Genetic and Epigenetic Regulation in Angus and Brahman Cattle

Thesis Submitted to The University of Adelaide in fulfilment of the requirement for the degree of Doctor of Philosophy

Callum Robin Lindsay MacPhillamy

BSc Biological Sciences (Hons)

School of Animal and Veterinary Sciences

The University of Adelaide

September 2023

Table of Contents

Abstract

Angus and Brahman cattle represent two economically important subspecies of cattle with contrasting phenotypes. The Angus cattle breed is representative of the taurine subspecies and has been bred for excellent meat production traits, and the Brahman cattle breed is representative of the indicine subspecies and has been bred for its ability to thrive in harsh conditions. Knowledge of genetic regulation is fundamental to our understanding of what causes these contrasting phenotypes in cattle breeds. Gene regulatory differences can arise because of genetic and epigenetic differences among the breeds, which can shed light on what contributes to the different phenotypes. Despite this knowledge being crucial to understanding how complex traits are controlled, relatively little is known about genetic and epigenetic regulatory differences between the cattle subspecies. This thesis investigated genetic and epigenetic differences between cattle subspecies by using Angus to represent taurine cattle and Brahman to represent indicine cattle in an effort to elucidate factors responsible for their distinct phenotypes.

Enhancers are a key genetic regulatory element, but relatively little is known about this DNA element in the cattle genome. The performance of nine machine learning (ML) models and four DNA representations was evaluated to determine the best combination to predict enhancers across species to cattle using models trained on high-quality enhancers in human and mouse. To evaluate the usefulness of cross-species prediction in general, the ML models were also applied to find pig and dog enhancers. For identifying enhancers in cattle, pig and dog, the combination of convolutional neural networks and one-hot encoding to represent the DNA sequence performed the best. They predicted a similar proportion of enhancers in these genomes as what has been estimated to be the proportion of enhancers in the human genome.

Whole genome bisulfite sequencing (WGBS) data was generated from Brahman, Angus and reciprocally crossed progeny using fetal liver samples to investigate differential methylation between the Brahman and Angus breeds. The reciprocal crosses were used to investigate the parent-of-origin effects on DNA methylation to determine what role dam and sire genetics had on the progeny's methylome. As breed-specific reference genomes are available for Brahman and Angus, the impact of reference genome choice was investigated to determine how this affects downstream analyses. The methylation analysis identified tens of thousands of differentially methylated regions (DMRs) that were breed-specific and parent-of-origin-specific. One of the DMRs may be controlling the expression of *Dgat1* in a breed and sire-of-origin manner. Genome comparison revealed around 75% of CpGs were shared between Brahman and Angus, with around 5% (~one million CpGs) being breed-specific. Moreover, single nucleotide polymorphisms (SNPs) and structural variants (SVs) between Brahman and Angus were 8-fold (p-value < 0.05) and 1.13-fold (p-value < 0.05) higher in CpGs, respectively, and a quantification bias of 2% was observed when the incorrect reference genome was used for analysis.

MicroRNA (miRNA) expression data was generated from the same samples that were used to generate WGBS data. This expression data was used to identify differentially expressed, breed-specific and parent-of-origin-specific miRNAs. Fourteen differentially expressed miRNAs (DEMs) were observed between the breeds, with the dam-of-origin and sire-of-origin comparisons identifying one and five DEMs, respectively. Genes that were predicted to be targets of the DEMs were significantly (p-value <0.05) more likely to be differentially expressed than genes not predicted to be targets of the DEMs. The expression of these miRNAs was then correlated with mRNA expression from the same samples and used to identify gene regulatory pathways that may be under microRNA control. MiRNAs that may be involved in regulating heat tolerance in Brahman and fat gain in Angus were identified, as well as a series of signalling pathways that, through differential gene regulation, may contribute to phenotypic differences between Brahman and Angus cattle.

Overall, this thesis identified genetic and epigenetic regions of interest between Brahman and Angus that can help shed light on the causes of the contrasting phenotypes observed between Brahman and Angus cattle. This information will benefit future functional studies that look to pinpoint causative elements controlling these traits.

Acknowledgements

I want to take this opportunity to express my most profound appreciation and gratitude to all those who have supported and encouraged me throughout my PhD journey. First and foremost, I am indebted to my supervisor, Dr Wai Yee Low. Lloyd, I cannot express how fortunate I am to have had you as my supervisor. The first 12 months of my PhD were a tremendous emotional and mental struggle; on more than one occasion, I questioned whether I should be doing a PhD. I am not sure I would have stayed if it had been for your patience, kindness, enthusiasm, encouragement, and guidance. For all this and more, I will be eternally grateful. Furthermore, your expertise in a wide range of fields and ability to always come up with a perspective or question I had not considered has given me an excellent foundation for the researcher I want to be.

To Dr Hamid Alinejad-Rokny, your support, guidance, and expertise were invaluable to shaping my research. The camaraderie and collaborative nature you foster in your lab were tremendously encouraging. I will forever aspire to emulate the culture you cultivate in all aspects of my work life.

Finally, I would like to thank Professor Wayne Pitchford. Your constant asking, "How does this help with a breeding programme?" instilled in me a sense to always think about the big picture and where my research fits in the broader context of positively contributing to society. To all my supervisors, your constructive feedback and insightful suggestions have been invaluable in shaping my research.

As the adage goes, “Misery loves company”, so I would like to thank my fellow PhD students for their support and for reminding me that no one’s PhD goes

smoothly 100% of the time. I want to thank Leesa-Joy Flanagan, who was always happy to share in a rant about the trials and tribulations of completing a PhD.

