal sufficient to block apoptosis by activation of PI3K/AKT pathway (Perez et al., 2002; Ishigami et al., 2008). Activation of this pathway stimulates protein synthesis and inhibits apoptosis (Clemmons, 2009). Interestingly, this pathway is also activated by insulin-like growth factors (IGFs) including IGF1 (Rommel et al., 2001), which is a central component of the somatotropic axis important in the regulation of myogenesis (Straface et al., 2009). Beef animals exhibit faster and more pronounced postnatal growth than dairy counterparts. As such, angiogenesis and inhibition of apoptosis through ICAM2 may be undergoing selection in tandem with genes involved in muscle growth (such as GHR and PAPPA2) and that these genes have experienced faster rates of frequency change in Beef breeds.

6.4.3 Fertility

Two genes (dynein, axonemal, heavy chain (DNAH) 5 and DNAH11) related to fertility were found to be evolving at a faster rate in Dairy animals compared to Beef animals. Both genes are members of the dynein heavy chain family of proteins encoding ciliary outer dynein arm proteins. Both DNAH5 and DNAH11 function as force generating proteins with ATPase activity involved in the movement of cilia (Chodhari et al., 2004). In cattle, very little has been reported regarding the function of either DNAH5 or DNAH11. In spite of this, DNAH5 and DNAH11 have both been implicated in primary ciliary dyskinesia (PCD) and *situs inversus* (reversed internal organs) (Bartoloni et al., 2002; Olbrich et al., 2002; Hornef et al., 2006). Several studies have identified mutations in DNAH5 and DNAH11 that are believed to cause PCD (Olbrich et al., 2002; Hornef et al., 2006; Pifferi et al., 2010; Knowles et al., 2012; Zhang et al., 2013a). PCD of males is characterised by reduced fertility due to sperm immotility (Chodhari et al., 2004; Hornef et al., 2006).

For many years selection in dairy cattle focused solely on increasing milk yield (Miglior et al., 2005). This was the case for most countries with the exception of Scandinavian countries, in which functionally important traits (such as health and reproduction) were included in breeding goals. Numerous studies have demonstrated that selection for only production traits, such as milk yield, causes negative effects on reproductive performance (Veerkamp et al., 2001; Kadarmideen et al., 2003). However, in recent years, breeding goals have diversified to include functionally important traits in an effort to reduce and reverse the decline in these traits (Miglior et al., 2005). The faster rates of evolution of DNAH5 and DNAH11 may be due to the historical indirect selection against fertility and the recent recovery of fertility rates through breeding however, further research is required.

6.4.4 Unintentional capture of a GLUT3 pseudogene

One gene which was not included on the targeted list but was found to be evolving at a significantly faster rate in Beef animals was a processed pseudogene, ENSBTAG00000030416. Although, this gene was not intended to be sequenced, it shares a high level of similarity with a gene which was targeted for sequencing (GLUT3). In humans, GLUT3 is predominantly expressed in brain and neural tissue although expression in is also found in skeletal muscle tissue (Bilan et al., 1992; Copland et al., 2007). Interestingly, muscle GLUT3 expression has been shown to be controlled by IGF1 (Copland et al., 2007). In spite of the similarity (92%) between GLUT3 and ENSBTAG00000030416, little can be said about the significance of this finding without further research. Although many pseudogenes have been identified in humans (Zhang et al., 2003) (annotations in cattle are not as complete), the common assumption is that a pseudogene is non-functional (Zheng et al., 2007). Pseudogenes are generally considered the evolutionary endpoint of genomic material that will eventually be removed from a genome (Zheng et al., 2007). Significant differences in this gene may represent a false positive of the study due to sequencing error (such as a low specificity primer) or alignment error (in which reads originating from a gene are mapped to a different gene with high similarity)(Copland et al., 2007).

6.4.5 Search for adaptive evolution

The McDonald-Kreitman (MK) test, which compares the ratio of polymorphic to fixed (divergent) SNPs at non-synonymous and synonymous sites, is a common method for the detection of positive (or balancing) selection on protein coding sequences (McDonald and Kreitman, 1991; MacCallum and

Hill, 2006; Parsch et al., 2009). The MK test has been used to investigate adaptation in several species including humans and cattle (Bustamante et al., 2005; MacEachern et al., 2009b). The test is relative straightforward to perform: A two-by-two contingency is created in which the columns are the fixed and polymorphic SNPs, and the rows are for the synonymous and non-synonymous SNPs. Thus, the four entries in the table are the number of fixed synonymous (D_s) and non-synonymous (D_n) SNPs, and the number of polymorphic synonymous (P_s) and non-synonymous (P_n). The MK test compares the ratio of D_n/D_s to P_n/P_s to evaluate whether neutral evolution can be rejected as the evolutionary force dictating change in a coding sequence (Messer and Petrov, 2013). Deviations from the expected ratio (i.e. $D_n/D_s = P_n/P_s$) may indicate either positive selection ($D_n/D_s > P_n/P_s$) or purifying (or balancing) selection ($D_n/D_s < P_n/P_s$). However, no significant deviations from the expected ratio were identified in any of the genes under investigation in this study.

A limitation of this approach for individual genes is the possibility that too few SNPs will be detected in the coding region of target genes (MacEachern et al., 2006), which was the case in this study. Furthermore, many sites under selection are likely to be outside of coding regions, and therefore are not covered by the MK test (Walsh, 2008). For many of the genes analysed, the MK test did not have sufficient power to detect significant deviations as the number of SNPs identified in the coding regions was quite low. For example, the majority of genes (104/173) had 5 or less SNPs annotated to the coding region.

One possibility which may result in a low number of non-synonymous fixed mutations, is that genes that are functionally important may be highly conserved across breeds and that even minor changes in protein structure are under purifying selection (Schmitt et al., 2007). This is further complicated by domestication (MacEachern et al., 2009a; MacEachern et al., 2009b). As a consequence of artificial selection and a small effective population, selective constraint may be reduced meaning that unfavourable mutations are not removed as quickly as in a population under natural selection (MacEachern et al., 2009b). This would explain the low number of fixed non-synonymous SNPs and the large number of polymorphic non-synonymous SNPs observed in both Beef and Dairy. For example, in the Beef population 745/1164 SNPs were non-synonymous polymorphic. Although a large number of polymorphic non-synonymous SNPs may be indicative of balancing selection, it may also be possible, that alleles within a gene are under positive selection but have yet to be fixed in either population. Furthermore, domestication and a reduction in effective population size may make the MK test unsuitable for cattle data (or any domestic animals) (MacEachern et al., 2009b).

Chapter 7: Thesis Discussion

7.1 Introduction

Systems biology is not only concerned with the analysis of raw data, it also entails the development of new tools and approaches for efficiently analysing data. Especially in an era of rapid high-throughput sequencing, methodological advances coupled with increased sequencing data can provide a solid foundation to integrative analysis of complex traits. The aim of this thesis was to combine these key concepts, of software development, methodological improvement and data analysis, to address questions pertaining to bovine muscle growth and development.

Chapter two describes the development and implementation of a SNP annotation tool, SNPdat. SNPdat was developed to address a need for such tools in the analysis of non-model organisms. In this chapter, it was demonstrated that SNPdat can be used to analyse data and provide similar results to currently available software, but that SNPdat can also be used to analyse data from many organisms not currently supported by the vast majority of currently available tools. Additionally, in this chapter, the development and implementation of additional scripts that are part of the SNPdat package is described. This tool is user friendly and can easily be incorporated into existing SNP discovery pipelines. SNPdat fills a niche for analyses involving non-model organisms that are not supported by many available SNP annotation tools. SNPdat will be of great interest to scientists involved in SNP discovery and analysis projects, particularly those with limited bioinformatics experience.

In the third chapter, two different statistical approaches were used to identify regions of the bovine genome associated with four economically important traits related to animal growth. Regions found significantly associated with a trait were then further investigated to identify potential biological processes involved in different aspects of animal growth. Results from this chapter will be of particular interest to quantitative geneticists involved in genomic selection. Significant associations could be incorporated into a genomic selection programme to identify animals with better potential for growth. Also, candidate genes and biological processes identified in this chapter have shed light on some of the regulatory mechanisms involved in animal growth. Particularly, the involvement of the peroxisome proliferator-activated receptor signalling pathway in a number of traits from this study confirmed the multi-faceted influence pathways can have on different aspects of each trait. Candidate genes and pathways identified could also form the basis for further studies that aim to identify causative mutations influencing each of these traits.

In recent years, Bayesian approaches have gained considerable momentum in the area of genetic prediction. Accurate assessment of convergence is an important step in Bayesian analysis and posterior inference. However, the concept of convergence diagnostics has largely been over-looked in the area of genetic prediction in livestock species such as cattle. Chapter four focused on convergence assessment techniques for a Bayesian model used in genetic prediction of complex traits. A number of metrics were implemented to assess convergence. It is hoped that the metrics, and approaches outlined, will be adopted by the larger community of researchers involved in genetic prediction of complex traits.

RNA-seq is a next-generation technology that has quickly been adapted by many researchers investigating transcriptional responses in experimental conditions. In chapter five, the transcriptional responses of skeletal muscle to nutritional restriction and compensatory growth were investigated using RNA-seq technology. Paired-end sequencing and bioinformatic approaches were utilised in this chapter. Paired-end sequencing is advantageous compared to single read sequencing or fluorescence based approaches (such microarrays). Interestingly, the peroxisome proliferator-activated receptor signalling pathway was found significantly over-represented in study. This supported evidence from chapter 3, that this pathway is involved in many aspects of animal growth.

Obvious morphological differences exist between dairy and beef breeds of cattle, in particular in the ability to accumulate muscle. In chapter six it was hypothesised that genes involved in regulating muscle growth might be evolving at different rates in beef cattle compared to dairy, possibly due to domestication and selection. Interestingly, the growth hormone receptor was found to be evolving at a faster rate in beef animals compared to dairy. This result further underlines the importance of growth hormone and its receptor in driving bovine muscle growth, and represents a candidate gene for inclusion in genetic breeding programmes.

7.2 General discussion

A major challenge in the analysis of SNP data derived from the analysis of non-model organisms is the dearth of species non-specific software. The majority of tools available for the analysis of SNP data are solely for the purpose of analysing human SNP data. Consequently, the interrogation of data derived from non-model organisms is often challenging, particularly so for researchers with limited bioinformatics experience. In chapter 2, the development and implementation of SNPdat is described. SNPdat was developed to fill a gap in available software for the analysis of SNP data originating from non-model organisms. SNPdat can be used to analysis SNP data for any organism with a reference sequence and annotation. The usefulness of SNPdat was also demonstrated in chapter 2 using a published dataset of both novel and known bovine SNPs (Mullen et al., 2012).

At various stages throughout this thesis, SNPdat was used to facilitate the analysis of bovine data. In chapter 3, SNPdat was used to annotate significantly associated SNPs to known QTL. Although SNPdat was designed to annotate SNPs to genomic features such as exons, it can be used to annotate SNPs to any information contained in a GTF/GFF file. Currently, to annotate SNPs to QTL information stored in databases, such as cattle QTLdb (Hu et al., 2013), requires the creation of custom scripts. However, as information stored in cattle QTLdb can be retrieved in GFF3 format, SNPdat can be used to annotate SNPs to known QTL. As such, SNPdat can facilitate the identification of candidate genes surrounding significantly associated SNPs and provide valuable insight into the functional roles of those genes.

As described in chapter 2, SNPdat can also identify synonymous and non-synonymous SNPs. This function of SNPdat was exploited in chapter 6, in which synonymous and non-synonymous SNPs were identified in somatotropic axis genes of dairy, beef and Brahman animals. This facilitated the identification of genes under significantly different rates of evolution in beef animals compared to dairy animals. Additionally, SNPdat was used to identify SNPs in coding regions which enabled the examination of genes from the somatotropic axis which may be under evolutionary selection pressure.

As outlined in chapter 1, systems biology approaches also entail the development and investigation of novel methods to improve the analysis of experimental data. This concept was addressed in chapter 4, in which graphical approaches to diagnose convergence were assessed. Although it may never be entirely possible to guarantee convergence of the MCMC chain in Bayesian inference, lack of convergence can often be identified (Cowles and Carlin, 1996). As such, convergence assessment will remain an important consideration in Bayesian inference. The conclusions outlined in chapter 4 were important in driving the approaches taken to assess convergence in chapter 3.

Muscle growth is clearly a complex process, influenced by many genetic factors. Integrating data from several sources, as in a systems biology approach, may provide valuable insight as to the biological mechanisms controlling various phenotypic outcomes related to growth. In total, 12 genes involved in the PPAR signalling pathway were identified in chapters 3 and 5. However, these 12 genes comprised results from the analysis of CWT, CFAT and CONF data in a genome-wide association (chapter 3) coupled with significantly differentially expressed genes from a compensatory growth model (chapter 5).

In chapter 3, the multi-functional role that biological pathways may play in linking different aspects of bovine growth was highlighted. This was supported by the identification of 6 pathways that were significantly over-represented in the analysis of a single carcass trait and also for the combined trait analysis. For example, the PPAR signalling pathway contained genes surrounding SNPs significantly associated CWT, CFAT and CONF. PPARs are a group of transcription factors that play a central role in controlling skeletal muscle lipid utilisation (Berger and Moller, 2002; Ehrenborg and Krook, 2009). There are three members of the PPAR family; PPARα, PPARγ and PPARδ, each of which is encoded by a separate gene (Abbott, 2009). PPARs regulate transcription by binding with retinoid X receptors (Tien et al., 2006). This heterodimer binds to peroxisome proliferator response elements in the promoter region of target genes which then stimulates expression (Tan et al., 2005). In chapter 3, PPARα and retinoid X receptor α were identified as candidate genes involved in regulating CWT and CONF, respectively.

PPARα is also involved in controlling the expression of fatty acid binding proteins (FABPs), which are a family of carrier proteins involved in fatty acid metabolism (Furuhashi and Hotamisligil, 2008). Interestingly, FABP4 was significantly down-regulated during the differential feeding period when animals are experiencing a depressed growth phase due to nutritional restriction (chapter 5). In fact, 6 genes from the PPAR signalling pathway were significantly differentially expressed at this time point, the majority of which were down-regulated. The PPAR signalling pathway was also significantly over-represented at this time point suggesting that this pathway may play an important role in partitioning of energy during a depressed growth phase due to energy restriction. In line with the findings of chapter 3 and chapter 5, it is clear that the PPAR signalling pathway plays a crucial role in influencing several aspects of bovine growth.

Furthermore, the compensatory growth study (chapter 5) provided several novel insights into the biological mechanisms regulating bovine muscle growth. Although several efforts have previously been made to elucidate the mechanisms underpinning accelerated growth in response to nutrient availability following nutritional restriction (Park, 2005; Lehnert et al., 2006; Fiems et al., 2007), many of the genes and biological pathways involved in CG remained concealed. The application of RNA-seq, a highly sensitive approach, allowed the examination of differential gene expression in animals exhibiting CG compared to control animals. From the RNA-seq analysis of *M. longissimus thoracis et lumborum* tissue, which is an economically important tissue in beef production, at two separate time points it is evident that several complex processes are activated, involving up and down-regulation of many genes. Genes that are differentially expressed during re-alimentation offer revealing insights as the mechanisms controlling rapid muscle growth. Additionally, these genes and pathways are potential targets to assess CG in other muscle tissues. In beef production systems, CG may be utilised to offset high feed costs over Winter however, the impact of CG on economically desirable traits such as meat quality and taste warrants further investigation.

Since earliest domestication, cattle have been utilised for meat products and consequently have experienced long periods of soft selection. This was followed by intensification of selection in last 200 years (Taberlet et al., 2011). These periods of selection have undoubtedly had an influence on the underlying genetic architecture of taurine breeds (Gibbs et al., 2009), potentially leading to increased rates of evolution in genes relevant to a desirable phenotypic outcome (for example, milk or meat traits). This hypothesis is supported by the identification of several genes evolving at significantly different rates in beef breeds. This also supports the view that domestication and selection has had and is still having an impact of the genetic structure of beef and dairy breeds separately.

For example, GH was identified as a candidate gene influencing the CFAT and CONF traits, and GHR as a candidate gene involved in CONF and CULL (chapter 3). This supports evidence from several studies in which GH and GHR were found to be involved in various aspects of growth (Curi et al., 2005; Mullen et al., 2010; Waters et al., 2011). In addition, GHR was also found to be evolving at a significantly faster rate in beef animals compared to dairy animals (chapter 6). The primary function of GH is to stimulate growth, and through GHR this function is regulated (Kopchick and Andry, 2000). Upon ligand binding, GHR stimulates the transcription of IGF1 (among other genes) (Jiang et al., 2007). These genes; GH, GHR and IGF1 form the central components of the somatotropic axis (Keady et al., 2013), so it is not surprising that they are involved in several aspects of bovine growth and have been under selection.

Another gene that was identified as evolving at a significantly different rate, PAPPA2, was also significantly down-regulated during the differential feeding period of the compensatory growth study (chapter 5). PAPPA2 has been shown to increase bioavailability of IGF1 through cleavage of IGFBP5, thus stimulating muscle proliferation (Overgaard et al., 2001a). Down-regulation of PPAPA2 during a prolonged period of nutrient restriction would seem to suggest a role regulating muscle growth. This highlights the importance of incorporating information from several sources, thus enabling the identification of genes involved in different aspects of bovine growth, namely the short-term down-regulation of tissue proliferation and the long-term evolutionary importance of this gene in growth.

The mechanisms regulating animal growth are undoubtedly extremely complex. A decade ago, the analyses outlined in this thesis would not have been possible. However, the recent publication of the bovine genome, coupled with improving sequencing technology and methodological advances, has allowed new levels of understanding to be reached. This thesis is evidence of the obvious progression in bovine research, and as such has offered many revealing insights into the key genes and pathways regulating growth in cattle.

Chapter 8: Bibliography

ABBOTT, B. D. (2009). Review of the expression of peroxisome proliferator-activated receptors alpha (PPAR alpha), beta (PPAR beta), and gamma (PPAR gamma) in rodent and human development. *Reproductive toxicology,* 27**,** 246-57.

ACHILLI, A., BONFIGLIO, S., OLIVIERI, A., MALUSA, A., PALA, M., HOOSHIAR KASHANI, B., PEREGO, U. A., AJMONE-MARSAN, P., LIOTTA, L., SEMINO, O., BANDELT, H. J., FERRETTI, L. & TORRONI, A. (2009). The multifaceted origin of taurine cattle reflected by the mitochondrial genome. *PloS one,* 4**,** e5753.

ADEREM, A. (2005). Systems biology: its practice and challenges. *Cell,* 121**,** 511-3.

ALI, M., NICIEZA, A. & WOOTTON, R. J. (2003). Compensatory growth in fishes: a response to growth depression. *Fish and Fisheries,* 4**,** 147-190.

ALLAIS, S., LEVEZIEL, H., PAYET-DUPRAT, N., HOCQUETTE, J. F., LEPETIT, J., ROUSSET, S., DENOYELLE, C., BERNARD-CAPEL, C., JOURNAUX, L., BONNOT, A. & RENAND, G. (2010). The two mutations, Q204X and nt821, of the myostatin gene affect carcass and meat quality in young heterozygous bulls of French beef breeds. *Journal of animal science,* 88**,** 446-54.

ALLAN, M. F. & SMITH, T. P. (2008). Present and future applications of DNA technologies to improve beef production. *Meat science,* 80**,** 79-85.

ALLEN, D. L. & UNTERMAN, T. G. (2007). Regulation of myostatin expression and myoblast differentiation by FoxO and SMAD transcription factors. *American journal of physiology. Cell physiology,* 292**,** C188-99.

ALTSHULER, D., POLLARA, V. J., COWLES, C. R., VAN ETTEN, W. J., BALDWIN, J., LINTON, L. & LANDER, E. S. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature,* 407**,** 513-6.

AMER, P. R., SIMM, G., KEANE, M. G., DISKIN, M. G. & WICKHAM, B. W. (2001). Breeding objectives for beef cattle in Ireland. *Livestock Production Science,* 67**,** 223-239.

ANDERS, S. & HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome biology,* 11**,** R106.

ANDERSSON, L. & GEORGES, M. (2004). Domestic-animal genomics: deciphering the genetics of complex traits. *Nature reviews. Genetics,* 5**,** 202-12.

ANSORGE, W. J. (2009). Next-generation DNA sequencing techniques. *New biotechnology,* 25**,** 195-203.

ARONESTY, E. 2013. *ea-utils* [Online]. Available: http://code.google.com/p/ea-utils/.

ASHWORTH, A. (1969). Growth rates in children recovering from protein-calorie malnutrition. *The British journal of nutrition,* 23**,** 835-45.

AUFFRAY, C., IMBEAUD, S., ROUX-ROUQUIE, M. & HOOD, L. (2003). From functional genomics to systems biology: concepts and practices. *Comptes rendus biologies,* 326**,** 879-92.

BAGNATO, A., SCHIAVINI, F., ROSSONI, A., MALTECCA, C., DOLEZAL, M., MEDUGORAC, I., SOLKNER, J., RUSSO, V., FONTANESI, L., FRIEDMANN, A., SOLLER, M. & LIPKIN, E. (2008). Quantitative trait loci affecting milk yield and protein percentage in a three-country Brown Swiss population. *Journal of dairy science,* 91**,** 767-83.

BAKER, J., LIU, J. P., ROBERTSON, E. J. & EFSTRATIADIS, A. (1993). Role of insulin-like growth factors in embryonic and postnatal growth. *Cell,* 75**,** 73-82.

BALDING, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature reviews. Genetics,* 7**,** 781-91.

BARRETT, T., WILHITE, S. E., LEDOUX, P., EVANGELISTA, C., KIM, I. F., TOMASHEVSKY, M., MARSHALL, K. A., PHILLIPPY, K. H., SHERMAN, P. M., HOLKO, M., YEFANOV, A., LEE, H., ZHANG, N., ROBERTSON, C. L., SEROVA, N., DAVIS, S. & SOBOLEVA, A. (2013). NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research,* 41**,** D991-5.

BARTOLONI, L., BLOUIN, J. L., PAN, Y., GEHRIG, C., MAITI, A. K., SCAMUFFA, N., ROSSIER, C., JORISSEN, M., ARMENGOT, M., MEEKS, M., MITCHISON, H. M., CHUNG, E. M., DELOZIER-BLANCHET, C. D., CRAIGEN, W. J. & ANTONARAKIS, S. E. (2002). Mutations in the DNAH11 (axonemal heavy chain dynein type 11) gene cause one form of situs inversus totalis and most likely primary ciliary dyskinesia. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 10282-6.

BARWICK, S. A. & HENZELL, A. L. (2005). Development successes and issues for the future in deriving and applying selection indexes for beef breeding. *Australian Journal of Experimental Agriculture,* 45**,** 923-933.

BASTIAANSEN, J. W. M., COSTER, A., CALUS, M. P. L., VAN ARENDONK, J. A. M. & BOVENHUIS, H. (2012). Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genetics Selection Evolution,* 44.

BEIKO, R. G., KEITH, J. M., HARLOW, T. J. & RAGAN, M. A. (2006). Searching for convergence in phylogenetic Markov chain Monte Carlo. *Systematic biology,* 55**,** 553-65.

BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological,* 57**,** 289-300.

BENNETT, S. T., BARNES, C., COX, A., DAVIES, L. & BROWN, C. (2005). Toward the 1,000 dollars human genome. *Pharmacogenomics,* 6**,** 373-82.

BERGER, J. & MOLLER, D. E. (2002). The mechanisms of action of PPARs. *Annual review of medicine,* 53**,** 409-35.

BERNUES, A., RUIZ, R., OLAIZOLA, A., VILLALBA, D. & CASASUS, I. (2011). Sustainability of pasture-based livestock farming systems in the European Mediterranean context: Synergies and trade-offs. *Livestock science,* 139**,** 44-57.

BERRY, D. P., MEADE, K. G., MULLEN, M. P., BUTLER, S., DISKIN, M. G., MORRIS, D. & CREEVEY, C. J. (2011). The integration of 'omic' disciplines and systems biology in cattle breeding. *Animal : an international journal of animal bioscience,* 5**,** 493-505.

BILAN, P. J., MITSUMOTO, Y., MAHER, F., SIMPSON, I. A. & KLIP, A. (1992). Detection of the GLUT3 facilitative glucose transporter in rat L6 muscle cells: regulation by cellular differentiation, insulin and insulin-like growth factor-I. *Biochemical and biophysical research communications,* 186**,** 1129-37.

BLACK, W. C. T., BAER, C. F., ANTOLIN, M. F. & DUTEAU, N. M. (2001). Population genomics: genome-wide sampling of insect populations. *Annual review of entomology,* 46**,** 441-69.

BLANDIN, G., MARCHAND, S., CHARTON, K., DANIELE, N., GICQUEL, E., BOUCHETEIL, J. B., BENTAIB, A., BARRAULT, L., STOCKHOLM, D., BARTOLI, M. & RICHARD, I. (2013). A human skeletal muscle interactome centered on proteins involved in muscular dystrophies: LGMD interactome. *Skeletal muscle,* 3**,** 3.

BLEIWEISS, R. (1998). Relative-rate tests and biological causes of molecular evolution in hummingbirds. *Molecular biology and evolution,* 15**,** 481-491.

BOISCLAIR, Y. R., RHOADS, R. P., UEKI, I., WANG, J. & OOI, G. T. (2001). The acid-labile subunit (ALS) of the 150 kDa IGF-binding protein complex: an important but forgotten component of the circulating IGF system. *The Journal of endocrinology,* 170**,** 63-70.

BOLORMAA, S., NETO, L. R., ZHANG, Y. D., BUNCH, R. J., HARRISON, B. E., GODDARD, M. E. & BARENDSE, W. (2011). A genome-wide association study of meat and carcass traits in Australian cattle. *Journal of animal science,* 89**,** 2297-309.

BONALDO, P. & SANDRI, M. (2013). Cellular and molecular mechanisms of muscle atrophy. *Disease Models & Mechanisms,* 6**,** 25-39.

BOURDON, R. M. (1998). Shortcomings of current genetic evaluation systems. *Journal of animal science,* 76**,** 2308-2323.

BREUER, K., FOROUSHANI, A. K., LAIRD, M. R., CHEN, C., SRIBNAIA, A., LO, R., WINSOR, G. L., HANCOCK, R. E., BRINKMAN, F. S. & LYNN, D. J. (2013). InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic acids research,* 41**,** D1228-33.

BROOKES, A. J. (1999). The essence of SNPs. *Gene,* 234**,** 177-86.

BROOKS, S. P. & GELMAN, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics,* 7**,** 434-455.

BROOKS, S. P. & ROBERTS, G. O. (1998). Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing,* 8**,** 319-335.

BRUFORD, M. W., BRADLEY, D. G. & LUIKART, G. (2003). DNA markers reveal the complexity of livestock domestication. *Nature reviews. Genetics,* 4**,** 900-10.

BRUGGEMAN, F. J. & WESTERHOFF, H. V. (2007). The nature of systems biology. *Trends in microbiology,* 15**,** 45-50.

BURKS, T. N. & COHN, R. D. (2011). Role of TGF-beta signaling in inherited and acquired myopathies. *Skeletal muscle,* 1**,** 19.

BURT, D. W. (2009). The cattle genome reveals its secrets. *Journal of biology,* 8**,** 36.

BUSTAMANTE, C. D., FLEDEL-ALON, A., WILLIAMSON, S., NIELSEN, R., HUBISZ, M. T., GLANOWSKI, S., TANENBAUM, D. M., WHITE, T. J., SNINSKY, J. J., HERNANDEZ, R. D., CIVELLO, D., ADAMS, M. D., CARGILL, M. & CLARK, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature,* 437**,** 1153-7.

CAMPION, B., KEANE, M. G., KENNY, D. A. & BERRY, D. P. (2009). Evaluation of estimated genetic merit for carcass weight in beef cattle: Live weights, feed intake, body measurements, skeletal and muscular scores, and carcass characteristics. *Livestock science,* 126**,** 87-99.

CANAVEZ, F. C., LUCHE, D. D., STOTHARD, P., LEITE, K. R., SOUSA-CANAVEZ, J. M., PLASTOW, G., MEIDANIS, J., SOUZA, M. A., FEIJAO, P., MOORE, S. S. & CAMARA-LOPES, L. H. (2012). Genome sequence and assembly of Bos indicus. *The Journal of heredity,* 103**,** 342-8.

CANOVAS, A., QUINTANILLA, R., AMILLS, M. & PENA, R. N. (2010). Muscle transcriptomic profiles in pigs with divergent phenotypes for fatness traits. *BMC genomics,* 11**,** 372.

CASSMAN, M. (2005). Barriers to progress in systems biology. *Nature,* 438**,** 1079.

CEDDIA, R. B., WILLIAM, W. N., JR. & CURI, R. (2001). The response of skeletal muscle to leptin. *Frontiers in bioscience : a journal and virtual library,* 6**,** D90-7.

CHAN, E. K., NAGARAJ, S. H. & REVERTER, A. (2010). The evolution of tropical adaptation: comparing taurine and zebu cattle. *Animal genetics,* 41**,** 467-77.

CHICUREL, M. (2002). Bioinformatics: Bringing it all together. *Nature,* 419**,** 751-+.

CHODHARI, R., MITCHISON, H. M. & MEEKS, M. (2004). Cilia, primary ciliary dyskinesia and molecular genetics. *Paediatric respiratory reviews,* 5**,** 69-76.

CHRISTIANS, J. K., DE ZWAAN, D. R. & FUNG, S. H. (2013). Pregnancy associated plasma protein A2 (PAPP-A2) affects bone size and shape and contributes to natural variation in postnatal growth in mice. *PloS one,* 8**,** e56260.

CHRISTIANS, J. K. & GRUSLIN, A. (2010). Altered levels of insulin-like growth factor binding protein proteases in preeclampsia and intrauterine growth restriction. *Prenatal diagnosis,* 30**,** 815-20.

CHRISTIANS, J. K., HOEFLICH, A. & KEIGHTLEY, P. D. (2006). PAPPA2, an enzyme that cleaves an insulin-like growth-factor-binding protein, is a candidate gene for a quantitative trait locus affecting body size in mice. *Genetics,* 173**,** 1547-53.

CLEMENT, K., VAISSE, C., LAHLOU, N., CABROL, S., PELLOUX, V., CASSUTO, D., GOURMELEN, M., DINA, C., CHAMBAZ, J., LACORTE, J. M., BASDEVANT, A., BOUGNERES, P., LEBOUC, Y., FROGUEL, P. & GUY-GRAND, B. (1998). A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature,* 392**,** 398-401.

CLEMMONS, D. R. (2009). Role of IGF-I in skeletal muscle mass maintenance. *Trends in Endocrinology and Metabolism,* 20**,** 349-356.

CLEMPSON, A. M., POLLOTT, G. E., BRICKELL, J. S., BOURNE, N. E., MUNCE, N. & WATHES, D. C. (2011). Evidence that leptin genotype is associated with fertility, growth, and milk production in Holstein cows. *Journal of dairy science,* 94**,** 3618-28.

COHUET, A., KRISHNAKUMAR, S., SIMARD, F., MORLAIS, I., KOUTSOS, A., FONTENILLE, D., MINDRINOS, M. & KAFATOS, F. C. (2008). SNP discovery and molecular evolution in Anopheles gambiae, with special emphasis on innate immune system. *BMC genomics,* 9**,** 227.

COLE, J. B., LEWIS, R. M., MALTECCA, C., NEWMAN, S., OLSON, K. M. & TAIT, R. G., JR. (2013). Breeding and Genetics Symposium: systems biology in animal breeding: Identifying relationships among markers, genes, and phenotypes. *Journal of animal science,* 91**,** 521-2.

CONNOR, E. E., KAHL, S., ELSASSER, T. H., PARKER, J. S., LI, R. W., VAN TASSELL, C. P., BALDWIN, R. L. & BARAO, S. M. (2010). Enhanced mitochondrial complex gene function and reduced liver size may mediate improved feed efficiency of beef cattle during compensatory growth. *Functional & integrative genomics,* 10**,** 39-51.

CONOVER, C. A. (2012). Key questions and answers about pregnancy-associated plasma protein-A. *Trends in endocrinology and metabolism: TEM,* 23**,** 242-9.

CONOVER, C. A., BOLDT, H. B., BALE, L. K., CLIFTON, K. B., GRELL, J. A., MADER, J. R., MASON, E. J. & POWELL, D. R. (2011). Pregnancy-associated plasma protein-A2 (PAPP-A2): tissue expression and biological consequences of gene knockout in mice. *Endocrinology,* 152**,** 2837-44.

COPLAND, J. A., PARDINI, A. W., WOOD, T. G., YIN, D., GREEN, A., BODENBURG, Y. H., URBAN, R. J. & STUART, C. A. (2007). IGF-1 controls GLUT3 expression in muscle via the transcriptional factor Sp1. *Biochimica et biophysica acta,* 1769**,** 631-40.

CORONA, E., DUDLEY, J. T. & BUTTE, A. J. (2010). Extreme evolutionary disparities seen in positive selection across seven complex diseases. *PloS one,* 5**,** e12236.

COWLES, M. K. & CARLIN, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association,* 91**,** 883-904.

CURI, R. A., OLIVEIRA, H. N., SILVEIRA, A. C. & LOPES, C. R. (2005). Effects of polymorphic microsatellites in the regulatory region of IGF1 and GHR on growth and carcass traits in beef cattle. *Animal genetics,* 36**,** 58-62.

CUTTER, A. D. & CHOI, J. Y. (2010). Natural selection shapes nucleotide polymorphism across the genome of the nematode Caenorhabditis briggsae. *Genome research,* 20**,** 1103-11.

CUVELIER, C., CLINQUART, A., HOCQUETTE, J. F., CABARAUX, J. F., DUFRASNE, I., ISTASSE, L. & HORNICK, J. L. (2006). Comparison of composition and quality traits of meat from young finishing bulls from Belgian Blue, Limousin and Aberdeen Angus breeds. *Meat science,* 74**,** 522-31.

D'ALESSANDRO, A. & ZOLLA, L. (2013). Meat science: From proteomics to integrated omics towards system biology. *Journal of proteomics,* 78**,** 558-77.

DAETWYLER, H. D., PONG-WONG, R., VILLANUEVA, B. & WOOLLIAMS, J. A. (2010). The impact of genetic architecture on genome-wide evaluation methods. *Genetics,* 185**,** 1021-31.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G. & DURBIN, R. (2011). The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-8.

DAVIES, K. E. & NOWAK, K. J. (2006). Molecular mechanisms of muscular dystrophies: old and new players. *Nature reviews. Molecular cell biology,* 7**,** 762-73.

DAYEM ULLAH, A. Z., LEMOINE, N. R. & CHELALA, C. (2012). SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic acids research,* 40**,** W65-70.

DE LOS CAMPOS, G., HICKEY, J. M., PONG-WONG, R., DAETWYLER, H. D. & CALUS, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics,* 193**,** 327-45.

DE ROOS, A. P., HAYES, B. J., SPELMAN, R. J. & GODDARD, M. E. (2008). Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. *Genetics,* 179**,** 1503-12.

DE ROOS, A. P., SCHROOTEN, C., VEERKAMP, R. F. & VAN ARENDONK, J. A. (2011). Effects of genomic selection on genetic improvement, inbreeding, and merit of young versus proven bulls. *Journal of dairy science,* 94**,** 1559-67.

DEKKERS, J. C. M. & HOSPITAL, F. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics,* 3**,** 22-32.

DELAVAUD, C., FERLAY, A., FAULCONNIER, Y., BOCQUIER, F., KANN, G. & CHILLIARD, Y. (2002). Plasma leptin concentration in adult cattle: effects of breed, adiposity, feeding level, and meal intake. *Journal of animal science,* 80**,** 1317-28.

DEPARTMENT OF AGRICULTURE FOOD AND THE MARINE (2012). AIM Bovine Statistics Report.

DHAHBI, J. M., KIM, H. J., MOTE, P. L., BEAVER, R. J. & SPINDLER, S. R. (2004). Temporal linkage between the phenotypic and genomic responses to caloric restriction. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 5524-9.

DODDS, M. G. & VICINI, P. (2004). Assessing convergence of Markov chain Monte Carlo simulations in hierarchical Bayesian models for population pharmacokinetics. *Annals of biomedical engineering,* 32**,** 1300-13.

DORAN, A. G. & CREEVEY, C. J. (2013). Snpdat: Easy and rapid annotation of results from de novo snp discovery projects for model and non-model organisms. *BMC bioinformatics,* 14**,** 45.

DOS REIS, M. & YANG, Z. (2011). Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. *Molecular biology and evolution,* 28**,** 2161-72.