Furthermore, I would like to sincerely thank my family and friends for their endless support and encouragement. I am grateful for their unwavering belief in my abilities. I would be remiss not to mention my partner, Michaela, whose (sometimes reluctant) willingness to listen to presentations multiple times helped me better communicate my research more concisely. I have hugely appreciated her support and encouragement throughout the PhD.

Lastly, I would like to thank Ben, Marcus and Henry of the Last Podcast on the Left, whose humour and storytelling kept me company on long drives to and from campus and during long programming and analysis sessions.

List of Publications and Expected Publications

1. Ren Y., **MacPhillamy C.**, To T.-H., Smith T.P.L., Williams J.L. & Low W.Y. (2021) Adaptive selection signatures in river buffalo with emphasis on immune and major histocompatibility complex genes. Genomics 113, 3599-609. https://doi.org/10.1016/j.ygeno.2021.08.021
2. **MacPhillamy C.**, Pitchford W.S., Alinejad-Rokny H. & Low W.Y. (2021) Opportunity to improve livestock traits using 3D genomics. Animal Genetics 52, 785-98. 10.1111/age.13135
3. **MacPhillamy C.**, Alinejad-Rokny H., Pitchford W.S. & Low W.Y. (2022) Cross-species enhancer prediction using machine learning. Genomics 114, 110454. 10.1016/j.ygeno.2022.110454
4. **MacPhillamy C.**, Chen T., Hiendleder S., Williams J.S., Alinejad-Rokny H. & Low W.Y. (2023) The genetics of epigenetics in the bovine pangenome era. **Under review with BMC Biology**
5. **MacPhillamy C.**, Yan R., Chen T., Hiendleder S., Williams J.L. & Low W.Y. (2023) MicroRNA breed and parent-of-origin effects provide insights into biological pathways differentiating cattle subspecies. **Manuscript ready for submission**

Conference Presentations

1. ‘A comparative framework for comparing Hi-C datasets across ruminant livestock species’. International Society of Animal Genetics. 26-30 July 2021. Virtual
2. ‘Predicting enhancers across species using machine learning’. XXIII International Congress of Genetics, 16-21 July 2023. Melbourne, Australia

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within the thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed:

Callum MacPhillamy

Date: 4th of September 2023

Chapter 1: General Introduction

Thesis Introduction

Australian beef is regarded as some of the best in the world, which is partly thanks to years of research conducted in breed improvement. The Australian beef herd comprises several different breeds, with Brahman and Angus cattle making up a substantial proportion of the total herd. Brahman is a beef breed and a member of the *Bos taurus indicus* subspecies of cattle that is known for its heat and disease tolerance and its ability to maintain productivity in harsh conditions with low-quality feed. Angus is another beef breed that is a member of the *B. taurus taurus* subspecies and is known for its high-quality beef, good intramuscular fat content, fertility, and quick maturity. Brahman and Angus cattle have opposing phenotypes; Brahman cattle thrive in hot climates, whereas Angus cattle struggle to maintain productivity in the same conditions. Angus cattle mature quickly, while Brahman cattle mature much more slowly. Angus cattle produce tender, high-value beef; conversely, Brahman cattle tend to produce tougher, lower-quality beef. Despite being separate subspecies, Brahman and Angus cattle remain sexually compatible and can produce fertile offspring. Many composite breeds have been established based on different combinations of Brahman and Angus genetics. However, the exact genetic and epigenetic mechanisms underpinning their contrasting phenotypes have yet to be fully understood. Therefore, this project aimed to investigate and identify the genetic and epigenetic differences between Brahman and Angus cattle that may contribute to their contrasting phenotypes.

Changes in gene expression drive phenotypic differences between cattle breeds. One key regulator of gene expression are enhancers but very little is known about this important class of cis-regulatory element. Substantial volumes of data have been generated to identify and validate enhancers throughout the human and mouse genomes. The increased interest in machine learning for enhancer prediction, particularly across species, coupled with high-quality data sets describing these genetic elements, means there is an opportunity to leverage these tools and data to improve our understanding of where enhancers are within the cattle genome and how these elements may be controlling gene expression. Additionally, given the ability of Brahman and Angus to produce hybrid offspring, there is an opportunity to investigate not only the effect that breed has on the gene expression of the progeny but also how the genetics of the sire and dam individually influences gene regulation in the progeny.

This thesis first evaluates enhancer prediction models and then uses the best one to identify enhancers genome-wide in the cattle reference genome. This evaluation was performed to provide an improved genome annotation for cattle. It then compares DNA methylation differences between Brahman and Angus around a variety of genomic features including the predicted enhancers, to identify associations between DNA methylation and regions of interest associated with breed and parent-of-origin differences. Finally, this thesis examines miRNA expression differences between the breeds and parent-of-origin groups to determine whether any associations between miRNA expression and breed differences exist. By comparing these genetic and epigenetic differences between Brahman, Angus and reciprocally crossed progeny liver samples, there is an opportunity to identify regions of interest that may contribute to the contrasting phenotypes of these two breeds.

Thesis Structure

This thesis contains six chapters comprised of a general thesis introduction (Chapter 1), a review of the literature (Chapter 2), three research chapters in publication format (Chapters 3-5), one of which has been published (Chapter 3) and another one (Chapter 4) is currently under review with a journal. Lastly, it concludes with a general discussion (Chapter 6). An overview of the thesis is presented in Figure 1.1.

The literature review chapter, Chapter 2, aims to introduce the reader to the main themes of the thesis; the evolutionary history of Brahman and Angus cattle, the role of enhancers, DNA methylation and microRNAs in gene regulation and what is known about them in cattle. It also introduces the reader to the concept of the "genetics of epigenetics" and the potential applications of the thesis to the beef industry in Australia.