DRENNAN, M. J. & MCGEE, M. (2009). Performance of spring-calving beef suckler cows and their progeny to slaughter on intensive and extensive grassland management systems. *Livestock science,* 120**,** 1-12.

DRUMMOND, A. J., SUCHARD, M. A., XIE, D. & RAMBAUT, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution,* 29**,** 1969-73.

DUAN, C., REN, H. & GAO, S. (2010). Insulin-like growth factors (IGFs), IGF receptors, and IGF-binding proteins: roles in skeletal muscle growth and differentiation. *General and comparative endocrinology,* 167**,** 344-51.

EDWARDS, C. J., MAGEE, D. A., PARK, S. D., MCGETTIGAN, P. A., LOHAN, A. J., MURPHY, A., FINLAY, E. K., SHAPIRO, B., CHAMBERLAIN, A. T., RICHARDS, M. B., BRADLEY, D. G., LOFTUS, B. J. & MACHUGH, D. E. (2010). A complete mitochondrial genome sequence from a mesolithic wild aurochs (Bos primigenius). *PloS one,* 5**,** e9255.

EGGEN, A. (2012). The development and application of genomic selection as a new breeding paradigm. *Animal Frontiers*, 2**,** 10-15.

EHRENBORG, E. & KROOK, A. (2009). Regulation of skeletal muscle physiology and metabolism by peroxisome proliferator-activated receptor delta. *Pharmacological reviews,* 61**,** 373-93.

ELSIK, C. G., TELLAM, R. L., WORLEY, K. C., GIBBS, R. A., MUZNY, D. M., WEINSTOCK, G. M., ADELSON, D. L., EICHLER, E. E., ELNITSKI, L., GUIGO, R., HAMERNIK, D. L., KAPPES, S. M., LEWIN, H. A., LYNN, D. J., NICHOLAS, F. W., REYMOND, A., RIJNKELS, M., SKOW, L. C., ZDOBNOV, E. M., SCHOOK, L., WOMACK, J., ALIOTO, T., ANTONARAKIS, S. E., ASTASHYN, A., CHAPPLE, C. E., CHEN, H. C., CHRAST, J., CAMARA, F., ERMOLAEVA, O., HENRICHSEN, C. N., HLAVINA, W., KAPUSTIN, Y., KIRYUTIN, B., KITTS, P., KOKOCINSKI, F., LANDRUM, M., MAGLOTT, D., PRUITT, K., SAPOJNIKOV, V., SEARLE, S. M., SOLOVYEV, V., SOUVOROV, A., UCLA, C., WYSS, C., ANZOLA, J. M., GERLACH, D., ELHAIK, E., GRAUR, D., REESE, J. T., EDGAR, R. C., MCEWAN, J. C., PAYNE, G. M., RAISON, J. M., JUNIER, T., KRIVENTSEVA, E. V., EYRAS, E., PLASS, M., DONTHU, R., LARKIN, D. M., REECY, J., YANG, M. Q., CHEN, L., CHENG, Z., CHITKO-MCKOWN, C. G., LIU, G. E., MATUKUMALLI, L. K., SONG, J., ZHU, B., BRADLEY, D. G., BRINKMAN, F. S., LAU, L. P., WHITESIDE, M. D., WALKER, A., WHEELER, T. T., CASEY, T., GERMAN, J. B., LEMAY, D. G., MAQBOOL, N. J., MOLENAAR, A. J., SEO, S., STOTHARD, P., BALDWIN, C. L., BAXTER, R., BRINKMEYER-LANGFORD, C. L., BROWN, W. C., CHILDERS, C. P., CONNELLEY, T., ELLIS, S. A., FRITZ, K., GLASS, E. J., HERZIG, C. T., IIVANAINEN, A., LAHMERS, K. K., BENNETT, A. K., DICKENS, C. M., GILBERT, J. G., HAGEN, D. E., SALIH, H., AERTS, J., CAETANO, A. R., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science,* 324**,** 522-8.

ENNS, R. M. & NICOLL, G. B. (2008). Genetic change results from selection on an economic breeding objective in beef cattle. *Journal of animal science,* 86**,** 3348-3357.

ESMAILIZADEH, A. K., BOTTEMA, C. D., SELLICK, G. S., VERBYLA, A. P., MORRIS, C. A., CULLEN, N. G. & PITCHFORD, W. S. (2008). Effects of the myostatin F94L substitution on beef traits. *Journal of animal science,* 86**,** 1038-46.

ETHERTON, T. D. (2004). Somatotropic function: the somatomedin hypothesis revisited. *Journal of animal science,* 82 E-Suppl**,** E239-244.

EVERSHED, R. P., PAYNE, S., SHERRATT, A. G., COPLEY, M. S., COOLIDGE, J., UREM-KOTSU, D., KOTSAKIS, K., OZDOGAN, M., OZDOGAN, A. E., NIEUWENHUYSE, O., AKKERMANS, P. M. M. G., BAILEY, D., ANDEESCU, R. R., CAMPBELL, S., FARID, S., HODDER, I., YALMAN, N., OZBASARAN, M., BICAKCI, E., GARFINKEL, Y., LEVY, T. & BURTON, M. M. (2008). Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature,* 455**,** 528-531.

FAN, B., ONTERU, S. K., DU, Z. Q., GARRICK, D. J., STALDER, K. J. & ROTHSCHILD, M. F. (2011). Genome-wide association study identifies Loci for body composition and structural soundness traits in pigs. *PloS one,* 6**,** e14726.

FAN, B., ONTERU, S. K. & ROTHSCHILD, M. F. (2009). The GGT1 and IGFBP5 genes are associated with fat deposition traits in the pig. *Archiv Fur Tierzucht-Archives of Animal Breeding,* 52**,** 337-339.

FARNIR, F., COPPIETERS, W., ARRANZ, J. J., BERZI, P., CAMBISANO, N., GRISART, B., KARIM, L., MARCQ, F., MOREAU, L., MNI, M., NEZER, C., SIMON, P., VANMANSHOVEN, P., WAGENAAR, D. & GEORGES, M. (2000). Extensive genome-wide linkage disequilibrium in cattle. *Genome research,* 10**,** 220-7.

FASTQC. *A quality control tool for high throughput sequence data.* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

FEDER, A. F., PETROV, D. A. & BERGLAND, A. O. (2012). LDx: estimation of linkage disequilibrium from high-throughput pooled resequencing data. *PloS one,* 7**,** e48588.

FENG, F., SALES, A. P. & KEPLER, T. B. (2011). A Bayesian approach for estimating calibration curves and unknown concentrations in immunoassays. *Bioinformatics,* 27**,** 707-12.

FIEMS, L. O., VANACKER, J. M., DE BOEVER, J. L., VAN CAELENBERGH, W., AERTS, J. M. & DE BRABANDER, D. L. (2007). Effect of energy restriction and re-alimentation in Belgian Blue double-muscled beef cows on digestibility and metabolites. *Journal of Animal Physiology and Animal Nutrition,* 91**,** 54-61.

FINNERAN, E., CROSSON, P., O'KIELY, P., SHALLOO, L., FORRISTAL, D. & WALLACE, M. (2010). Simulation Modelling of the Cost of Producing and Utilising Feeds for Ruminants on Irish Farms *Journal of Farm Management,* 14**,** 95-116.

FLICEK, P., AMODE, M. R., BARRELL, D., BEAL, K., BRENT, S., CARVALHO-SILVA, D., CLAPHAM, P., COATES, G., FAIRLEY, S., FITZGERALD, S., GIL, L., GORDON, L., HENDRIX, M., HOURLIER, T., JOHNSON, N., KAHARI, A. K., KEEFE, D., KEENAN, S., KINSELLA, R., KOMOROWSKA, M., KOSCIELNY, G., KULESHA, E., LARSSON, P., LONGDEN, I., MCLAREN, W., MUFFATO, M., OVERDUIN, B., PIGNATELLI, M., PRITCHARD, B., RIAT, H. S., RITCHIE, G. R., RUFFIER, M., SCHUSTER, M., SOBRAL, D., TANG, Y. A., TAYLOR, K., TREVANION, S., VANDROVCOVA, J., WHITE, S., WILSON, M., WILDER, S. P., AKEN, B. L., BIRNEY, E., CUNNINGHAM, F., DUNHAM, I., DURBIN, R., FERNANDEZ-SUAREZ, X. M., HARROW, J., HERRERO, J., HUBBARD, T. J., PARKER, A., PROCTOR, G., SPUDICH, G., VOGEL, J., YATES, A., ZADISSA, A. & SEARLE, S. M. (2012). Ensembl 2012. *Nucleic acids research,* 40**,** D84-90.

FOLEY, C., CHAPWANYA, A., CREEVEY, C. J., NARCIANDI, F., MORRIS, D., KENNY, E. M., CORMICAN, P., CALLANAN, J. J., O'FARRELLY, C. & MEADE, K. G. (2012). Global endometrial transcriptomic profiling: transient immune activation precedes tissue proliferation and repair in healthy beef cows. *BMC genomics,* 13**,** 489.

FONTANESI, L., SCHIAVO, G., GALIMBERTI, G., CALO, D. G., SCOTTI, E., MARTELLI, P. L., BUTTAZZONI, L., CASADIO, R. & RUSSO, V. (2012). A genome wide association study for backfat thickness in Italian Large White pigs highlights new regions affecting fat deposition including neuronal genes. *BMC genomics,* 13**,** 583.

FRANCESCHINI, A., SZKLARCZYK, D., FRANKILD, S., KUHN, M., SIMONOVIC, M., ROTH, A., LIN, J., MINGUEZ, P., BORK, P., VON MERING, C. & JENSEN, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research,* 41**,** D808-15.

FUJIKAWA, K., TAKAHASHI, A., NISHIMURA, A., ITOH, M., TAKANO-SHIMIZU, T. & OZAKI, M. (2009). Characteristics of genes up-regulated and down-regulated after 24 h starvation in the head of Drosophila. *Gene,* 446**,** 11-7.

FURUHASHI, M. & HOTAMISLIGIL, G. S. (2008). Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets. *Nature reviews. Drug discovery,* 7**,** 489-503.

GAO, Q., YUE, G. D., LI, W. Q., WANG, J. Y., XU, J. H. & YIN, Y. (2012). Recent Progress Using High-throughput Sequencing Technologies in Plant Molecular Breeding. *Journal of integrative plant biology,* 54**,** 215-227.

GAO, X., BECKER, L. C., BECKER, D. M., STARMER, J. D. & PROVINCE, M. A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic epidemiology,* 34**,** 100-5.

GARRICK, D. J. & GOLDEN, B. L. (2009). Producing and using genetic evaluations in the United States beef industry of today. *Journal of animal science,* 87**,** E11-E18.

GASSMANN, R., CARVALHO, A., HENZING, A. J., RUCHAUD, S., HUDSON, D. F., HONDA, R., NIGG, E. A., GERLOFF, D. L. & EARNSHAW, W. C. (2004). Borealin: a novel chromosomal passenger required for stability of the bipolar mitotic spindle. *The Journal of cell biology,* 166**,** 179-91.

GE, X., ZHANG, Y. & JIANG, H. (2013). Signaling pathways mediating the effects of insulin-like growth factor-I in bovine muscle satellite cells. *Molecular and cellular endocrinology,* 372**,** 23-9.

GEARY, T. W., MCFADIN, E. L., MACNEIL, M. D., GRINGS, E. E., SHORT, R. E., FUNSTON, R. N. & KEISLER, D. H. (2003). Leptin as a predictor of carcass composition in beef cattle. *Journal of animal science,* 81**,** 1-8.

GELMAN, A. & RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *statistical Science,* 7**,** 457-472.

GEYER, C. J. (1992). Practical Markov Chain Monte Carlo. *Statistical Science,* 7**,** 473-483.

GIANOLA, D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. *Genetics,* 194**,** 573-96.

GIANOLA, D., DE LOS CAMPOS, G., HILL, W. G., MANFREDI, E. & FERNANDO, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics,* 183**,** 347-63.

GIBBS, R. A., TAYLOR, J. F., VAN TASSELL, C. P., BARENDSE, W., EVERSOLE, K. A., GILL, C. A., GREEN, R. D., HAMERNIK, D. L., KAPPES, S. M., LIEN, S., MATUKUMALLI, L. K., MCEWAN, J. C., NAZARETH, L. V., SCHNABEL, R. D., WEINSTOCK, G. M., WHEELER, D. A., AJMONE-MARSAN, P., BOETTCHER, P. J., CAETANO, A. R., GARCIA, J. F., HANOTTE, O., MARIANI, P., SKOW, L. C., SONSTEGARD, T. S., WILLIAMS, J. L., DIALLO, B., HAILEMARIAM, L., MARTINEZ, M. L., MORRIS, C. A., SILVA, L. O., SPELMAN, R. J., MULATU, W., ZHAO, K., ABBEY, C. A., AGABA, M., ARAUJO, F. R., BUNCH, R. J., BURTON, J., GORNI, C., OLIVIER, H., HARRISON, B. E., LUFF, B., MACHADO, M. A., MWAKAYA, J., PLASTOW, G., SIM, W., SMITH, T., THOMAS, M. B., VALENTINI, A., WILLIAMS, P., WOMACK, J., WOOLLIAMS, J. A., LIU, Y., QIN, X., WORLEY, K. C., GAO, C., JIANG, H., MOORE, S. S., REN, Y., SONG, X. Z., BUSTAMANTE, C. D., HERNANDEZ, R. D., MUZNY, D. M., PATIL, S., SAN LUCAS, A., FU, Q., KENT, M. P., VEGA, R., MATUKUMALLI, A., MCWILLIAM, S., SCLEP, G., BRYC, K., CHOI, J., GAO, H., GREFENSTETTE, J. J., MURDOCH, B., STELLA, A., VILLA-ANGULO, R., WRIGHT, M., AERTS, J., JANN, O., NEGRINI, R., GODDARD, M. E., HAYES, B. J., BRADLEY, D. G., BARBOSA DA SILVA, M., LAU, L. P., LIU, G. E., LYNN, D. J., PANZITTA, F. & DODDS, K. G. (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science*, 324**,** 528-32.

GILL, J. (2008). Is partial-dimension convergence a problem for inferences from MCMC algorithms? *Political Analysis*, 16**,** 153-178.

GILMOUR, A. R., CULLIS, B. R., GOGEL, B. J., WELHAM, S. J. & THOMPSON, R. (2009a). ASReml User Guide 3.0. UK: VSN International Ltd, Hemel Hempstead, HP1 1ES.

GILMOUR, A. R., CULLIS, B. R., GOGEL, B. J., WELHAM, S. J. & THOMPSON, R. (2009b). ASReml User Guide Release 3.0. UK: VSN International Ltd, Hemel Hempstead, HP1 1ES.

GLASS, D. J. (2003). Molecular mechanisms modulating muscle mass. *Trends in Molecular Medicine,* 9**,** 344-50.

GLASS, D. J. (2010). Signaling pathways perturbing muscle mass. *Current opinion in clinical nutrition and metabolic care,* 13**,** 225-9.

GODDARD, M. E. & HAYES, B. J. (2007). Genomic selection. *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie,* 124**,** 323-30.

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**,** 381-91.

GOLDEN, B. L., GARRICK, D. J. & BENYSHEK, L. L. (2009). Milestones in beef cattle genetic evaluation. *Journal of animal science,* 87**,** E3-E10.

GOODSWEN, S. J., GONDRO, C., WATSON-HAIGH, N. S. & KADARMIDEEN, H. N. (2010). FunctSNP: an R package to link SNPs to functional knowledge and dbAutoMaker: a suite of Perl scripts to build SNP databases. *BMC bioinformatics,* 11**,** 311.

GROBET, L., MARTIN, L. J., PONCELET, D., PIROTTIN, D., BROUWERS, B., RIQUET, J., SCHOEBERLEIN, A., DUNNER, S., MENISSIER, F., MASSABANDA, J., FRIES, R., HANSET, R. & GEORGES, M. (1997). A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nature genetics,* 17**,** 71-4.

GUTIERREZ-GIL, B., WILLIAMS, J. L., HOMER, D., BURTON, D., HALEY, C. S. & WIENER, P. (2009). Search for quantitative trait loci affecting growth and carcass traits in a cross population of beef and dairy cattle. *Journal of animal science,* 87**,** 24-36.

HAN, H. Q., ZHOU, X., MITCH, W. E. & GOLDBERG, A. L. (2013). Myostatin/activin pathway antagonism: molecular basis and therapeutic potential. *The international journal of biochemistry & cell biology,* 45**,** 2333-47.

HARRIS, B. & JOHNSON, D. (1998). Approximate reliability of genetic evaluations under an animal model. *Journal of dairy science,* 81**,** 2723-8.

HAYES, B. J., CHAMBERLAIN, A. J., MACEACHERN, S., SAVIN, K., MCPARTLAN, H., MACLEOD, I., SETHURAMAN, L. & GODDARD, M. E. (2009). A genome map of divergent artificial selection between Bos taurus dairy cattle and Bos taurus beef cattle. *Animal genetics,* 40**,** 176-84.

HELMBERG, W. (2012). Bioinformatic databases and resources in the public domain to aid HLA research. *Tissue antigens,* 80**,** 295-304.

HERINGSTAD, B., CHANG, Y. M., GIANOLA, D. & KLEMETSDAL, G. (2003). Genetic analysis of longitudinal trajectory of clinical mastitis in first-lactation Norwegian cattle. *Journal of dairy science,* 86**,** 2676-2683.

HERRINGTON, J. & CARTER-SU, C. (2001). Signaling pathways activated by the growth hormone receptor. *Trends in endocrinology and metabolism: TEM,* 12**,** 252-7.

HERTZEL, A. V. & BERNLOHR, D. A. (2000). The mammalian fatty acid-binding protein multigene family: molecular and genetic insights into function. *Trends in endocrinology and metabolism: TEM,* 11**,** 175-80.

HINDORFF, L. A., MACARTHER, J., MORALES, J., JUNKINS, H. A., HALL, P. N., KLEMM, A. H. & MANOLIO, T. A. 2013. *A Catalog of*

Published Genome-Wide Association Studies [Online]. Available: http://www.genome.gov/gwastudies/.

HOLMANS, P., GREEN, E. K., PAHWA, J. S., FERREIRA, M. A., PURCELL, S. M., SKLAR, P., OWEN, M. J., O'DONOVAN, M. C. & CRADDOCK, N. (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *American journal of human genetics,* 85**,** 13-24.

HOOD, L. (2003). Systems biology: integrating technology, biology, and computation. *Mechanisms of ageing and development*, 124**,** 9-16.

HOOPER, S. L., HOBBS, K. H. & THUMA, J. B. (2008). Invertebrate muscles: Thin and thick filament structure; molecular basis of contraction and its regulation, catch and asynchronous muscle. *Progress in Neurobiology,* 86**,** 72-127.

HOOPER, S. L. & THUMA, J. B. (2005). Invertebrate muscles: Muscle specific genes and proteins. *Physiological Reviews,* 85**,** 1001-1060.

HORNEF, N., OLBRICH, H., HORVATH, J., ZARIWALA, M. A., FLIEGAUF, M., LOGES, N. T., WILDHABER, J., NOONE, P. G., KENNEDY, M., ANTONARAKIS, S. E., BLOUIN, J. L., BARTOLONI, L., NUSSLEIN, T., AHRENS, P., GRIESE, M., KUHL, H., SUDBRAK, R., KNOWLES, M. R., REINHARDT, R. & OMRAN, H. (2006). DNAH5 mutations are a common cause of primary ciliary dyskinesia with outer dynein arm defects. *American journal of respiratory and critical care medicine,* 174**,** 120-6.

HORNICK, J. L., VAN EENAEME, C., GERARD, O., DUFRASNE, I. & ISTASSE, L. (2000). Mechanisms of reduced and compensatory growth. *Domestic animal endocrinology,* 19**,** 121-32.

HOSKINS, R. A., PHAN, A. C., NAEEMUDDIN, M., MAPA, F. A., RUDDY, D. A., RYAN, J. J., YOUNG, L. M., WELLS, T., KOPCZYNSKI, C. & ELLIS, M. C. (2001). Single nucleotide polymorphism markers for genetic mapping in Drosophila melanogaster. *Genome research,* 11**,** 1100-13.

HTSEQ. *Analysing high-throughput sequencing data with Python.* [Online]. Available: http://www-huber.embl.de/users/anders/HTSeq/.

HU, Z. L., DRACHEVA, S., JANG, W., MAGLOTT, D., BASTIAANSEN, J., ROTHSCHILD, M. F. & REECY, J. M. (2005). A QTL resource and comparison tool for pigs: PigQTLDB. *Mammalian genome : official journal of the International Mammalian Genome Society,* 16**,** 792-800.

HU, Z. L., FRITZ, E. R. & REECY, J. M. (2007). AnimalQTLdb: a livestock QTL database tool set for positional QTL information mining and beyond. *Nucleic acids research,* 35**,** D604-9.

HU, Z. L., PARK, C. A., WU, X. L. & REECY, J. M. (2013). Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic acids research,* 41**,** D871-9.

HUANG, M. T., MASON, J. C., BIRDSEY, G. M., AMSELLEM, V., GERWIN, N., HASKARD, D. O., RIDLEY, A. J. & RANDI, A. M. (2005). Endothelial intercellular adhesion molecule (ICAM)-2 regulates angiogenesis. *Blood,* 106**,** 1636-1643.

HUELSENBECK, J. P., LARGET, B., MILLER, R. E. & RONQUIST, F. (2002). Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic biology,* 51**,** 673-688.

HUXLEY, A. F. & NIEDERGERKE, R. (1954). Structural Changes in Muscle during Contraction - Interference Microscopy of Living Muscle Fibres. *Nature,* 173**,** 971-973.

HUXLEY, H. & HANSON, J. (1954). Changes in the Cross-Striations of Muscle during Contraction and Stretch and Their Structural Interpretation. *Nature,* 173**,** 973-976.

ICBF. *Irish Cattle Breeding Federation* [Online]. Available: www.icbf.com.

IHGSC (2004). Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-45.

ISHIGAMI, T., UZAWA, K., FUSHIMI, K., SAITO, K., KATO, Y., NAKASHIMA, D., HIGO, M., KOUZU, Y., BUKAWA, H., KAWATA, T., ITO, H. & TANZAWA, H. (2008). Inhibition of ICAM2 induces radiosensitisation in oral squamous cell carcinoma cells. *British journal of cancer,* 98**,** 1357-1365.

JENUTH, J. P. (2000). The NCBI. Publicly available tools and resources on the Web. *Methods in molecular biology,* 132**,** 301-12.

JIANG, H. & GE, X. (2013). Mechanism of growth hormone stimulation of skeletal muscle growth in cattle. *Journal of animal science*.

JIANG, H., WANG, Y., WU, M., GU, Z., FRANK, S. J. & TORRES-DIAZ, R. (2007). Growth hormone stimulates hepatic expression of bovine growth hormone receptor messenger ribonucleic acid through signal transducer and activator of transcription 5 activation of a major growth hormone receptor gene promoter. *Endocrinology,* 148**,** 3307-15.

JIANG, J., JIANG, L., ZHOU, B., FU, W., LIU, J. F. & ZHANG, Q. (2011). Snat: a SNP annotation tool for bovine by integrating various sources of genomic information. *BMC genetics,* 12**,** 85.

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. *PloS one,* 5**,** e13661.

JOBLING, M. (2010). Are compensatory growth and catch-up growth two sides of the same coin? *Aquaculture International,* 18**,** 501-510.

JOHANSEN, K. A. & OVERTURF, K. (2006). Alterations in expression of genes associated with muscle metabolism and growth during nutritional restriction and refeeding in rainbow trout. *Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology,* 144**,** 119-27.

JOHNSON, R. C., NELSON, G. W., TROYER, J. L., LAUTENBERGER, J. A., KESSING, B. D., WINKLER, C. A. & O'BRIEN, S. J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC genomics,* 11**,** 724.

KADARMIDEEN, H. N., THOMPSON, R., COFFEY, M. P. & KOSSAIBATI, M. A. (2003). Genetic parameters and evaluations from single- and multiple-trait analysis of dairy cow fertility and milk production. *Livestock Production Science,* 81**,** 183-195.

KADARMIDEEN, H. N., VON ROHR, P. & JANSS, L. L. (2006). From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. *Mammalian genome : official journal of the International Mammalian Genome Society,* 17**,** 548-64.

KAMBADUR, R., SHARMA, M., SMITH, T. P. & BASS, J. J. (1997). Mutations in myostatin (GDF8) in double-muscled Belgian Blue and Piedmontese cattle. *Genome research,* 7**,** 910-6.

KANEHISA, M., GOTO, S., SATO, Y., FURUMICHI, M. & TANABE, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research,* 40**,** D109-14.

KAROLCHIK, D., BARBER, G. P., CASPER, J., CLAWSON, H., CLINE, M. S., DIEKHANS, M., DRESZER, T. R., FUJITA, P. A., GURUVADOO, L., HAEUSSLER, M., HARTE, R. A., HEITNER, S., HINRICHS, A. S., LEARNED, K., LEE, B. T., LI, C. H., RANEY, B. J., RHEAD, B., ROSENBLOOM, K. R., SLOAN, C. A., SPEIR, M. L., ZWEIG, A. S., HAUSSLER, D., KUHN, R. M. & KENT, W. J. (2013). The UCSC Genome Browser database: 2014 update. *Nucleic acids research*.

KEADY, S. M., KENNY, D. A., KEANE, M. G. & WATERS, S. M. (2011). Effect of sire breed and genetic merit for carcass weight on the transcriptional regulation of the somatotropic axis in longissimus dorsi of crossbred steers. *Journal of animal science,* 89**,** 4007-16.

KEADY, S. M., KENNY, D. A., OHLENDIECK, K., DOYLE, S., KEANE, M. G. & WATERS, S. M. (2013). Proteomic profiling of bovine M. longissimus lumborum from Crossbred Aberdeen Angus and Belgian Blue sired steers varying in genetic merit for carcass weight. *Journal of animal science,* 91**,** 654-65.

KHATKAR, M. S., NICHOLAS, F. W., COLLINS, A. R., ZENGER, K. R., CAVANAGH, J. A., BARRIS, W., SCHNABEL, R. D., TAYLOR, J. F. & RAADSMA, H. W. (2008). Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. *BMC genomics,* 9**,** 187.

KIJAS, J. M., WALES, R., TORNSTEN, A., CHARDON, P., MOLLER, M. & ANDERSSON, L. (1998). Melanocortin receptor 1 (MC1R) mutations and coat color in pigs. *Genetics,* 150**,** 1177-85.

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**,** R36.

KIM, J. J., FARNIR, F., SAVELL, J. & TAYLOR, J. F. (2003). Detection of quantitative trait loci for growth and beef carcass fatness traits in a cross between Bos taurus (Angus) and Bos indicus (Brahman) cattle. *Journal of animal science,* 81**,** 1933-42.

KIM, Y., RYU, J., WOO, J., KIM, J. B., KIM, C. Y. & LEE, C. (2011). Genome-wide association study reveals five nucleotide sequence variants for carcass traits in beef cattle. *Animal genetics,* 42**,** 361-5.

KITANO, H. (2002). Systems biology: a brief overview. *Science,* 295**,** 1662-4.

KLEIN, R. J., ZEISS, C., CHEW, E. Y., TSAI, J. Y., SACKLER, R. S., HAYNES, C., HENNING, A. K., SANGIOVANNI, J. P., MANE, S. M., MAYNE, S. T., BRACKEN, M. B., FERRIS, F. L., OTT, J., BARNSTABLE, C. & HOH, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science,* 308**,** 385-9.

KLUNGLAND, H., VAGE, D. I., GOMEZ-RAYA, L., ADALSTEINSSON, S. & LIEN, S. (1995). The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. *Mammalian genome : official journal of the International Mammalian Genome Society,* 6**,** 636-9.

KNOWLES, M. R., LEIGH, M. W., CARSON, J. L., DAVIS, S. D., DELL, S. D., FERKOL, T. W., OLIVIER, K. N., SAGEL, S. D., ROSENFELD, M., BURNS, K. A., MINNIX, S. L., ARMSTRONG, M. C., LORI, A., HAZUCHA, M. J., LOGES, N. T., OLBRICH, H., BECKER-HECK, A., SCHMIDTS, M., WERNER, C., OMRAN, H. & ZARIWALA, M. A. (2012). Mutations of DNAH11 in patients with primary ciliary dyskinesia with normal ciliary ultrastructure. *Thorax,* 67**,** 433-41.

KNURR, T., LAARA, E. & SILLANPAA, M. J. (2013). Impact of prior specifications in ashrinkage-inducing Bayesian model for quantitative trait mapping and genomic prediction. *Genetics, selection, evolution : GSE,* 45**,** 24.

KO, Y., ZHAI, C. & RODRIGUEZ-ZAS, S. (2009). Inference of gene pathways using mixture Bayesian networks. *BMC systems biology,* 3**,** 54.

KOLKMAN, I., OPSOMER, G., AERTS, S., HOFLACK, G., LAEVENS, H. & LIPS, D. (2010). Analysis of body measurements of newborn purebred Belgian Blue calves. *Animal : an international journal of animal bioscience,* 4**,** 661-71.

KOLLIAS, H. D. & MCDERMOTT, J. C. (2008). Transforming growth factor-beta and myostatin signaling in skeletal muscle. *Journal of applied physiology,* 104**,** 579-587.

KONIG, I. R. (2011). Validation in genetic association studies. *Briefings in bioinformatics,* 12**,** 253-8.

KOPCHICK, J. J. & ANDRY, J. M. (2000). Growth hormone (GH), GH receptor, and signal transduction. *Molecular genetics and metabolism,* 71**,** 293-314.

KORTE, A. & FARLOW, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods,* 9**,** 29.

KRUGLYAK, L. & NICKERSON, D. A. (2001). Variation is the spice of life. *Nature genetics,* 27**,** 234-6.

KUBER, P. S., BUSBOOM, J. R., DUCKETT, S. K., MIR, P. S., MIR, Z., MCCORMICK, R. J., GASKINS, C. T., CRONRATH, J. D., MARKS, D. J. & REEVES, J. J. (2004). Effects of biological type and dietary fat treatment on factors associated with tenderness: II. Measurements on beef semitendinosus muscle. *Journal of animal science,* 82**,** 779-84.

LAMETSCH, R., KRISTENSEN, L., LARSEN, M. R., THERKILDSEN, M., OKSBJERG, N. & ERTBJERG, P. (2006). Changes in the muscle proteome after compensatory growth in pigs. *Journal of animal science,* 84**,** 918-24.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, N., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LANGMEAD, B. & SALZBERG, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods,* 9**,** 357-9.

LANGO ALLEN, H., ESTRADA, K., LETTRE, G., BERNDT, S. I., WEEDON, M. N., RIVADENEIRA, F., WILLER, C. J., JACKSON, A. U., VEDANTAM, S., RAYCHAUDHURI, S., FERREIRA, T., WOOD, A. R., WEYANT, R. J., SEGRE, A. V., SPELIOTES, E. K., WHEELER, E., SORANZO, N., PARK, J. H., YANG, J., GUDBJARTSSON, D., HEARD-COSTA, N. L., RANDALL, J. C., QI, L., VERNON SMITH, A., MAGI, R., PASTINEN, T., LIANG, L., HEID, I. M., LUAN, J., THORLEIFSSON, G., WINKLER, T. W., GODDARD, M. E., SIN LO, K., PALMER, C., WORKALEMAHU, T., AULCHENKO, Y. S., JOHANSSON, A., ZILLIKENS, M. C., FEITOSA, M. F., ESKO, T., JOHNSON, T., KETKAR, S., KRAFT, P., MANGINO, M., PROKOPENKO, I., ABSHER, D., ALBRECHT, E., ERNST, F., GLAZER, N. L., HAYWARD, C., HOTTENGA, J. J., JACOBS, K. B., KNOWLES, J. W., KUTALIK, Z., MONDA, K. L., POLASEK, O., PREUSS, M., RAYNER, N. W., ROBERTSON, N. R., STEINTHORSDOTTIR, V., TYRER, J. P., VOIGHT, B. F., WIKLUND, F., XU, J., ZHAO, J. H., NYHOLT, D. R., PELLIKKA, N., PEROLA, M., PERRY, J. R., SURAKKA, I., TAMMESOO, M. L., ALTMAIER, E. L., AMIN, N., ASPELUND, T., BHANGALE, T., BOUCHER, G., CHASMAN, D. I., CHEN, C., COIN, L., COOPER, M.

N., DIXON, A. L., GIBSON, Q., GRUNDBERG, E., HAO, K., JUHANI JUNTTILA, M., KAPLAN, L. M., KETTUNEN, J., KONIG, I. R., KWAN, T., LAWRENCE, R. W., LEVINSON, D. F., LORENTZON, M., MCKNIGHT, B., MORRIS, A. P., MULLER, M., SUH NGWA, J., PURCELL, S., RAFELT, S., SALEM, R. M., SALVI, E., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature,* 467**,** 832-8.

LAOUKILI, J., KOOISTRA, M. R., BRAS, A., KAUW, J., KERKHOVEN, R. M., MORRISON, A., CLEVERS, H. & MEDEMA, R. H. (2005). FoxM1 is required for execution of the mitotic programme and chromosome stability. *Nature cell biology,* 7**,** 126-36.

LEE, S. H., VAN DER WERF, J. H., HAYES, B. J., GODDARD, M. E. & VISSCHER, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS genetics,* 4**,** e1000231.

LEE, S. J. (2004). Regulation of muscle mass by myostatin. *Annual Review of Cell and Developmental Biology,* 20**,** 61-86.

LEE, S. J. & MCPHERRON, A. C. (2001). Regulation of myostatin activity and muscle growth. *Proceedings of the National Academy of Sciences of the United States of America,* 98**,** 9306-11.

LEHNERT, S. A., BYRNE, K. A., REVERTER, A., NATTRASS, G. S., GREENWOOD, P. L., WANG, Y. H., HUDSON, N. J. & HARPER, G. S. (2006). Gene expression profiling of bovine skeletal muscle in response to and during recovery from chronic and severe undernutrition. *Journal of animal science,* 84**,** 3239-50.

LEMAY, D. G., LYNN, D. J., MARTIN, W. F., NEVILLE, M. C., CASEY, T. M., RINCON, G., KRIVENTSEVA, E. V., BARRIS, W. C., HINRICHS, A. S., MOLENAAR, A. J., POLLARD, K. S., MAQBOOL, N. J., SINGH, K., MURNEY, R., ZDOBNOV, E. M., TELLAM, R. L., MEDRANO, J. F., GERMAN, J. B. & RIJNKELS, M. (2009). The bovine lactation genome: insights into the evolution of mammalian milk. *Genome biology,* 10**,** R43.

LEWIS, K. A., GRAY, P. C., BLOUNT, A. L., MACCONNELL, L. A., WIATER, E., BILEZIKJIAN, L. M. & VALE, W. (2000). Betaglycan binds inhibin and can mediate functional antagonism of activin signalling. *Nature,* 404**,** 411-414.

LI, H. & DURBIN, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, S., MA, L., LI, H., VANG, S., HU, Y., BOLUND, L. & WANG, J. (2007). Snap: an integrated SNP annotation platform. *Nucleic acids research,* 35**,** D707-10.

LIU, C. K., CHEN, Y. H., TANG, C. Y., CHANG, S. C., LIN, Y. J., TSAI, M. F., CHEN, Y. T. & YAO, A. (2008). Functional analysis of novel SNPs and mutations in human and mouse genomes. *BMC bioinformatics,* 9 Suppl 12**,** S10.

LIU, Y., QIN, X., SONG, X. Z. H., JIANG, H. Y., SHEN, Y. F., DURBIN, K. J., LIEN, S., KENT, M. P., SODELAND, M., REN, Y. R., ZHANG, L., SODERGREN, E., HAVLAK, P., WORLEY, K. C., WEINSTOCK, G. M. & GIBBS, R. A. (2009). Bos taurus genome assembly. *BMC genomics,* 10.

LOFTUS, R. T., MACHUGH, D. E., BRADLEY, D. G., SHARP, P. M. & CUNNINGHAM, P. (1994). Evidence for two independent domestications of cattle. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 2757-61.

LUCY, M. C. (2008). Functional differences in the growth hormone and insulin-like growth factor axis in cattle and pigs: implications for post-partum nutrition and reproduction. *Reproduction in domestic animals = Zuchthygiene,* 43 Suppl 2**,** 31-9.

LUCY, M. C., VERKERK, G. A., WHYTE, B. E., MACDONALD, K. A., BURTON, L., CURSONS, R. T., ROCHE, J. R. & HOLMES, C. W. (2009). Somatotropic axis components and nutrient partitioning in genetically diverse dairy cows managed under different feed allowances in a pasture system. *Journal of dairy science,* 92**,** 526-39.