In Chapter 3, a published research paper (MacPhillamy *et al*. 2022, *Genomics*) evaluated the performance of machine learning models in cross-species enhancer prediction was evaluated. This chapter involved testing several different machine-learning models and DNA representations first on human and mouse enhancer data and then applying them to predict enhancers in three less well-studied mammalian species, specifically cattle, pig and dog. The results of this chapter suggest that machine learning models can identify enhancers in these species better than random chance. Furthermore, machine learning models predicted a similar proportion of enhancers within the genome of these three species as what has been predicted by ENCODE for the human genome.

The reduction in sequencing costs has led to a substantial increase in breed-specific genomes. However, the impact this has on downstream analyses, particularly of DNA methylation, which is susceptible to mutations between breeds, has not been well characterised. In Chapter 4 (currently under review with BMC Biology), whole genome bisulfite sequencing (WGBS) data generated from Brahman, Angus and reciprocally crossed progeny liver samples was used to investigate DNA methylation differences between the two breeds, including at enhancer sites identified in Chapter 3. WGBS reads were mapped to the reference genomes of Angus and Brahman to determine what impact the incorrect reference genome had on methylome analysis, as well as determine regions of differential methylation between the two breeds. Additionally, including reciprocal crosses enabled the elucidation of parent-of-origin-specific effects on DNA methylation. This chapter identified a small but significant quantitative bias introduced when mapping WGBS data to the incorrect reference genome and identified DMR that may be regulating *Dgat1*, a gene important for fat metabolism, in a breed and sire-of-origin-specific manner.

Following on from Chapter 4, the microRNA expression differences between Brahman, Angus and reciprocally crossed progeny were investigated to determine biological pathways that may be regulated by differentially expressed and breed-specific microRNAs (Chapter 5). MicroRNA expression correlated with mRNA expression and identified several microRNAs that may regulate genes involved in heat tolerance in Brahman cattle and fat gain in Angus cattle.

The final chapter presents a summary of the thesis and offers directions for future research. Specifically, it discusses how Chapters 3,4 and 5 could be used in genomic prediction now and what further research would be needed to improve their utility in genomic prediction tasks.

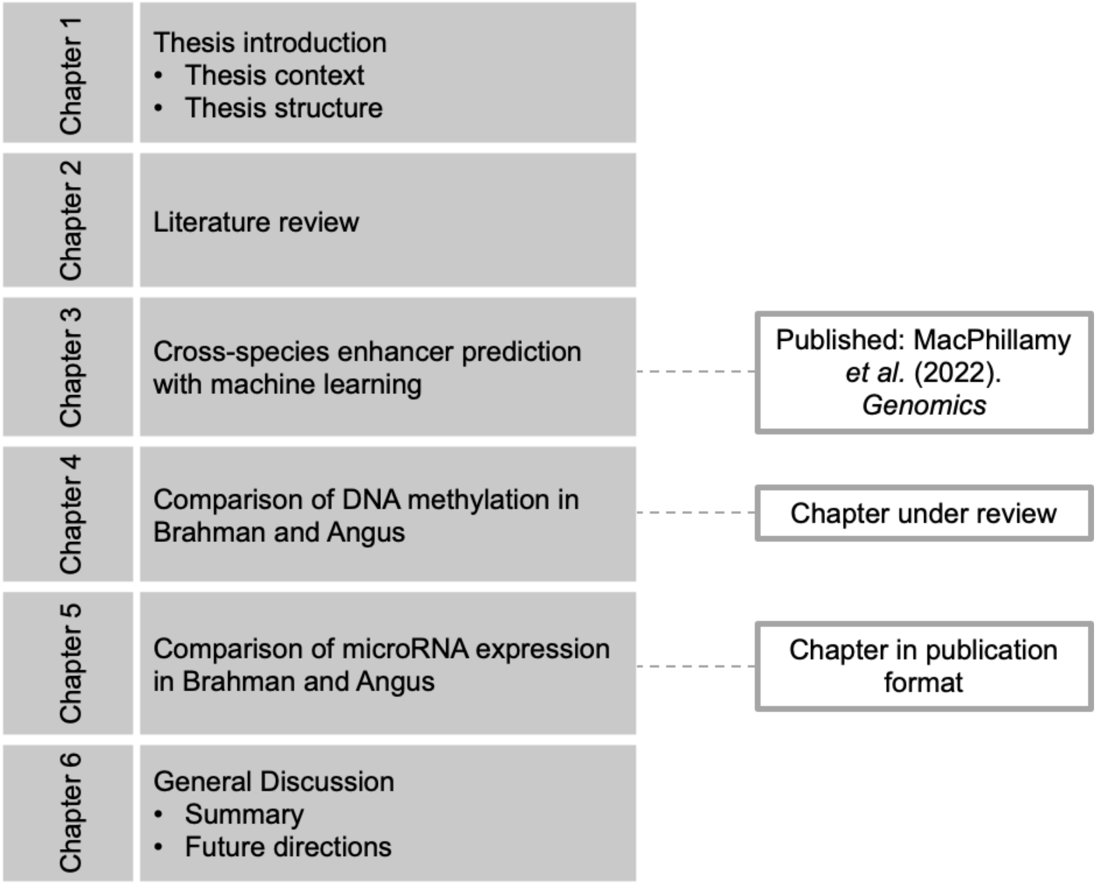

Chapter 1
Thesis introduction
• Thesis context
• Thesis structure
Chapter 2
Literature review
Chapter 3
Cross-species enhancer prediction with machine learning
Published: MacPhillamy *et al.* (2022). *Genomics*
Chapter 4
Comparison of DNA methylation in Brahman and Angus
Chapter under review
Chapter 5
Comparison of microRNA expression in Brahman and Angus
Chapter in publication format
Chapter 6
General Discussion
• Summary
• Future directions

Figure 1.1 – Overview of thesis structure.