LUNA-NEVAREZ, P., RINCON, G., MEDRANO, J. F., RILEY, D. G., CHASE, C. C., JR., COLEMAN, S. W., VANLEEUWEN, D. M., DEATLEY, K. L., ISLAS-TREJO, A., SILVER, G. A. & THOMAS, M. G. (2011). Single nucleotide polymorphisms in the growth hormone-insulin-like growth factor axis in straightbred and crossbred Angus, Brahman, and Romosinuano heifers: population genetic analyses and association of genotypes with reproductive phenotypes. *Journal of animal science,* 89**,** 926-34.

LUNA-PINTO, G. & CRONJE, P. B. (2000). The roles of the insulin-like growth factor system and leptin as possible mediators of the effects of nutritional restriction on age at puberty and compensatory growth in dairy heifers. *South African Journal of Animal Science-Suid-Afrikaanse Tydskrif Vir Veekunde,* 30**,** 155-163.

LYNN, D. J., WINSOR, G. L., CHAN, C., RICHARD, N., LAIRD, M. R., BARSKY, A., GARDY, J. L., ROCHE, F. M., CHAN, T. H., SHAH, N., LO, R., NASEER, M., QUE, J., YAU, M., ACAB, M., TULPAN, D., WHITESIDE, M. D., CHIKATAMARLA, A., MAH, B., MUNZNER, T., HOKAMP, K., HANCOCK, R. E. & BRINKMAN, F. S. (2008). InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Molecular systems biology,* 4**,** 218.

MACCALLUM, C. & HILL, E. (2006). Being positive about selection. *PLoS biology,* 4**,** 293-295.

MACEACHERN, S., HAYES, B., MCEWAN, J. & GODDARD, M. (2009a). An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in Domestic cattle. *BMC genomics,* 10**,** 181.

MACEACHERN, S., MCEWAN, J., MATHER, A., MCCULLOCH, A., SUNNUCKS, P. & GODDARD, M. (2006). Testing the neutral theory of molecular evolution using genomic data: a comparison of the human and bovine transcriptome. *Genetics, selection, evolution : GSE,* 38**,** 321-41.

MACEACHERN, S., MCEWAN, J., MCCULLOCH, A., MATHER, A., SAVIN, K. & GODDARD, M. (2009b). Molecular evolution of the Bovini tribe (Bovidae, Bovinae): is there evidence of rapid evolution or reduced selective constraint in Domestic cattle? *BMC genomics,* 10**,** 179.

MACHADO, M. B. B., ALENCAR, M. M., PEREIRA, A. P., OLIVEIRA, H. N., CASAS, E., COUTINHO, L. L. & REGITANO, L. C. A. (2003). QTL affecting body weight in a candidate region of cattle chromosome 5. *Genetics and Molecular Biology,* 26**,** 259-265.

MACNEIL, M. D. & GROSZ, M. D. (2002). Genome-wide scans for QTL affecting carcass traits in Hereford x composite double backcross populations. *Journal of animal science,* 80**,** 2316-24.

MANGEL, M. & MUNCH, S. B. (2005). A life-history perspective on short- and long-term consequences of compensatory growth. *The American naturalist,* 166**,** E155-76.

MANOLIO, T. A. (2013). Published GWA Reports, 2005-2012.

MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., CHO, J. H., GUTTMACHER, A. E., KONG, A., KRUGLYAK, L., MARDIS, E., ROTIMI, C. N., SLATKIN, M., VALLE, D., WHITTEMORE, A. S., BOEHNKE, M., CLARK, A. G., EICHLER, E. E., GIBSON, G., HAINES, J. L., MACKAY, T. F., MCCARROLL, S. A. & VISSCHER, P. M. (2009). Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.

MAO, F., CHEN, L., VINSKY, M., OKINE, E., WANG, Z., BASARAB, J., CREWS, D. H., JR. & LI, C. (2013). Phenotypic and genetic relationships of feed efficiency with growth performance, ultrasound, and carcass merit traits in Angus and Charolais steers. *Journal of animal science,* 91**,** 2067-76.

MARDIS, E. R. (2006). Anticipating the 1,000 dollar genome. *Genome biology,* 7**,** 112.

MARDIS, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG,* 24**,** 133-41.

MARJORAM, P. & TAVARE, S. (2006). Modern computational approaches for analysing molecular genetic variation data. *Nature reviews. Genetics,* 7**,** 759-70.

MARKLUND, L., MOLLER, M. J., SANDBERG, K. & ANDERSSON, L. (1996). A missense mutation in the gene for melanocyte-stimulating hormone receptor (MC1R) is associated with the chestnut coat color in horses. *Mammalian genome : official journal of the International Mammalian Genome Society,* 7**,** 895-9.

MARTINEZ, A., ALDAI, N., CELAYA, R. & OSORO, K. (2010). Effect of breed body size and the muscular hypertrophy gene in the production and carcass traits of concentrate-finished yearling bulls. *Journal of animal science,* 88**,** 1229-39.

MATUKUMALLI, L. K., LAWLEY, C. T., SCHNABEL, R. D., TAYLOR, J. F., ALLAN, M. F., HEATON, M. P., O'CONNELL, J., MOORE, S. S., SMITH, T. P., SONSTEGARD, T. S. & VAN TASSELL, C. P. (2009). Development and characterization of a high density SNP genotyping assay for cattle. *PloS one,* 4**,** e5350.

MCCABE, M., WATERS, S., MORRIS, D., KENNY, D., LYNN, D. & CREEVEY, C. (2012). RNA-seq analysis of differential gene expression in liver from lactating dairy cows divergent in negative energy balance. *BMC genomics,* 13**,** 193.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. & HIRSCHHORN, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews. Genetics,* 9**,** 356-69.

MCCLURE, M. C., MORSCI, N. S., SCHNABEL, R. D., KIM, J. W., YAO, P., ROLF, M. M., MCKAY, S. D., GREGG, S. J., CHAPPLE, R. H., NORTHCUTT, S. L. & TAYLOR, J. F. (2010). A genome scan for quantitative trait loci influencing carcass, post-natal growth and reproductive traits in commercial Angus cattle. *Animal genetics,* 41**,** 597-607.

MCDONALD, J. H. & KREITMAN, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature,* 351**,** 652-4.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research,* 20**,** 1297-303.

MCPHERRON, A. C., LAWLER, A. M. & LEE, S. J. (1997). Regulation of skeletal muscle mass in mice by a new TGF-beta superfamily member. *Nature,* 387**,** 83-90.

MCPHERRON, A. C. & LEE, S. J. (1997). Double muscling in cattle due to mutations in the myostatin gene. *Proceedings of the National Academy of Sciences of the United States of America,* 94**,** 12457-61.

MCRAE, A. F., MCEWAN, J. C., DODDS, K. G., WILSON, T., CRAWFORD, A. M. & SLATE, J. (2002). Linkage disequilibrium in domestic sheep. *Genetics,* 160**,** 1113-22.

MCTAVISH, E. J., DECKER, J. E., SCHNABEL, R. D., TAYLOR, J. F. & HILLIS, D. M. (2013). New World cattle show ancestry from multiple independent domestication events. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** E1398-406.

MEADOWS, J. R., CHAN, E. K. & KIJAS, J. W. (2008). Linkage disequilibrium compared between five populations of domestic sheep. *BMC genetics,* 9**,** 61.

MELHAM, T. (2013). Modelling, abstraction, and computation in systems biology: A view from computer science. *Progress in biophysics and molecular biology,* 111**,** 129-36.

MEREDITH, B. K., KEARNEY, F. J., FINLAY, E. K., BRADLEY, D. G., FAHEY, A. G., BERRY, D. P. & LYNN, D. J. (2012). Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. *BMC genetics,* 13**,** 21.

MESSER, P. W. & PETROV, D. A. (2013). Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 8615-20.

METZKER, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics,* 11**,** 31-46.

MEUWISSEN, T. H., HAYES, B. J. & GODDARD, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics,* 157**,** 1819-29.

MEYER, L. R., ZWEIG, A. S., HINRICHS, A. S., KAROLCHIK, D., KUHN, R. M., WONG, M., SLOAN, C. A., ROSENBLOOM, K. R., ROE, G., RHEAD, B., RANEY, B. J., POHL, A., MALLADI, V. S., LI, C. H., LEE, B. T., LEARNED, K., KIRKUP, V., HSU, F., HEITNER, S., HARTE, R. A., HAEUSSLER, M., GURUVADOO, L., GOLDMAN, M., GIARDINE, B. M., FUJITA, P. A., DRESZER, T. R., DIEKHANS, M., CLINE, M. S., CLAWSON, H., BARBER, G. P., HAUSSLER, D. & KENT, W. J. (2013). The UCSC Genome Browser database: extensions and updates 2013. *Nucleic acids research,* 41**,** D64-9.

MIGLIOR, F., MUIR, B. L. & VAN DOORMAAL, B. J. (2005). Selection indices in Holstein cattle of various countries. *Journal of dairy science,* 88**,** 1255-1263.

MORGULIS, A., COULOURIS, G., RAYTSELIS, Y., MADDEN, T. L., AGARWALA, R. & SCHAFFER, A. A. (2008). Database indexing for production MegaBLAST searches. *Bioinformatics,* 24**,** 1757-64.

MUERS, M. (2009). GENOMICS Milking the cow genome. *Nature Reviews Genetics,* 10**,** 346-346.

MULLEN, M. P., BERRY, D. P., HOWARD, D. J., DISKIN, M. G., LYNCH, C. O., BERKOWICZ, E. W., MAGEE, D. A., MACHUGH, D. E. & WATERS, S. M. (2010). Associations between novel single nucleotide polymorphisms in the Bos taurus growth hormone gene and performance traits in Holstein-Friesian dairy cattle. *Journal of dairy science,* 93**,** 5959-5969.

MULLEN, M. P., BERRY, D. P., HOWARD, D. J., DISKIN, M. G., LYNCH, C. O., BERKOWICZ, E. W., MAGEE, D. A., MACHUGH, D. E. & WATERS, S. M. (2011a). Associations between novel single nucleotide polymorphisms in the Bos taurus growth hormone gene and performance traits in Holstein-Friesian dairy cattle (vol 93, pg 5959, 2010). *Journal of dairy science,* 94**,** 1069-1069.

MULLEN, M. P., CREEVEY, C. J., BERRY, D. P., MCCABE, M. S., MAGEE, D. A., HOWARD, D. J., KILLEEN, A. P., PARK, S. D., MCGETTIGAN, P. A., LUCY, M. C., MACHUGH, D. E. & WATERS, S. M. (2012). Polymorphism discovery and allele frequency estimation using high-throughput DNA sequencing of target-enriched pooled DNA samples. *BMC genomics,* 13**,** 16.

MULLEN, M. P., LYNCH, C. O., WATERS, S. M., HOWARD, D. J., O'BOYLE, P., KENNY, D. A., BUCKLEY, F., HORAN, B. & DISKIN, M. G. (2011b). Single nucleotide polymorphisms in the growth hormone and insulin-like growth factor-1 genes are associated with milk production, body condition score and fertility traits in dairy cows. *Genetics and Molecular Research,* 10**,** 1819-1830.

MURIGNEUX, V., SAULIERE, J., CROLLIUS, H. R. & LE HIR, H. (2013). Transcriptome-wide identification of RNA binding sites by CLIP-seq. *Methods,* 63**,** 32-40.

MUSARO, A., MCCULLAGH, K., PAUL, A., HOUGHTON, L., DOBROWOLNY, G., MOLINARO, M., BARTON, E. R., SWEENEY, H. L. & ROSENTHAL, N. (2001). Localized Igf-1 transgene expression sustains hypertrophy and regeneration in senescent skeletal muscle. *Nature genetics,* 27**,** 195-200.

NEWCOMBE, P. J., RECK, B. H., SUN, J. L., PLATEK, G. T., VERZILLI, C., KADER, A. K., KIM, S. T., HSU, F. C., ZHANG, Z., ZHENG, S. L., MOOSER, V. E., CONDREAY, L. D., SPRAGGS, C. F., WHITTAKER, J. C., RITTMASTER, R. S. & XU, J. F. (2012). A Comparison of Bayesian and Frequentist Approaches to Incorporating External Information for the Prediction of Prostate Cancer Risk. *Genetic epidemiology,* 36**,** 71-83.

NISHIMURA, S., WATANABE, T., MIZOSHITA, K., TATSUDA, K., FUJITA, T., WATANABE, N., SUGIMOTO, Y. & 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**,** 40.

NSENGIMANA, J., BARET, P., HALEY, C. S. & VISSCHER, P. M. (2004). Linkage disequilibrium in the domesticated pig. *Genetics,* 166**,** 1395-404.

NYLANDER, J. A., RONQUIST, F., HUELSENBECK, J. P. & NIEVES-ALDREY, J. L. (2004). Bayesian phylogenetic analysis of combined data. *Systematic biology,* 53**,** 47-67.

NYLANDER, J. A., WILGENBUSCH, J. C., WARREN, D. L. & SWOFFORD, D. L. (2008). AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics,* 24**,** 581-3.

O'HARA, R. B., CANO, J. M., OVASKAINEN, O., TEPLITSKY, C. & ALHO, J. S. (2008). Bayesian approaches in evolutionary quantitative genetics. *Journal of evolutionary biology,* 21**,** 949-57.

O'LOUGHLIN, A., LYNN, D. J., MCGEE, M., DOYLE, S., MCCABE, M. & EARLEY, B. (2012). Transcriptomic analysis of the stress response to weaning at housing in bovine leukocytes using RNA-seq technology. *BMC genomics,* 13**,** 250.

O'NEILL, C. J., SWAIN, D. L. & KADARMIDEEN, H. N. (2010). Evolutionary process of Bos taurus cattle in favourable versus unfavourable environments and its implications for genetic selection. *Evolutionary Applications,* 3**,** 422-433.

OLBRICH, H., HAFFNER, K., KISPERT, A., VOLKEL, A., VOLZ, A., SASMAZ, G., REINHARDT, R., HENNIG, S., LEHRACH, H., KONIETZKO, N., ZARIWALA, M., NOONE, P. G., KNOWLES, M., MITCHISON, H. M., MEEKS, M., CHUNG, E. M., HILDEBRANDT, F., SUDBRAK, R. & OMRAN, H. (2002). Mutations in DNAH5 cause primary ciliary dyskinesia and randomization of left-right asymmetry. *Nature genetics,* 30**,** 143-4.

OLDMEADOW, C. & KEITH, J. M. (2011). Model selection in Bayesian segmentation of multiple DNA alignments. *Bioinformatics,* 27**,** 604-10.

OLSEN, H. G., HAYES, B. J., KENT, M. P., NOME, T., SVENDSEN, M., LARSGARD, A. G. & LIEN, S. (2011). Genome-wide association mapping in Norwegian Red cattle identifies quantitative trait loci for fertility and milk production on BTA12. *Animal genetics,* 42**,** 466-74.

OSHLACK, A. & WAKEFIELD, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology direct,* 4**,** 14.

OSZKIEWICZ, D., MUINONEN, K., VIRTANEN, J., GRANVIK, M. & BOWELL, E. (2012). Modeling collision probability for Earth-impactor 2008 TC3. *Planetary and Space Science,* 73**,** 30-38.

OTTO, A. & PATEL, K. (2010). Signalling and the control of skeletal muscle size. *Experimental cell research,* 316**,** 3059-66.

OVERGAARD, M. T., BOLDT, H. B., LAURSEN, L. S., SOTTRUP-JENSEN, L., CONOVER, C. A. & OXVIG, C. (2001a). Pregnancy-associated plasma protein-A2 (PAPP-A2), a novel insulin-like growth factor-binding protein-5 proteinase. *Journal of Biological Chemistry,* 276**,** 21849-21853.

OVERGAARD, M. T., BOLDT, H. B., LAURSEN, L. S., SOTTRUP-JENSEN, L., CONOVER, C. A. & OXVIG, C. (2001b). Pregnancy-associated plasma protein-A2 (PAPP-A2), a novel insulin-like growth factor-binding protein-5 proteinase. *The Journal of biological chemistry,* 276**,** 21849-53.

PABIOU, T., FIKSE, W. F., AMER, P. R., CROMIE, A. R., NASHOLM, A. & BERRY, D. P. (2012). Genetic relationships between carcass cut weights predicted from video image analysis and other performance traits in cattle. *Animal : an international journal of animal bioscience,* 6**,** 1389-97.

PARK, C. S. (2005). Role of compensatory mammary growth in epigenetic control of gene expression. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology,* 19**,** 1586-91.

PARSCH, J., ZHANG, Z. & BAINES, J. F. (2009). The Influence of Demography and Weak Selection on the McDonald-Kreitman Test: An Empirical Study in Drosophila. *Molecular biology and evolution,* 26**,** 691-698.

PEREZ, O. D., KINOSHITA, S., HITOSHI, Y., PAYAN, D. G., KITAMURA, T., NOLAN, G. P. & LORENS, J. B. (2002). Activation of the PKB/AKT pathway by ICAM-2. *Immunity,* 16**,** 51-65.

PICHA, M. E., TURANO, M. J., TIPSMARK, C. K. & BORSKI, R. J. (2008). Regulation of endocrine and paracrine sources of Igfs and Gh receptor during compensatory growth in hybrid striped bass (Morone chrysops X Morone saxatilis). *The Journal of endocrinology,* 199**,** 81-94.

PIFFERI, M., MICHELUCCI, A., CONIDI, M. E., CANGIOTTI, A. M., SIMI, P., MACCHIA, P. & BONER, A. L. (2010). New DNAH11 mutations in primary ciliary dyskinesia with normal axonemal ultrastructure. *The European respiratory journal,* 35**,** 1413-6.

PILECKA, I., WHATMORE, A., HOOFT VAN HUIJSDUIJNEN, R., DESTENAVES, B. & CLAYTON, P. (2007). Growth hormone signalling: sprouting links between pathways, human genetics and therapeutic options. *Trends in endocrinology and metabolism: TEM,* 18**,** 12-8.

PRITCHARD, J. K. & PRZEWORSKI, M. (2001). Linkage disequilibrium in humans: models and data. *American journal of human genetics,* 69**,** 1-14.

PRYCE, J. E., BOLORMAA, S., CHAMBERLAIN, A. J., BOWMAN, P. J., SAVIN, K., GODDARD, M. E. & HAYES, B. J. (2010). A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *Journal of dairy science,* 93**,** 3331-45.