Chapter 2: Literature Review

Introduction

The taurine and indicine cattle subspecies possess contrasting phenotypes, with Angus and Brahman cattle being two economically important cattle breeds that are representative of these distinct subspecies, respectively. Understanding the gene regulatory differences between these two cattle breeds is crucial to understanding what might be driving these distinct phenotypes. The first step in understanding what might be causing phenotypic differences between the two breeds is to investigate their gene expression differences. While this will provide insight into what genes might be responsible for breed differences, it is also necessary to gain an understanding of what is driving those gene expression differences. Both genetic and epigenetic differences can alter gene expression between Angus and Brahman cattle. Genetic elements known to regulate gene expression are enhancers and microRNAs (miRNAs), and DNA methylation is a well-known epigenetic modifier of gene expression.

This literature review introduces the central theme of this thesis – investigating the genetic and epigenetic differences between Brahman and Angus cattle to shed light on possible causes of phenotypic differences in important traits like heat tolerance and meat quality. This thesis builds on recent work that has investigated gene expression differences between Brahman and Angus cattle (Liu *et al.* 2021). Specifically, this thesis details the use of machine learning techniques and enhancer data from well-studied species like human and mouse to identify enhancers in cattle. Additionally, whole-genome bisulfite sequencing (WGBS) and miRNA sequencing were performed to investigate DNA methylation and miRNA expression differences between Brahman and Angus cattle. This literature review provides an overview of the evolutionary history of taurine and indicine cattle, the role of enhancers in gene regulation, machine learning techniques used in genomics for enhancer prediction, the role of DNA methylation in gene regulation, and finally, the role of miRNAs in modulating gene expression patterns.

The evolutionary history of taurine and indicine cattle

It is widely believed that the two main lineages of modern cattle breeds arose from two separate domestication events of the wild auroch (*Bos primigenius*) (McTavish *et al.* 2013). The first domestication event occurred in the Fertile Crescent approximately 10,000 years ago and gave rise to *Bos taurus taurus* from the wild auroch, *B. p. primigenius* (Bruford *et al.* 2003; Ajmone-Marsan *et al.* 2010; MacHugh *et al.* 2017). A more recent, subsequent domestication event occurred in the Indus Valley approximately 1,500 years later and involved *B. p. nomadicus*, which itself separated from the *B. p. primigenius* around 250-330,000 years ago (Loftus *et al.* 1994). This second domestication event gave rise to *Bos taurus indicus*. The subspecies are herein referred to as taurine and indicine cattle, respectively, where the Angus breed is representative of taurine cattle, and Brahman is representative of indicine cattle.

Angus and Brahman have contrasting phenotypes. Angus cattle have been bred for meat production and growth traits (Elzo *et al.* 2012; Low *et al.* 2020), shorter gestation length and lower calving difficulty (Casas *et al.* 2011). In contrast, Brahman cattle have superior heat and disease tolerance traits and can subsist on low-quality feed but mature more slowly (Dikmen *et al.* 2018; Goszczynski *et al.* 2018). Despite these phenotypic differences, the Angus and Brahman genomes differ by ~1% when measured by single nucleotide polymorphisms (SNPs) (Koren *et al.* 2018). This relatively high genetic similarity but divergent traits suggests the possibility that a combination of genetic and epigenetic differences between the two breeds contribute to their distinct phenotypes.

The role of enhancers in gene regulation

Enhancers are short (100-1000bp) sequences of non-coding DNA that activate the expression of target genes by recruiting transcription factors (TFs) (Claringbould & Zaugg 2021), RNA polymerase II and forming a loop with the target promoter (Panigrahi & O'Malley 2021). The role of enhancers in gene regulation is well-characterised in species like human and mouse, especially in the context of disease. For example, a study of limb development in mice observed limb malformations caused by the *Epha4* enhancer when a topologically associating domain boundary was removed, allowing *Epha4* to target genes in both domains leading to ectopic expression (Lupiáñez *et al.* 2015). Another example is seen where mutations within the *Ptf1a* enhancer have been observed to cause isolated pancreatic agenesis, where the pancreas fails to develop in utero (Weedon *et al.* 2014). There are further examples in cattle, with an investigation of the genetic causes of cholesterol deficiency in dairy cattle revealing possible enhancer activity introduced by a long-terminal repeat (Becker *et al.* 2022). Another study observed a putative enhancer contained within a 12kb copy number variant (CNV) associated with clinical mastitis resistance, milk yield and fertility in dairy cattle (Lee *et al.* 2021). These examples highlight the vital role enhancers play in orchestrating complex gene expression patterns, such as those required during limb development. Furthermore, the examples in cattle highlight the need to understand the locations of enhancers within the cattle genome. For example, by identifying a possible enhancer associated with mastitis resistance and milk yield, Lee *et al.* (2021) have provided a possible causative region for contrasting phenotypes within a dairy herd. This example and that of Becker *et al.* (2022) highlight why it is important to identify these regions within the cattle genome, as they can shed light on how different phenotypes are controlled.

Challenges in identifying enhancers

While it is clear enhancers have a critical role in maintaining the correct gene regulatory framework of an organism, they remain a challenging genomic element to identify. Unlike promoters which exist immediately upstream of their target gene (Khambata-Ford *et al.* 2003), the genomic position of an enhancer is not indicative of the gene it targets. Enhancers have been found to target genes in any orientation, up or downstream of their location (Banerji *et al.* 1981), up to one million base pairs away (Lettice *et al.* 2003), and even on different chromosomes (Geyer *et al.* 1990; Lomvardas *et al.* 2006). Moreover, enhancers can act in a tissue- and time-point-specific manner (Lupiáñez *et al.* 2015; De Vas *et al.* 2023), further complicating their identification. Despite their elusive nature, enhancers are often associated with open chromatin, and so assays like assay of transposase accessible chromatin and sequencing (ATAC-seq), DNase I hypersensitivity sequencing (DNase-seq), micrococcal nuclease sequencing (MNase-seq), and formaldehyde-assisted isolation of regulatory elements sequencing (FAIRE-seq) can be used to identify putative locations of enhancers (Cao *et al.* 2021). Furthermore, assays like chromatin immunoprecipitation and sequencing (ChIP-seq), targeting H3K27ac and H3K4me1, can also be used to identify active enhancers (Heintzman *et al.* 2007; Heintzman & Ren 2009; Creyghton *et al.* 2010; Rada-Iglesias *et al.* 2011).

Enhancers in cattle

Relatively little is known about enhancers in cattle compared to human and mouse and there are several enhancer databases available for human and mouse (Visel *et al.* 2007; Andersson *et al.* 2014; Fishilevich *et al.* 2017; Wang *et al.* 2018; Gao & Qian 2019; Consortium *et al.* 2020). However, none currently exist for cattle, despite more and more studies being completed aimed at identifying these regions (Alexandre *et al.* 2021; Cao *et al.* 2021; Forutan *et al.* 2021; Kern *et al.* 2021; Prowse-Wilkins *et al.* 2021; Prowse-Wilkins *et al.* 2022). While the increasing number of studies investigating enhancers in cattle is valuable to the community, some breeds such as those of indicine origin lack information on enhancers. As more breed-specific cattle genomes become available, there is an increasing need for tools that can identify enhancers within all cattle breeds as well as other livestock species.

Machine learning for enhancer prediction

Several methods have emerged to leverage the plethora of data available in model organisms like human and mouse to try and accelerate our understanding of where enhancers might be within the cattle genome. Recently, machine learning has emerged as a powerful tool for predicting enhancers from the DNA sequence alone (Firpi *et al.* 2010; Fletez-Brant *et al.* 2013; Ghandi *et al.* 2016; Liu *et al.* 2016; Liu *et al.* 2018; Nguyen *et al.* 2019; Kelley 2020; Cai *et al.* 2021; Inayat *et al.* 2021; Yang *et al.* 2021b; Zeng *et al.* 2021; Butt *et al.* 2022). Unfortunately, many of these studies have focused on predicting enhancers within the human genome, with few examining their performance in predicting enhancers from human or mouse to less well-studied species (Chen *et al.* 2018; Minnoye *et al.* 2020; Hong *et al.* 2021). Furthermore, none have evaluated models and different DNA representations in identifying enhancers within the cattle genome. Given their important role in regulating gene expression, the volume of enhancer data available in human and mouse, and the multitude of machine learning methods developed for enhancer prediction, there is an opportunity to evaluate which machine learning method is best suited to identify these essential regulatory regions within the cattle genome.

DNA methylation and its role in gene regulation

Epigenetic modifications like DNA methylation are heritable components that alter the accessibility of the genome to transcriptional machinery without changing the underlying sequence (Handy *et al.* 2011). DNA methylation is perhaps the most studied epigenetic modification, with vital roles in regulating gene expression, repression of transposable elements, and parental chromosome-specific regulation through genomic imprinting and X-chromosome inactivation (Li & Zhang 2014; Jansz 2019). DNA methylation within the mammalian genome typically refers to the methylation of a cytosine-phosphate-guanine (CpG) dinucleotide (Ramsahoye *et al.* 2000; Ziller *et al.* 2011). DNA methylation in the CpG context is by far the most common in mammals (Ramsahoye *et al.* 2000; Ziller *et al.* 2011), though it should be noted that other methylation contexts (CH and CHH, where H = A, C, T) are being increasingly examined for their role in neuronal development (Xie *et al.* 2012; Guo *et al.* 2014). DNA methylation involves the transfer of a methyl group to the C5 position of the cytosine to form 5-methylcytosine (Moore *et al.* 2013). Regardless of the methylation context, DNA methylation is known mainly as a repressive epigenetic modification, with regions showing a high degree of methylation (hypermethylation) being transcriptionally inactive and those with lower methylation (hypomethylation) being active (Razin & Cedar 1991; Bird 2002; Bommarito & Fry 2019). DNA methylation can inhibit gene expression through indirect and direct mechanisms. Indirect inhibition occurs when methylated DNA is preferentially bound by proteins that contain a methylated DNA binding domain which then physically blocks the binding of TFs (Hendrich & Bird 1998; Bird & Wolffe 1999). The direct mechanism with which DNA methylation induces gene silencing is by altering the TF binding site, whereby the addition of the methyl group to the cytosine directly blocks the binding of TFs leading to gene silencing (Bird 2002). Hence, DNA methylation can lead to phenotypic differences among individuals through the differential repression of transcription.

DNA methylation in cattle

Given that DNA methylation has a role in silencing gene expression, differential methylation between Brahman and Angus may shed light on epigenomic changes responsible for the contrasting phenotypes observed between these two breeds, with DNA methylation implicated in several cattle traits. For example, a study comparing DNA methylation levels between tender and tough meat from indicine cattle revealed higher levels of DNA methylation around the GNAS complex locus (*GNAS*) and EBF transcription factor 3 (*Ebf3*) gene in the tender group (de Souza *et al.* 2022). The authors posit that differences in methylation between the tough and tender groups affecting the *G* protein signalling pathway and *Ebf3* gene, both of which are involved in muscle homeostasis, may be contributing to the differences in meat tenderness. Similarly, a comparison of heat stress response between Angus (taurine) and Nellore (indicine) revealed that genes and pathways involved in stress response and cellular defence were more often hypomethylated and, thus, active, in Nellore cattle compared to Angus cattle (Del Corvo *et al.* 2021). Together, these examples highlight how investigating DNA methylation differences between contrasting phenotypes can be used to identify regions of interest that may be contributing to the trait of interest.