PRYCE, J. E., HAYES, B. J., BOLORMAA, S. & GODDARD, M. E. (2011). Polymorphic regions affecting human height also control stature in cattle. *Genetics,* 187**,** 981-4.

PURFIELD, D. C., BRADLEY, D. G., KEARNEY, J. F. & BERRY, D. P. (2013). Genome-wide association study for calving traits in Holstein-Friesian dairy cattle. *Animal : an international journal of animal bioscience***,** 1-12.

QI, D. L., CHAO, Y., GUO, S. C., ZHAO, L. Y., LI, T. P., WEI, F. L. & ZHAO, X. Q. (2012). Convergent, Parallel and Correlated Evolution of Trophic Morphologies in the Subfamily Schizothoracinae from the Qinghai-Tibetan Plateau. *PloS one,* 7.

R DEVELOPMENT CORE TEAM (2011). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

RAMANAN, V. K., SHEN, L., MOORE, J. H. & SAYKIN, A. J. (2012). Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends in genetics : TIG,* 28**,** 323-32.

RAMEY, H. R., DECKER, J. E., MCKAY, S. D., ROLF, M. M., SCHNABEL, R. D. & TAYLOR, J. F. (2013). Detection of selective sweeps in cattle using genome-wide SNP data. *BMC genomics,* 14**,** 382.

RENAVILLE, R., HAMMADI, M. & PORTETELLE, D. (2002). Role of the somatotropic axis in the mammalian metabolism. *Domestic animal endocrinology,* 23**,** 351-60.

RIVA, A. & KOHANE, I. S. (2002). SNPper: retrieval and analysis of human SNPs. *Bioinformatics,* 18**,** 1681-5.

ROBERT, C. & CASELLA, G. (2011). A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data. *Statistical Science,* 26**,** 102-115.

ROBINSON-RECHAVI, M. & HUCHON, D. (2000). RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. *Bioinformatics,* 16**,** 296-7.

ROMMEL, C., BODINE, S. C., CLARKE, B. A., ROSSMAN, R., NUNEZ, L., STITT, T. N., YANCOPOULOS, G. D. & GLASS, D. J. (2001). Mediation of IGF-1-induced skeletal myotube hypertrophy by PI(3)K/Akt/mTOR and PI(3)K/Akt/GSK3 pathways. *Nature cell biology,* 3**,** 1009-1013.

ROSEN, E. D. (2006). New drugs from fat bugs? *Cell metabolism,* 3**,** 1-2.

ROTH, S. M. & WALSH, S. (2004). Myostatin: a therapeutic target for skeletal muscle wasting. *Current opinion in clinical nutrition and metabolic care,* 7**,** 259-63.

SAATCHI, M., GARRICK, D. J., TAIT, R. G., JR., MAYES, M. S., DREWNOSKI, M., SCHOONMAKER, J., DIAZ, C., BEITZ, D. C. & REECY, J. M. (2013). Genome-wide association and prediction of direct genomic breeding values for composition of fatty acids in Angus beef cattlea. *BMC genomics,* 14**,** 730.

SADKOWSKI, T., JANK, M., ZWIERZCHOWSKI, L., OPRZADEK, J. & MOTYL, T. (2009). Comparison of skeletal muscle transcriptional profiles in dairy and beef breeds bulls. *Journal of applied genetics*, 50**,** 109-23.

SAHU, A. (2004). Minireview: A hypothalamic role in energy balance with special emphasis on leptin. *Endocrinology,* 145**,** 2613-20.

SAINZ, N., RODRIGUEZ, A., CATALAN, V., BECERRIL, S., RAMIREZ, B., GOMEZ-AMBROSI, J. & FRUHBECK, G. (2009). Leptin administration favors muscle mass accretion by decreasing FoxO3a and increasing PGC-1alpha in ob/ob mice. *PloS one,* 4**,** e6808.

SALEM, M., SILVERSTEIN, J., REXROAD, C. E., 3RD & YAO, J. (2007). Effect of starvation on global gene expression and proteolysis in rainbow trout (Oncorhynchus mykiss). *BMC genomics,* 8**,** 328.

SANGER, F., AIR, G. M., BARRELL, B. G., BROWN, N. L., COULSON, A. R., FIDDES, C. A., HUTCHISON, C. A., SLOCOMBE, P. M. & SMITH, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature,* 265**,** 687-95.

SARTORI, R., BASTOS, M. R., BARUSELLI, P. S., GIMENES, L. U., ERENO, R. L. & BARROS, C. M. (2010). Physiological differences and implications to reproductive management of Bos taurus and Bos indicus cattle in a tropical environment. *Society of Reproduction and Fertility supplement,* 67**,** 357-75.

SARTORI, R., MILAN, G., PATRON, M., MAMMUCARI, C., BLAAUW, B., ABRAHAM, R. & SANDRI, M. (2009). Smad2 and 3 transcription factors control muscle mass in adulthood. *American journal of physiology. Cell physiology,* 296**,** C1248-57.

SATIJA, R., NOVAK, A., MIKLOS, I., LYNGSO, R. & HEIN, J. (2009). BigFoot: Bayesian alignment and phylogenetic footprinting with MCMC. *Bmc Evolutionary Biology,* 9.

SBONER, A., MU, X. J., GREENBAUM, D., AUERBACH, R. K. & GERSTEIN, M. B. (2011). The real cost of sequencing: higher than you think! *Genome biology,* 12**,** 125.

SCHATZ, M. C., DELCHER, A. L. & SALZBERG, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome research,* 20**,** 1165-73.

SCHIAFFINO, S. & MAMMUCARI, C. (2011). Regulation of skeletal muscle growth by the IGF1-Akt/PKB pathway: insights from genetic models. *Skeletal muscle,* 1**,** 4.

SCHMITT, A. O., SCHUCHHARDT, J., LUDWIG, A. & BROCKMANN, G. A. (2007). Protein evolution within and between species. *Journal of theoretical biology,* 249**,** 376-83.

SCHUELKE, M., WAGNER, K. R., STOLZ, L. E., HUBNER, C., RIEBEL, T., KOMEN, W., BRAUN, T., TOBIN, J. F. & LEE, S. J. (2004). Myostatin mutation associated with gross muscle hypertrophy in a child. *The New England journal of medicine,* 350**,** 2682-8.

SCHULZE, P. C., FANG, J., KASSIK, K. A., GANNON, J., CUPESI, M., MACGILLIVRAY, C., LEE, R. T. & ROSENTHAL, N. (2005). Transgenic overexpression of locally acting insulin-like growth factor-1 inhibits ubiquitin-mediated muscle atrophy in chronic left-ventricular dysfunction. *Circulation research,* 97**,** 418-26.

SHEN, T. H., CARLSON, C. S. & TARCZY-HORNOCH, P. (2009). SNPit: a federated data integration system for the purpose of functional SNP annotation. *Computer methods and programs in biomedicine,* 95**,** 181-9.

SHENDURE, J. & JI, H. (2008). Next-generation DNA sequencing. *Nature biotechnology,* 26**,** 1135-45.

SHERRATT, A. (1983). The Secondary Exploitation of Animals in the Old-World. *World Archaeology,* 15**,** 90-104.

SHERRY, S. T., WARD, M. & SIROTKIN, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research,* 9**,** 677-9.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research,* 29**,** 308-11.

SHI, H., ZHANG, Q., WANG, Y., YANG, P., WANG, Q. & LI, H. (2011). Chicken adipocyte fatty acid-binding protein knockdown affects expression of peroxisome proliferator-activated receptor gamma gene during oleate-induced adipocyte differentiation. *Poultry science,* 90**,** 1037-44.

SIMES, R. J. (1986). An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika,* 73**,** 751-754.

SMITH, K. D. & BOLOURI, H. (2005). Dissecting innate immune responses with the tools of systems biology. *Current opinion in immunology,* 17**,** 49-54.

SNELLING, W. M., ALLAN, M. F., KEELE, J. W., KUEHN, L. A., MCDANELD, T., SMITH, T. P., SONSTEGARD, T. S., THALLMAN, R. M. & BENNETT, G. L. (2010). Genome-wide association study of growth in crossbred beef cattle. *Journal of animal science,* 88**,** 837-48.

SNELLING, W. M., CUSHMAN, R. A., KEELE, J. W., MALTECCA, C., THOMAS, M. G., FORTES, M. R. & REVERTER, A. (2013). Breeding and Genetics Symposium: networks and pathways to guide genomic selection. *Journal of animal science,* 91**,** 537-52.

STEPHENS, M. & BALDING, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature reviews. Genetics,* 10**,** 681-90.

STORCH, J. & MCDERMOTT, L. (2009). Structural and functional analysis of fatty acid-binding proteins. *Journal of lipid research,* 50 Suppl**,** S126-31.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* 64**,** 479-498.

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics,* 31**,** 2013-2035.

STOREY, J. D. & TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America,* 100**,** 9440-5.

STRAFACE, G., APRAHAMIAN, T., FLEX, A., GAETANI, E., BISCETTI, F., SMITH, R. C., PECORINI, G., POLA, E., ANGELINI, F., STIGLIANO, E., CASTELLOT, J. J., LOSORDO, D. W. & POLA, R. (2009). Sonic hedgehog regulates angiogenesis and myogenesis during post-natal skeletal muscle regeneration. *Journal of cellular and molecular medicine,* 13**,** 2424-2435.

STREINER, D. L. & NORMAN, G. R. (2011). Correction for multiple testing: is there a resolution? *Chest,* 140**,** 16-8.

TABERLET, P., COISSAC, E., PANSU, J. & POMPANON, F. (2011). Conservation genetics of cattle, sheep, and goats. *Comptes rendus biologies,* 334**,** 247-54.

TAKAHASHI, A., KUREISHI, Y., YANG, J., LUO, Z. Y., GUO, K., MUKHOPADHYAY, D., IVASHCHENKO, Y., BRANELLEC, D. & WALSH, K. (2002). Myogenic Akt signaling regulates blood vessel recruitment during myofiber growth. *Molecular and cellular biology,* 22**,** 4803-4814.

TAN, B., YIN, Y., LIU, Z., LI, X., XU, H., KONG, X., HUANG, R., TANG, W., SHINZATO, I., SMITH, S. B. & WU, G. (2009). Dietary L-arginine supplementation increases muscle gain and reduces body fat mass in growing-finishing pigs. *Amino acids,* 37**,** 169-75.

TAN, N. S., MICHALIK, L., DESVERGNE, B. & WAHLI, W. (2005). Multiple expression control mechanisms of peroxisome proliferator-activated receptors and their target genes. *Journal of Steroid Biochemistry and Molecular Biology,* 93**,** 99-105.

TELLAM, R. L., LEMAY, D. G., VAN TASSELL, C. P., LEWIN, H. A., WORLEY, K. C. & ELSIK, C. G. (2009). Unlocking the bovine genome. *BMC genomics,* 10**,** 193.

TENESA, A., NAVARRO, P., HAYES, B. J., DUFFY, D. L., CLARKE, G. M., GODDARD, M. E. & VISSCHER, P. M. (2007). Recent human effective population size estimated from linkage disequilibrium. *Genome research,* 17**,** 520-6.

THE PICARD TOOLKIT. Available: http://picard.sourceforge.net/.

THE R PROJECT. *The R Project for Statistical Computing* [Online]. Available: http://www.r-project.org/.

TIEN, E. S., HANNON, D. B., THOMPSON, J. T. & VANDEN HEUVEL, J. P. (2006). Examination of Ligand-Dependent Coactivator Recruitment by Peroxisome Proliferator-Activated Receptor-alpha (PPARalpha). *PPAR research,* 2006**,** 69612.

TISHKOFF, S. A. & VERRELLI, B. C. (2003). Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annual review of genomics and human genetics,* 4**,** 293-340.

TOLLA, N., MIRKENA, T. & YIMEGNUHAL, A. (2003). Effect of feed restriction on compensatory growth of Arsi (Bos indicus) bulls. *Animal Feed Science and Technology,* 103**,** 29-39.

TRENDELENBURG, A. U., MEYER, A., ROHNER, D., BOYLE, J., HATAKEYAMA, S. & GLASS, D. J. (2009). Myostatin reduces Akt/TORC1/p70S6K signaling, inhibiting myoblast differentiation and myotube size. *American journal of physiology. Cell physiology,* 296**,** C1258-70.

TROSTLER, N., ROMSOS, D. R., BERGEN, W. G. & LEVEILLE, G. A. (1979). Skeletal muscle accretion and turnover in lean and obese (ob/ob) mice. *Metabolism: clinical and experimental,* 28**,** 928-33.

TROY, C. S., MACHUGH, D. E., BAILEY, J. F., MAGEE, D. A., LOFTUS, R. T., CUNNINGHAM, P., CHAMBERLAIN, A. T., SYKES, B. C. & BRADLEY, D. G. (2001). Genetic evidence for Near-Eastern origins of European cattle. *Nature,* 410**,** 1088-91.

TSUCHIDA, K., NAKATANI, M., HITACHI, K., UEZUMI, A., SUNADA, Y., AGETA, H. & INOKUCHI, K. (2009). Activin signaling as an emerging target for therapeutic interventions. *Cell Communication and Signaling,* 7.

VAN DEN BERG, I., FRITZ, S. & BOICHARD, D. (2013). QTL fine mapping with Bayes C(pi): a simulation study. *Genetics, selection, evolution : GSE,* 45**,** 19.

VAN OMMEN, B. & STIERUM, R. (2002). Nutrigenomics: exploiting systems biology in the nutrition and health arena. *Current opinion in biotechnology,* 13**,** 517-21.

VANN, R. C., ALTHEN, T. G., SOLOMON, M. B., EASTRIDGE, J. S., PAROCZAY, E. W. & VEENHUIZEN, J. J. (2001). Recombinant bovine somatotropin (rbST) increases size and proportion of fast-glycolytic muscle fibers in semitendinosus muscle of creep-fed steers. *Journal of animal science,* 79**,** 108-14.

VANRADEN, P. M., JENSEN, E. L., LAWLOR, T. J. & FUNK, D. A. (1990). Prediction of transmitting abilities for Holstein type traits. *Journal of dairy science,* 73**,** 191-7.

VEERKAMP, R. F., DILLON, P., KELLY, E., CROMIE, A. R. & GROEN, A. F. (2002). Dairy cattle breeding objectives combining yield, survival and calving interval for pasture-based systems in Ireland under different milk quota scenarios. *Livestock Production Science,* 76**,** 137-151.

VEERKAMP, R. F., KOENEN, E. P. C. & DE JONG, G. (2001). Genetic correlations among body condition score, yield, and fertility in first-parity cows estimated by random regression models. *Journal of dairy science,* 84**,** 2327-2335.

VEGH, P., FOROUSHANI, A. B., MAGEE, D. A., MCCABE, M. S., BROWNE, J. A., NALPAS, N. C., CONLON, K. M., GORDON, S. V., BRADLEY, D. G., MACHUGH, D. E. & LYNN, D. J. (2013). Profiling microRNA expression in bovine alveolar macrophages using RNA-seq. *Veterinary immunology and immunopathology,* 155**,** 238-44.

VIGNE, J. D. (2011). The origins of animal domestication and husbandry: a major change in the history of humanity and the biosphere. *Comptes rendus biologies,* 334**,** 171-81.

VISSCHER, P. M. (2008). Sizing up human height variation. *Nature genetics,* 40**,** 489-90.

VOELKERDING, K. V., DAMES, S. A. & DURTSCHI, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry,* 55**,** 641-58.

WALSH, B. (2008). Using molecular markers for detecting domestication, improvement, and adaptation genes. *Euphytica,* 161**,** 1-17.

WANG, K., LI, M. & BUCAN, M. (2007). Pathway-based approaches for analysis of genomewide association studies. *American journal of human genetics,* 81**,** 1278-83.

WANG, K., LI, M. & HAKONARSON, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research,* 38**,** e164.

WANG, P., DAI, M., XUAN, W., MCEACHIN, R. C., JACKSON, A. U., SCOTT, L. J., ATHEY, B., WATSON, S. J. & MENG, F. (2006). SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics,* 22**,** e523-9.

WANG, Y. H., BOWER, N. I., REVERTER, A., TAN, S. H., DE JAGER, N., WANG, R., MCWILLIAM, S. M., CAFE, L. M., GREENWOOD, P. L. & LEHNERT, S. A. (2009a). Gene expression patterns during intramuscular fat development in cattle. *Journal of animal science,* 87**,** 119-30.

WANG, Z., GERSTEIN, M. & SNYDER, M. (2009b). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics,* 10**,** 57-63.

WATERS, S. M., BERRY, D. P. & MULLEN, M. P. (2012). Polymorphisms in genes of the somatotrophic axis are independently associated with milk production, udder health, survival and animal size in Holstein-Friesian dairy cattle. *Journal of animal breeding and genetics = Zeitschrift fur Tierzuchtung und Zuchtungsbiologie,* 129**,** 70-8.

WATERS, S. M., MCCABE, M. S., HOWARD, D. J., GIBLIN, L., MAGEE, D. A., MACHUGH, D. E. & BERRY, D. P. (2011). Associations between newly discovered polymorphisms in the Bos taurus growth hormone receptor gene and performance traits in Holstein-Friesian dairy cattle. *Animal genetics,* 42**,** 39-49.

WEEDON, M. N., LANGO, H., LINDGREN, C. M., WALLACE, C., EVANS, D. M., MANGINO, M., FREATHY, R. M., PERRY, J. R., STEVENS, S., HALL, A. S., SAMANI, N. J., SHIELDS, B., PROKOPENKO, I., FARRALL, M., DOMINICZAK, A., JOHNSON, T., BERGMANN, S., BECKMANN, J. S., VOLLENWEIDER, P., WATERWORTH, D. M., MOOSER, V., PALMER, C. N., MORRIS, A. D., OUWEHAND, W. H., ZHAO, J. H., LI, S., LOOS, R. J., BARROSO, I., DELOUKAS, P., SANDHU, M. S., WHEELER, E., SORANZO, N., INOUYE, M., WAREHAM, N. J., CAULFIELD, M., MUNROE, P. B., HATTERSLEY, A. T., MCCARTHY, M. I. & FRAYLING, T. M. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature genetics,* 40**,** 575-83.

WICKHAM, B. W. & DURR, J. W. (2011). A new international infrastructure for beef cattle breeding. *Animal Frontiers,* 1**,** 53-59.