Breed and parent-of-origin effects

Breed effects, as the name suggests, are those associated with a particular breed, and for offspring to have those effects, both parents must be from the same breed or genetic background. Parent-of-origin effects (POEs), on the other hand, are those associated with the genetics of either the mother or the father and the phenotype observed in the offspring depends on which parent contributed the allele (Lawson *et al.* 2013). An example of POEs would be if the progeny of an Angus dam and Brahman sire had the same muscle amount at birth as offspring born to an Angus dam and bull. It has been reported that maternally inherited genes disproportionately contribute to myofiber development in reciprocally crossed Angus and Brahman cattle (Xiang *et al.* 2013), suggesting that the breed of the dam is an important determinant of myofiber development, regardless of the breed of the sire. Similarly, a recent study of the effect of sire breed on meat quality found that sire breed significantly impacts meat quality (Cafferky *et al.* 2019), suggesting sire genetics are important for meat traits, regardless of the dam's genetics. These examples illustrate the importance of considering not only how the breed's genetics affects traits but also how the genetics of a particular parent can affect traits.

Genetics of epigenetics

As DNA methylation often occurs in the CpG context, a single nucleotide variant (SNV) can completely erase a methylation site. Moreover, the spontaneous deamination of a methylated CpG to a TpG is the most common dinucleotide mutation in the mammalian genome (Żemojtel *et al.* 2011; Yang *et al.* 2021a), and has even been suggested to produce novel transcription factor binding sites with high-efficiency (Żemojtel *et al.* 2011). This spontaneous deamination is significant when trying to evaluate genome-wide methylation levels of an individual due to the possibility that an individual loses CpG sites relative to the reference. This results in a locus being classified as unmethylated when it has no ability to be methylated. Structural variants (SVs) can also add or remove CpG sites, further impacting DNA methylation. These sequence variants can complicate epigenomic analyses, as sequencing reads may not map unambiguously, leading to spurious estimates of DNA methylation. Additionally, the loss of CpG in Brahman compared to Angus may be erroneously reported as a differentially methylated site when no shared methylation site exists between the two breeds. Lastly, in cases where individuals possess an insertion SV that introduces CpG sites, sequencing reads originating from this region may not be correctly mapped to the reference, leading to inaccurate estimates of DNA methylation. These examples are used to demonstrate how the changing, addition or removal of sequence (genetics) can impact the methylation (epigenetics) status of that region and thus illustrates the "genetics of epigenetics".

MicroRNAs and their role in gene regulation

MiRNAs are a class of small (~22bp) non-coding RNA that can have major impacts on gene expression via post-transcriptional gene silencing (Kim *et al.* 2008; Ha & Kim 2014; O'Brien *et al.* 2018). MiRNAs are strongly implicated in a diverse array of biological processes, such as metabolism (Rottiers & Näär 2012) and responding to the external environment (Vrijens *et al.* 2015; Liu *et al.* 2020). It is generally accepted that miRNAs are conserved across species (Macfarlane & Murphy 2010), though there are also species- and tissue-specific miRNA expression patterns (Jopling 2012; Sun *et al.* 2014). MiRNAs control gene expression via Watson-Crick base pairing between the miRNA-induced silencing complex (miRISC) and the three prime untranslated region (3’UTR) of the mRNA target (Eulalio *et al.* 2008; Bartel 2009; Fabian *et al.* 2010). The degree of complementarity between the miRNA and mRNA determines what silencing mechanism will be used, i.e., complete complementarity leads to mRNA cleavage, and partial complementarity leads to translational repression (Jo *et al.* 2015). This tolerance for incomplete base pairing gives a single miRNA the ability to target multiple mRNAs and thus also means a single mRNA can be targeted by numerous miRNAs (Peterson *et al.* 2014). Furthermore, SNVs have been demonstrated to alter miRNA-mRNA target specificity (Sun *et al.* 2009). The role of miRNAs in modulating a diverse array of biological processes, the ability of a single miRNA to target multiple mRNAs and the potential for SNVs to alter miRNA-mRNA target specificity highlights the importance of understanding miRNA expression differences between contrasting phenotypes, like those exhibited by Brahman and Angus cattle.

MicroRNAs in cattle

Many studies have investigated the role of miRNAs in gene regulation in cattle (Huang *et al.* 2011; Muroya *et al.* 2013; Li *et al.* 2014; Sun *et al.* 2014; Sengar *et al.* 2018; Gao *et al.* 2020; Kumar *et al.* 2021; Pacífico *et al.* 2022). Some examples include a recent study that compared rumen epithelial response to dietary changes and observed eight differentially expressed miRNAs between forage diet and high-grain diet, with an enrichment of genes associated with tricarboxylic acid and short-chain fatty acid metabolism (Pacífico *et al.* 2022). Another study identified circulating plasma miRNAs that were correlated with health, welfare and production performance traits like fertility and telomere length (Ioannidis *et al.* 2018). However, few have examined them in the context of differences between *B. taurus* and *B. indicus*. Deb and Sengar (2021) investigated miRNA expression differences between an indicine breed (Sahiwal) and an indicine-taurine hybrid (Frieswal) in response to heat tolerance. Similarly, Dong *et al.* (2023) investigated differential miRNA expression between Mongolian (taurine) and Hainan (indicine) cattle testes. While this was a comparison of taurine and indicine cattle, the tissue used is only likely to reflect fertility differences between the two subspecies. Despite miRNAs being relatively well studied in cattle, a comparison between taurine and indicine cattle using an essential metabolic organ like the liver is yet to be made. Given the role of the liver in many production traits, an investigation into the miRNA expression profiles of this organ can shed light on the causes of gene regulatory differences driving the contrasting phenotypes of Brahman and Angus cattle.

Contrasting phenotypes first appear in utero

Fetal development is a complex process that relies on the careful orchestration of gene expression. This orchestra is conducted via genetics and epigenetics like enhancers, DNA methylation and non-coding RNAs, such as miRNAs (Zhu *et al.* 2021). Extensive epigenetic reprogramming occurs in the transition from differentiated gametes to a totipotent embryo (Zhu *et al.* 2021). Once the pregnancy has been established, the growth rate is generally linear until around day 153 (~5 months), when it takes on a log form (Reynolds *et al.* 1990; Krog *et al.* 2018); this is comparable to humans (Kiserud *et al.* 2017), which share a similar gestation period to cattle. It is at this time point that phenotypic differences between breeds first appear (Xiang *et al.* 2013). Furthermore, several developmental processes important to beef cattle are underway at this time point, specifically, secondary myogenesis and adipogenesis.

Myogenesis is the formation of muscle cells which begins during the embryonic stage of life and is crucial to meat breeds like Brahman and Angus as there is no increase in the number of muscle fibres after birth (Stickland 1978; Zhu *et al.* 2006). Secondary myogenesis is underway at day 153 (Du *et al.* 2010), and this is where the bulk of skeletal muscle fibres are created (Beermann *et al.* 1978). Additionally, adipogenesis begins around this developmental stage (Du *et al.* 2010). Adipogenesis is the formation of adipocytes, is crucial to the palatability and value of the meat, and it differs from myogenesis in that while it starts in utero, the number of adipocytes continues to grow until around 250 days post birth (Du *et al.* 2015). As these critical processes begin in utero and phenotypic differences are observable at this time point (Xiang *et al.* 2013), investigating genetic and epigenetic differences between Brahman and Angus at day 153 has the potential to illuminate what gene regulatory differences may be occurring between these two breeds.

Potential applications for the beef industry

The beef cattle industry in Australia contributes ~$23.1 billion to the national economy (MLA 2022). Forecasts suggest that domestic utilisation of beef products will increase from 597,000 tonnes in 2023 to 633,000 tonnes by 2025, with beef consumption per capita in Australia expected to increase over the same period (MLA 2023). The national beef herd is broadly split into two regions, the Northern and

Southern herds. The Northern herd is predominantly made up of tropically adapted cattle like Brahman, where they are favoured for their hardiness, heat tolerance and ability to thrive on low-quality feed but produce low-value meat as a result (PwC 2011). Also present in the Northern herd are indicus-taurine crosses and composite breeds. Conversely, the Southern herd comprises European breeds like Angus, where the higher quality feed and more moderate climates allow the cattle to gain weight quickly and produce high-value beef (PwC 2011). The inability of these European breeds to maintain production in adverse conditions is a substantial concern to the beef industry, as high-quality beef is required to maintain the economic value of the industry. As a result, there is a need to understand how traits like heat tolerance, fat, and muscle gain are differentially regulated in these breeds. Improving our understanding of how these traits are controlled in the pure-bred animals will benefit efforts to create efficient hybrid animals. This understanding of the underlying biology will potentially help the industry maintains and improve productivity in increasingly unstable markets and climates.

Scope and aims of this thesis

This thesis presented an opportunity to investigate possible genetic and epigenetic differences between Brahman and Angus cattle at a critical developmental time point that could shed light on the gene expression differences that give rise to the distinct phenotypes. Questions raised and discussed in this thesis included:

1. *Can human and mouse enhancer data be used to develop a machine-learning model to predict enhancers in cattle?*

There is an opportunity to evaluate and compare numerous machine-learning models and DNA representations to determine the best combination for accurate cross-species enhancer prediction. In this thesis, I compared nine machine-learning models and four DNA representations. I evaluated their cross-species enhancer prediction performance using a highly curated enhancer dataset from human and mouse and publicly available ChIP-seq data for cattle, pig, and dog (Chapter 2). Pig and dog were selected to assess the models' performance on a variety of mammalian genomes, as well as to provide a least three replicates to ensure performance on the cattle genome was not an anomaly. This chapter aimed to determine how well enhancers could be predicted from one species to another to help elucidate the genomic location of possible regulatory elements.

2. *Do DNA methylation differences between the breeds and parent-of-origin comparisons during a critical developmental time point reveal regulatory regions that may relate to phenotypic differences between Brahman and Angus?*

Key developmental pathways like myogenesis and adipogenesis are underway at day 153 in Brahman and Angus. These pathways are responsible for many of the desirable carcass traits in these breeds. Understanding what DNA methylation differences exist between Brahman and Angus at this developmental timepoint can shed light on possible causes of phenotypic differences that are observed between Brahman and Angus adults. Furthermore, investigating parent-of-origin effects and their role in contributing to differential methylation of important regulatory regions can shed light on the importance of parental genetics in contributing to the progeny's phenotype (Chapter 3).

3. *Does differential miRNA expression between breeds and parent-of-origin group comparisons during a critical developmental time point reveal candidate miRNAs that could contribute to phenotypic differences?*

MiRNAs are known to play a critical role in fine-tuning the expression of genes related to important developmental pathways during fetal development. Understanding what miRNAs are differentially expressed between breeds can shed light on post-transcriptional gene regulation that may influence breed differences between Brahman and Angus. Moreover, investigating parent-of-origin effects and whether this contributes to differential miRNA expression may shed light on the importance of parental genetics in contributing to the phenotype of the offspring (Chapter 4).