WIERSTRA, I. & ALVES, J. (2007). FOXM1, a typical proliferation-associated transcription factor. *Biological chemistry,* 388**,** 1257-74.

WILLHAM, R. L. (1986). From Husbandry to Science - a Highly Significant Facet of Our Livestock Heritage. *Journal of animal science,* 62**,** 1742-1758.

WINTERS, R., WINTERS, A. & AMEDEE, R. G. (2010). Statistics: a brief overview. *The Ochsner journal,* 10**,** 213-6.

WRAY, N. R., GODDARD, M. E. & VISSCHER, P. M. (2007). Prediction of individual genetic risk to disease from genome-wide association studies. *Genome research,* 17**,** 1520-8.

WTCCC (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature,* 447**,** 661-78.

XU, X., ZHANG, Y., WILLIAMS, J., ANTONIOU, E., MCCOMBIE, W. R., WU, S., ZHU, W., DAVIDSON, N. O., DENOYA, P. & LI, E. (2013). Parallel comparison of Illumina RNA-Seq and Affymetrix microarray platforms on transcriptomic profiles generated from 5-aza-deoxy-cytidine treated HT-29 colon cancer cells and simulated datasets. *BMC bioinformatics,* 14 Suppl 9**,** S1.

YAMAGUCHI, Y. & HEARING, V. J. (2009). Physiological factors that regulate skin pigmentation. *BioFactors,* 35**,** 193-9.

YAN, X. L., BAXTER, R. C. & FIRTH, S. M. (2010). Involvement of Pregnancy-Associated Plasma Protein-A2 in Insulin-Like Growth Factor (IGF) Binding Protein-5 Proteolysis during Pregnancy: A Potential Mechanism for Increasing IGF Bioavailability. *Journal of Clinical Endocrinology & Metabolism,* 95**,** 1412-1420.

YANDELL, M. & ENCE, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature reviews. Genetics,* 13**,** 329-42.

YANG, W. & TEMPELMAN, R. J. (2012). A Bayesian antedependence model for whole genome prediction. *Genetics,* 190**,** 1491-501.

YOUNG, M. D., WAKEFIELD, M. J., SMYTH, G. K. & OSHLACK, A. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome biology,* 11**,** R14.

ZEYLAND, J., WOLKO, L., BOCIANOWSKI, J., SZALATA, M., SLOMSKI, R., DZIEDUSZYCKI, A. M., RYBA, M., PRZYSTALOWSKA, H. & LIPINSKI, D. (2013). Complete mitochondrial genome of wild aurochs

(Bos primigenius) reconstructed from ancient DNA. *Polish journal of veterinary sciences*, 16**,** 265-73.

ZHANG, J., GUAN, L., WEN, W., LU, Y., ZHU, Q., YUAN, H., CHEN, Y., WANG, H. & LI, H. (2013a). A novel mutation of DNAH5 in chronic rhinosinusitis and primary ciliary dyskinesia in a Chinese family. *European archives of oto-rhino-laryngology : official journal of the European Federation of Oto-Rhino-Laryngological Societies*.

ZHANG, L., LIU, J., ZHAO, F., REN, H., XU, L., LU, J., ZHANG, S., ZHANG, X., WEI, C., LU, G., ZHENG, Y. & DU, L. (2013b). Genome-wide association studies for growth and meat production traits in sheep. *PloS one,* 8**,** e66569.

ZHANG, Z., ERSOZ, E., LAI, C. Q., TODHUNTER, R. J., TIWARI, H. K., GORE, M. A., BRADBURY, P. J., YU, J., ARNETT, D. K., ORDOVAS, J. M. & BUCKLER, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics,* 42**,** 355-60.

ZHANG, Z., SCHWARTZ, S., WAGNER, L. & MILLER, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology,* 7**,** 203-14.

ZHANG, Z. L., HARRISON, P. M., LIU, Y. & GERSTEIN, M. (2003). Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome research,* 13**,** 2541-2558.

ZHENG, D., FRANKISH, A., BAERTSCH, R., KAPRANOV, P., REYMOND, A., CHOO, S. W., LU, Y., DENOEUD, F., ANTONARAKIS, S. E., SNYDER, M., RUAN, Y., WEI, C. L., GINGERAS, T. R., GUIGO, R., HARROW, J. & GERSTEIN, M. B. (2007). Pseudogenes in the

ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome research,* 17**,** 839-51.

ZHENG, G., TU, K., YANG, Q., XIONG, Y., WEI, C., XIE, L., ZHU, Y. & LI, Y. (2008). ITFP: an integrated platform of mammalian transcription factors. *Bioinformatics,* 24**,** 2416-7.

ZHOU, X., WANG, J. L., LU, J., SONG, Y., KWAK, K. S., JIAO, Q., ROSENFELD, R., CHEN, Q., BOONE, T., SIMONET, W. S., LACEY, D. L., GOLDBERG, A. L. & HAN, H. Q. (2010). Reversal of cancer cachexia and muscle wasting by ActRIIB antagonism leads to prolonged survival. *Cell,* 142**,** 531-43.

ZHOU, Y., XU, B. C., MAHESHWARI, H. G., HE, L., REED, M., LOZYKOWSKI, M., OKADA, S., CATALDO, L., COSCHIGAMO, K., WAGNER, T. E., BAUMANN, G. & KOPCHICK, J. J. (1997). A mammalian model for Laron syndrome produced by targeted disruption of the mouse growth hormone receptor/binding protein gene (the Laron mouse). *Proceedings of the National Academy of Sciences of the United States of America,* 94**,** 13215-20.

ZHU, M., YU, M. & ZHAO, S. (2009). Understanding quantitative genetics in the systems biology era. *International journal of biological sciences,* 5**,** 161-70.

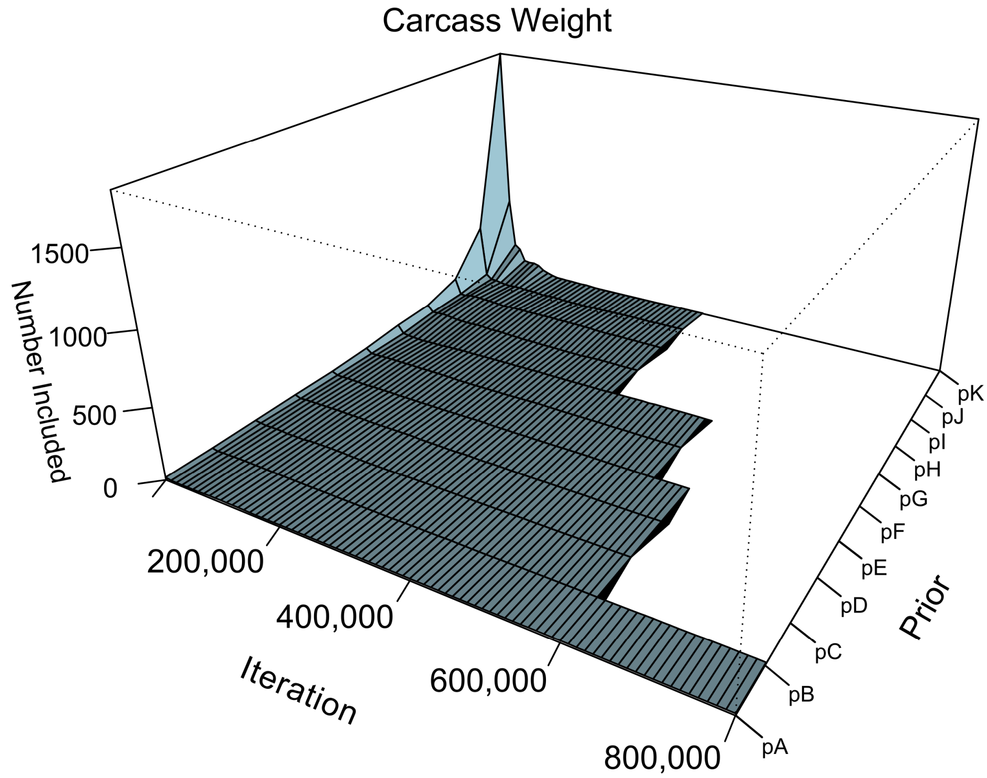
ZIEGLER, A., KONIG, I. R. & THOMPSON, J. R. (2008). Biostatistical aspects of genome-wide association studies. *Biometrical journal. Biometrische Zeitschrift,* 50**,** 8-28.

ZIMIN, A. V., DELCHER, A. L., FLOREA, L., KELLEY, D. R., SCHATZ, M. C., PUIU, D., HANRAHAN, F., PERTEA, G., VAN TASSELL, C. P., SONSTEGARD, T. S., MARCAIS, G., ROBERTS, M., SUBRAMANIAN, P., YORKE, J. A. & SALZBERG, S. L. (2009). A whole-genome assembly of the domestic cow, Bos taurus. *Genome biology*, 10**,** R42.

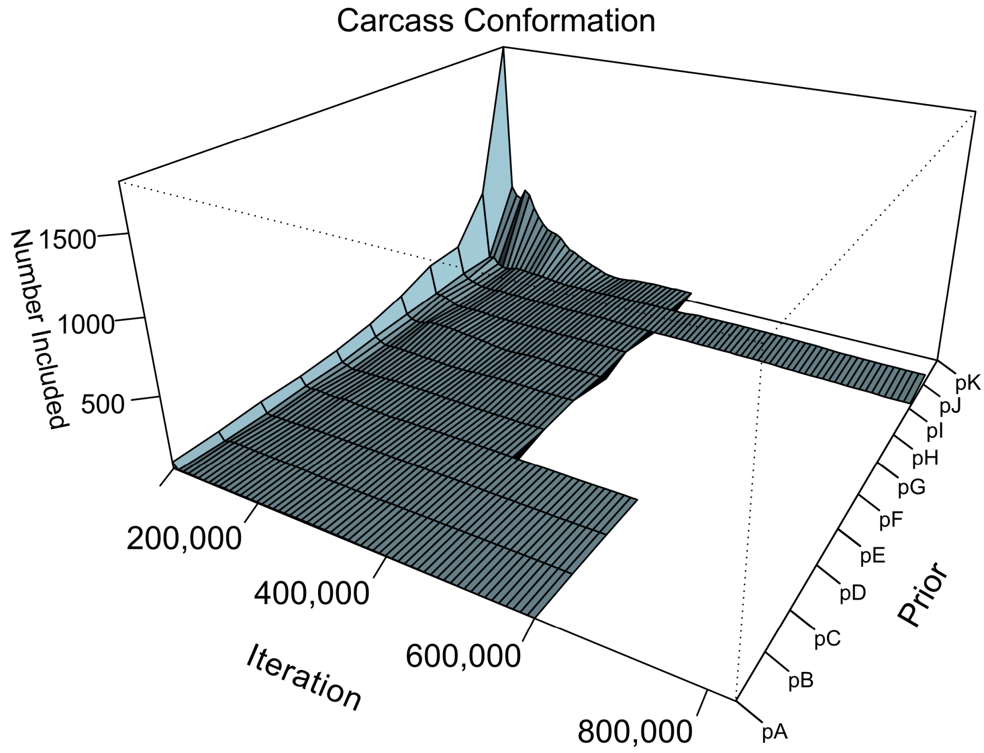
Appendix

Appendix 1 (Chapter 4)

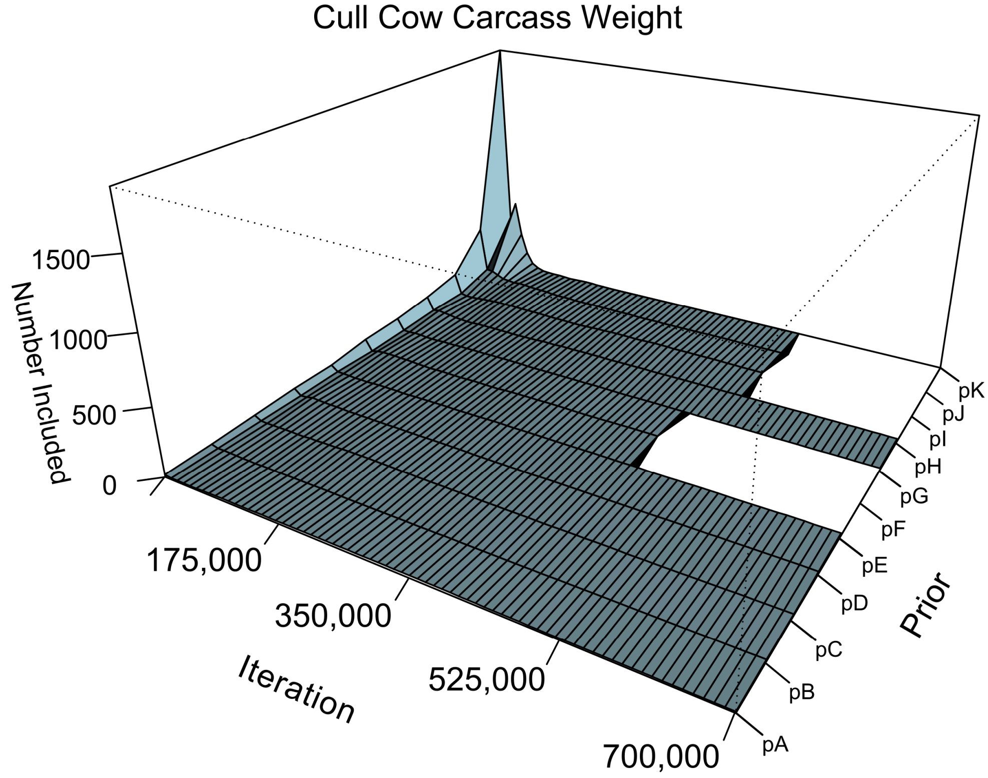
A1.1 Surface plot of the number of SNPs with a PP>0.5 for the carcass weight phenotype



A1.2 Surface plot of the number of SNPs with a PP>0.5 for the carcass conformation phenotype

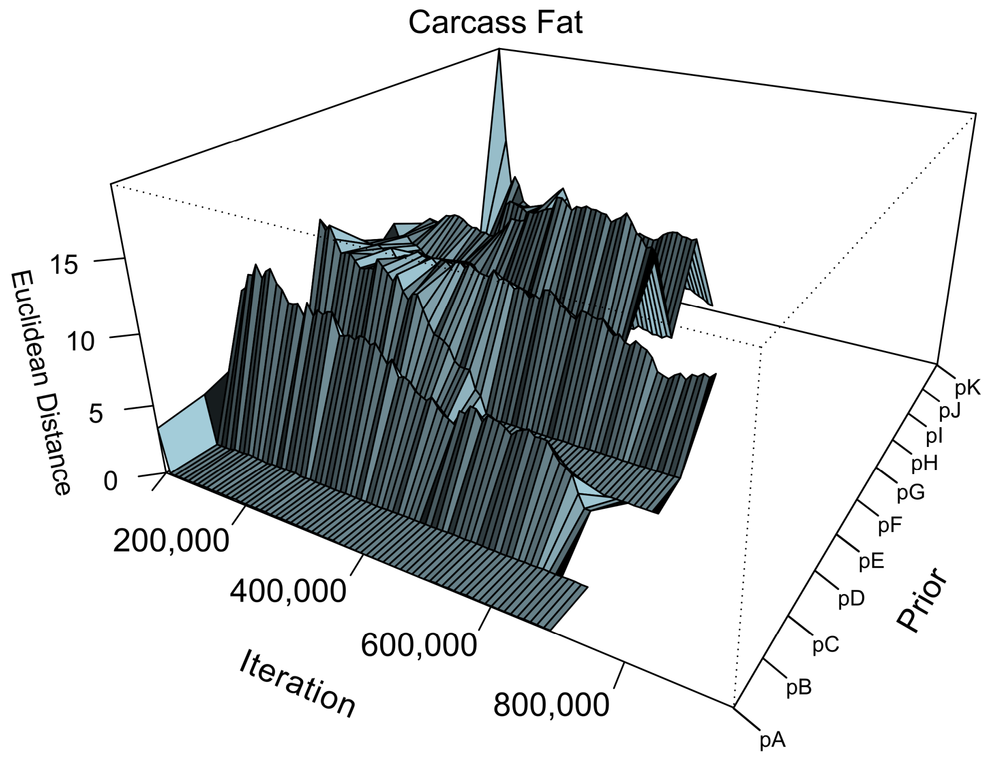


A1.3 Surface plot of the number of SNPs with a PP>0.5 for the cull cow carcass weight phenotype

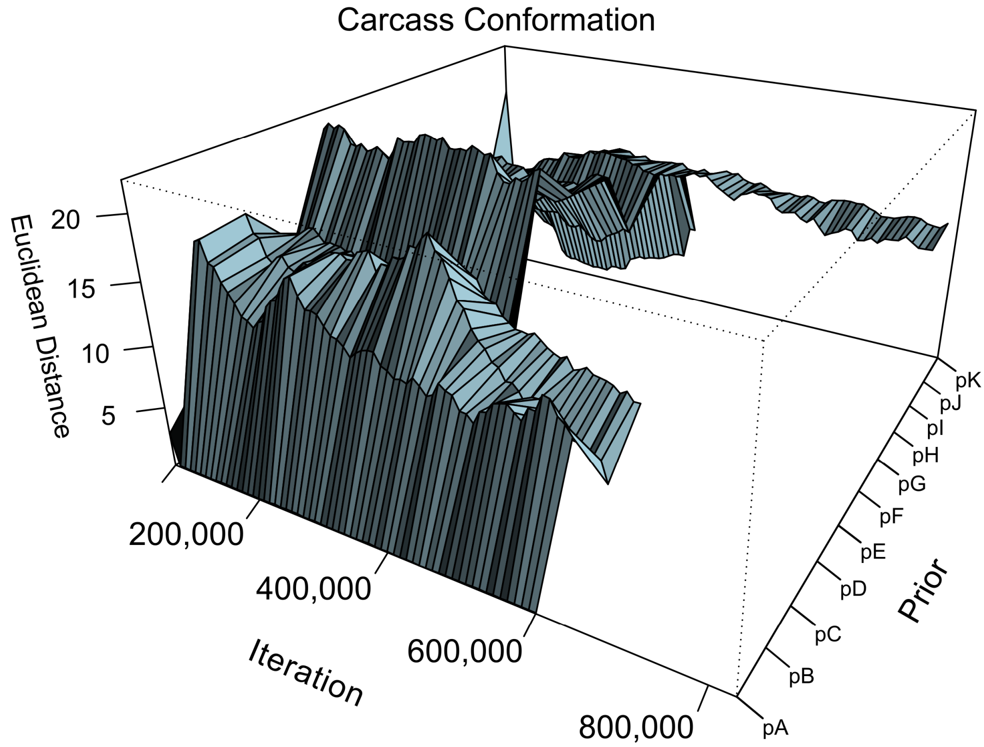


Appendix 2 (Chapter 4)

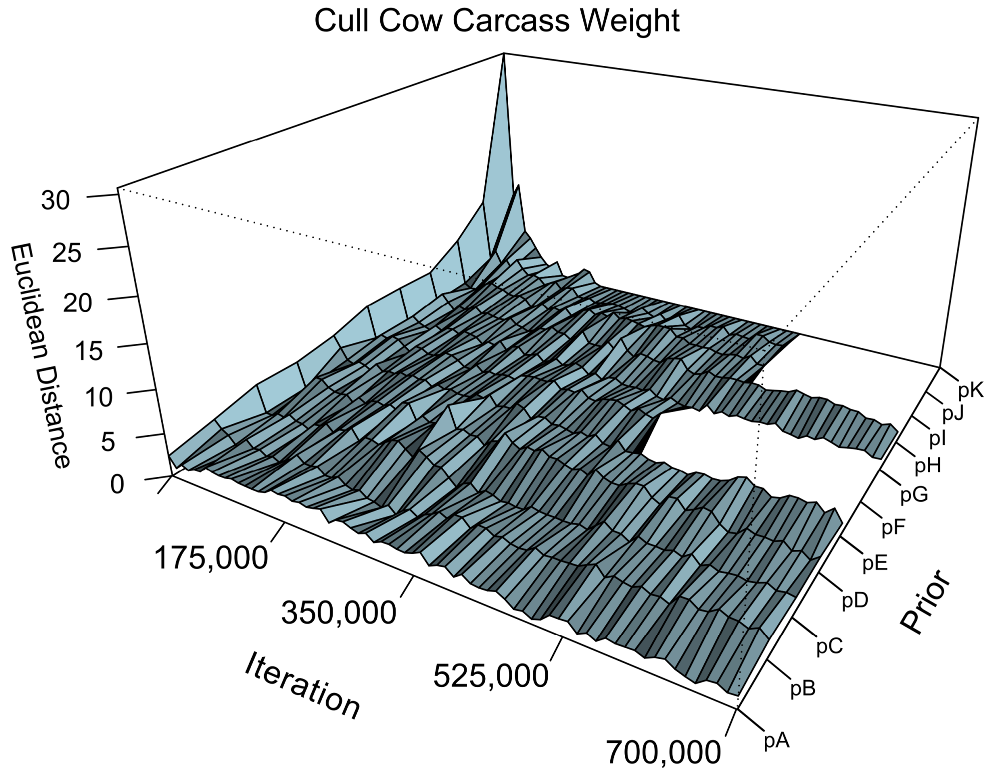
A2.1 Surface plot of the Euclidean distances for the carcass fat phenotype



A2.2 Surface plot of the Euclidean distances for the carcass conformation phenotype

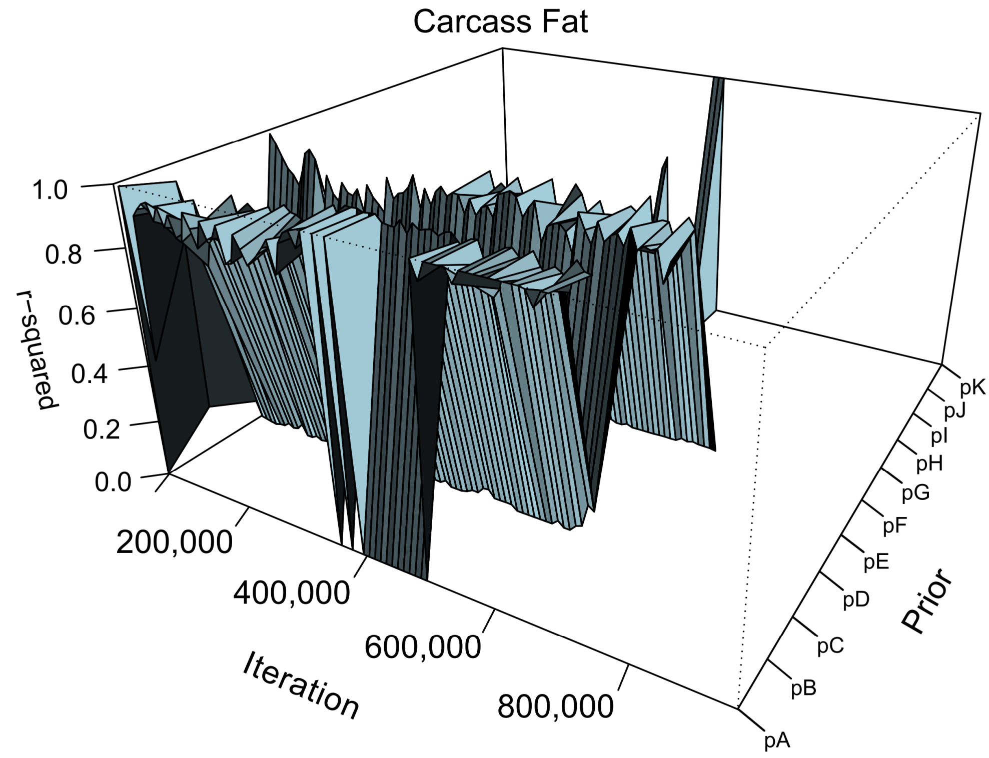


A2.3 Surface plot of the Euclidean distances for the cull cow carcass weight phenotype

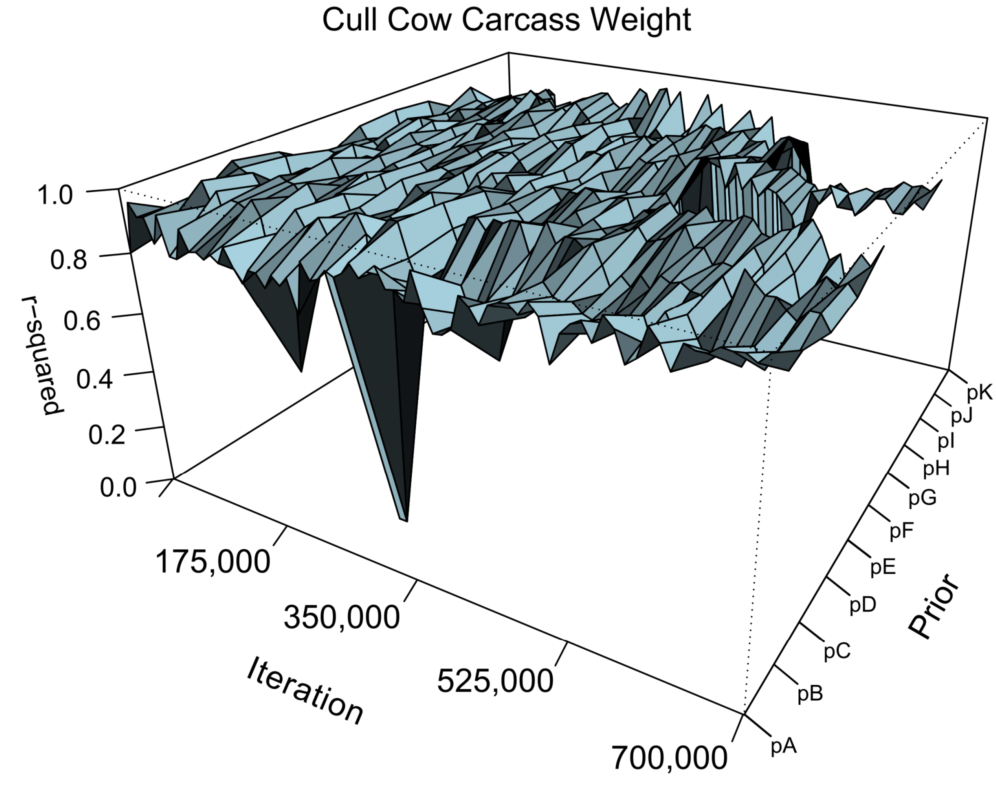


Appendix 3 (Chapter 4)

A3.1 Surface plot of the R^2 for the carcass fat phenotype



A3.2 Surface plot of the R^2 for the cull cow carcass weight phenotype



Appendix 4 (Chapter 5)

A4.1 Muscle sample collection

A biopsy (~ 0.5 g to 1 g) of muscle between the 12th and 13th rib (*M. longissimus thoracis et lumborum*) was taken with a trochar and cannula instrument. Each sample was washed in DPBS and snap frozen in liquid nitrogen. Samples were stored at -80 °C until total RNA was extracted.

A4.2 Extraction of total RNA from muscle

For each sample; total RNA was prepared from 100mg of frozen muscle tissue. Tissue samples were homogenised in 3 ml of TRIzol reagent. After homogenisation, the mixture was evenly divided into 3 Eppendorf tubes and 200 µl of chloroform was added. The tubes were then shaken vigorously and incubated at room temperature for 2 min. The resulting mixture was centrifuged at 12,000 g for 15 min at 4^{o}C. The aqueous phase was then transferred to a fresh 1.5 ml tube, and isopropanol was added and mixed by vortexing the tube for 10 sec. The sample was then centrifuged at 12,000 g for 10 min at 4^{o}C, after which the supernatant was removed and 1 ml of 75% ethanol was added to the remaining RNA pellet. The sample was vortexed and then centrifuged at 7,500 g for 5 min at 4^{o}C. Following centrifugation, the supernatant was removed and the pellet air-dried briefly ensuring that the pellet did not dry out completely. Nuclease-free water (20 µl) was added to each tube and the pellet dissolved completely by gently pipetting. The contents of each tube were pooled and added to a sterile 1.5 ml tube. The Nanodrop spectrophotometer was used to determine total RNA quantity in the sample. The concentration of total RNA for each sample in triplicate was recorded and averaged. The a260/a280 ratio, an indicator of protein contamination was also recorded. Samples with ratios ranging between 1.8 and 2.0 were accepted. The Agilent Bioanalyser 2100 (Agilent Technologies) was used to assess the quality of the RNA. Samples with a RNA integrity number (RIN) equal to or greater than 8 were deemed acceptable and used as input for the preparation of mRNA-seq cDNA libraries.

Appendix 5 (Chapter 5)

A5.1 Live weight, live weight gains and dry matter intake for the control and experimental group at each time-point.

Trait	Control group	Experimental group	SED
Live weight (kg)			
Start (day 0)	296	298	6.89
End of differential feeding period (day 99)	438	356	6.89
Re-alimentation period (day 131)	474	416	6.89
Live weight gain (kg/day)			
Differential feeding period (day 0-99)	1.55	0.63	0.05
Re-alimentation period (day 99-131)	1.26	1.74	0.093
DMI (kg/day)			
Differential feeding period (day 0-99)	9.48	4.41	0.08
Re-alimentation period (day 99-131)	10.34	10.25	0.09

Appendix 6 (Chapter 5)

A6.1 FASTQ file names and alignment statistics for GEO submission GSE48481

GEO counts file name	Corresponding FASTQ file	Strand	Raw Reads	Reads removed by trimming	Reads post trimming	Aligned exactly 1 time Reads
run_110218_1_00023_tp1_cont_r1_counts.txt	run_110218.s_1_1_sequence.fq.gz	Forward	29744590	2929632	26814958	11087004
	run_110218.s_1_2_sequence.fq.gz	Reverse	29744590	2929632	26814958	11087395
run_110218_2_00553_tp1_rest_r1_counts.txt	run_110218.s_2_1_sequence.fq.gz	Forward	33267602	1432866	31834736	18160670
	run_110218.s_2_2_sequence.fq.gz	Reverse	33267602	1432866	31834736	18134204
run_110218_3_00030_tp2_cont_r1_counts.txt	run_110218.s_3_1_sequence.fq.gz	Forward	36903677	3578101	33325576	17522406
	run_110218.s_3_2_sequence.fq.gz	Reverse	36903677	3578101	33325576	17041299
run_110218_4_00023_tp1_cont_r2_counts.txt	run_110218.s_4_1_sequence.fq.gz	Forward	34481888	3257689	31224199	13428818
	run_110218.s_4_2_sequence.fq.gz	Reverse	34481888	3257689	31224199	13088928
run_110218_6_00553_tp1_rest_r2_counts.txt	run_110218.s_6_1_sequence.fq.gz	Forward	34277957	1729668	32548289	18738853
	run_110218.s_6_2_sequence.fq.gz	Reverse	34277957	1729668	32548289	18408845
run_110218_7_00030_tp2_cont_r2_counts.txt	run_110218.s_7_1_sequence.fq.gz	Forward	36215753	1655711	34560042	18101039
	run_110218.s_7_2_sequence.fq.gz	Reverse	36215753	1655711	34560042	17941491
run_110218_8_00553_tp1_rest_r3_counts.txt	run_110218.s_8_1_sequence.fq.gz	Forward	35787810	1072539	34715271	19987018
	run_110218.s_8_2_sequence.fq.gz	Reverse	35787810	1072539	34715271	19829224
run_110321_1_00553_tp2_rest_counts.txt	run_110321.s_1_1_sequence.fq.gz	Forward	30649324	1078514	29570810	13844582
	run_110321.s_1_2_sequence.fq.gz	Reverse	30649324	1078514	29570810	13645278
run_110321_2_00030_tp1_cont_counts.txt	run_110321.s_2_1_sequence.fq.gz	Forward	30611937	895376	29716561	16662457
	run_110321.s_2_2_sequence.fq.gz	Reverse	30611937	895376	29716561	16476307
run_110321_3_00465_tp1_rest_counts.txt	run_110321.s_3_1_sequence.fq.gz	Forward	28016772	650007	27366765	13493740
	run_110321.s_3_2_sequence.fq.gz	Reverse	28016772	650007	27366765	13348691
run_110321_4_00414_tp2_cont_counts.txt	run_110321.s_4_1_sequence.fq.gz	Forward	32233953	918516	31315437	15212628
	run_110321.s_4_2_sequence.fq.gz	Reverse	32233953	918516	31315437	15096723

run_110321_6_00587_tp2_rest_counts.txt	run_110321.s_6_1_sequence.fq.gz	Forward	33579604	2896869	30682735	10923643
	run_110321.s_6_2_sequence.fq.gz	Reverse	33579604	2896869	30682735	9930809
run_110321_7_00468_tp1_cont_counts.txt	run_110321.s_7_1_sequence.fq.gz	Forward	38106809	1587028	36519781	20248894
	run_110321.s_7_2_sequence.fq.gz	Reverse	38106809	1587028	36519781	19956491
run_110321_8_00926_tp1_rest_counts.txt	run_110321.s_8_1_sequence.fq.gz	Forward	30708273	900407	29807866	16652524
	run_110321.s_8_2_sequence.fq.gz	Reverse	30708273	900407	29807866	16546043
run_110622_1_00023_tp2_cont_counts.txt	run_110622.s_1_1_sequence.fq.gz	Forward	30350428	628709	29721719	11584093
	run_110622.s_1_2_sequence.fq.gz	Reverse	30350428	628709	29721719	11635435
run_110622_2_00921_tp2_rest_counts.txt	run_110622.s_2_1_sequence.fq.gz	Forward	36577555	872238	35705317	14185395
	run_110622.s_2_2_sequence.fq.gz	Reverse	36577555	872238	35705317	14120204
run_110622_3_00414_tp1_cont_counts.txt	run_110622.s_3_1_sequence.fq.gz	Forward	37336377	974626	36361751	19645522
	run_110622.s_3_2_sequence.fq.gz	Reverse	37336377	974626	36361751	19564367
run_110622_4_00587_tp1_rest_counts.txt	run_110622.s_4_1_sequence.fq.gz	Forward	38310347	1087462	37222885	21153101
	run_110622.s_4_2_sequence.fq.gz	Reverse	38310347	1087462	37222885	21066344
run_110622_6_00584_tp2_cont_counts.txt	run_110622.s_6_1_sequence.fq.gz	Forward	34436732	677216	33759516	16030922
	run_110622.s_6_2_sequence.fq.gz	Reverse	34436732	677216	33759516	15967174
run_110622_7_00976_tp2_rest_counts.txt	run_110622.s_7_1_sequence.fq.gz	Forward	29191963	784974	28406989	16612087
	run_110622.s_7_2_sequence.fq.gz	Reverse	29191963	784974	28406989	16566322
run_110622_8_00521_tp1_cont_counts.txt	run_110622.s_8_1_sequence.fq.gz	Forward	35348221	1089664	34258557	18633181
	run_110622.s_8_2_sequence.fq.gz	Reverse	35348221	1089664	34258557	18535473
run_110715_1_00921_tp1_rest_counts.txt	run_110715.s_1_1_sequence.fq.gz	Forward	39415231	1218682	38196549	18337574
	run_110715.s_1_2_sequence.fq.gz	Reverse	39415231	1218682	38196549	18359401
run_110715_2_00521_tp2_cont_counts.txt	run_110715.s_2_1_sequence.fq.gz	Forward	32426502	323292	32103210	17423823
	run_110715.s_2_2_sequence.fq.gz	Reverse	32426502	323292	32103210	17360144
run_110715_3_00465_tp2_rest_counts.txt	run_110715.s_3_1_sequence.fq.gz	Forward	34480283	969410	33510873	15930298
	run_110715.s_3_2_sequence.fq.gz	Reverse	34480283	969410	33510873	15830837
run_110715_4_00584_tp1_cont_counts.txt	run_110715.s_4_1_sequence.fq.gz	Forward	26634475	253675	26380800	11787365
	run_110715.s_4_2_sequence.fq.gz	Reverse	26634475	253675	26380800	11728281
run_110715_6_00976_tp1_rest_counts.txt	run_110715.s_6_1_sequence.fq.gz	Forward	33738230	404085	33334145	20572938

	run_110715.s_6_2_sequence.fq.gz	Reverse	33738230	404085	33334145	20469177
run_110715_7_00468_tp2_cont_counts.txt	run_110715.s_7_1_sequence.fq.gz	Forward	33503514	390638	33112876	14991768
	run_110715.s_7_2_sequence.fq.gz	Reverse	33503514	390638	33112876	14882725
run_110715_8_00926_tp2_rest_counts.txt	run_110715.s_8_1_sequence.fq.gz	Forward	29670975	381935	29289040	16783695
	run_110715.s_8_2_sequence.fq.gz	Reverse	29670975	381935	29289040	16710238