References

Ajmone-Marsan P., Garcia J.F. & Lenstra J.A. (2010) On the origin of cattle: How aurochs became cattle and colonized the world. Evolutionary Anthropology: Issues, News, and Reviews 19, 148-57. https://doi.org/10.1002/evan.20267

Alexandre P.A., Naval-Sánchez M., Menzies M., Nguyen L.T., Porto-Neto L.R., Fortes M.R.S. & Reverter A. (2021) Chromatin accessibility and regulatory vocabulary across indicine cattle tissues. Genome Biology 22, 273. 10.1186/s13059-021-02489-7

Andersson R., Gebhard C., Miguel-Escalada I., Hoof I., Bornholdt J., Boyd M., Chen Y., Zhao X., Schmidl C., Suzuki T., Ntini E., Arner E., Valen E., Li K., Schwarzfischer L., Glatz D., Raithel J., Lilje B., Rapin N., Bagger F.O., Jørgensen M., Andersen P.R., Bertin N., Rackham O., Burroughs A.M., Baillie J.K., Ishizu Y., Shimizu Y., Furuhata E., Maeda S., Negishi Y., Mungall C.J., Meehan T.F., Lassmann T., Itoh M., Kawaji H., Kondo N., Kawai J., Lennartsson A., Daub C.O., Heutink P., Hume D.A., Jensen T.H., Suzuki H., Hayashizaki Y., Müller F., Forrest A.R.R., Carninci P., Rehli M., Sandelin A. & The F.C. (2014) An atlas of active enhancers across human cell types and tissues. Nature 507, 455-61. 10.1038/nature12787

Banerji J., Rusconi S. & Schaffner W. (1981) Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299-308. https://doi.org/10.1016/0092-8674(81)90413-X

Bartel D.P. (2009) MicroRNAs: target recognition and regulatory functions. Cell 136, 215-33.

Becker D., Weikard R., Heimes A., Hadlich F., Hammon H.M., Meyerholz M.M., Petzl W., Zerbe H., Schuberth H.-J., Hoedemaker M., Schmicke M., Engelmann S. & Kühn C. (2022) Allele-biased expression of the bovine APOB gene associated with the cholesterol deficiency defect suggests cis-regulatory enhancer effects of the LTR retrotransposon insertion. Scientific Reports 12, 13469. 10.1038/s41598-022-17798-5

Beermann D., Cassens R. & Hausman G. (1978) A second look at fiber type differentiation in porcine skeletal muscle. Journal of Animal Science 46, 125-32.

Bird A. (2002) DNA methylation patterns and epigenetic memory. Genes & Development 16, 6-21. 10.1101/gad.947102

Bird A.P. & Wolffe A.P. (1999) Methylation-induced repression—belts, braces, and chromatin. Cell 99, 451-4.

Bommarito P.A. & Fry R.C. (2019) Chapter 2-1 - The Role of DNA Methylation in Gene Regulation. In: Toxicoepigenetics (ed. by S.D. McCullough & D.C. Dolinoy), pp. 127-51. Academic Press.

Bruford M.W., Bradley D.G. & Luikart G. (2003) DNA markers reveal the complexity of livestock domestication. Nature Reviews Genetics 4, 900-10. 10.1038/nrg1203

Butt A.H., Alkhalifah T., Alturise F. & Khan Y.D. (2022) A machine learning technique for identifying DNA enhancer regions utilizing CIS-regulatory element patterns. Scientific Reports 12, 15183. 10.1038/s41598-022-19099-3

Cafferky J., Hamill R.M., Allen P., O'Doherty J.V., Cromie A. & Sweeney T. (2019) Effect of Breed and Gender on Meat Quality of M. longissimus thoracis et lumborum Muscle from Crossbred Beef Bulls and Steers. Foods 8, 173.

Cai L.J., Ren X.B., Fu X.Z., Peng L., Gao M.Y. & Zeng X.X. (2021) iEnhancer-XG: interpretable sequence-based enhancers and their strength predictor. Bioinformatics 37, 1060-7. 10.1093/bioinformatics/btaa914

Cao X., Cheng J., Huang Y., Lan X., Lei C. & Chen H. (2021) Comparative Enhancer Map of Cattle Muscle Genome Annotated by ATAC-Seq. Frontiers in Veterinary Science 8. 10.3389/fvets.2021.782409

Casas E., Thallman R. & Cundiff L. (2011) Birth and weaning traits in crossbred cattle from Hereford, Angus, Brahman, Boran, Tuli, and Belgian Blue sires. Journal of Animal Science 89, 979-87.

Chen L., Fish A.E. & Capra J.A. (2018) Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. PLoS Computational Biology 14, e1006484. 10.1371/journal.pcbi.1006484

Claringbould A. & Zaugg J.B. (2021) Enhancers in disease: molecular basis and emerging treatment strategies. Trends in Molecular Medicine 27, 1060-73. 10.1016/j.molmed.2021.07.012

Consortium E.P., Moore J.E., Purcaro M.J., Pratt H.E., Epstein C.B., Shoresh N., Adrian J., Kawli T., Davis C.A., Dobin A., Kaul R., Halow J., Van Nostrand E.L., Freese P., Gorkin D.U., Shen Y., He Y., Mackiewicz M., Pauli-Behn F., Williams B.A., Mortazavi A., Keller C.A., Zhang X.O., Elhajjajy S.I., Huey J., Dickel D.E., Snetkova V., Wei X., Wang X., Rivera-Mulia J.C., Rozowsky J., Zhang J., Chhetri S.B., Zhang J., Victorsen A., White K.P., Visel A., Yeo G.W., Burge C.B., Lecuyer E., Gilbert D.M., Dekker J., Rinn J., Mendenhall E.M., Ecker J.R., Kellis M., Klein R.J., Noble W.S., Kundaje A., Guigo R., Farnham P.J., Cherry J.M., Myers R.M., Ren B., Graveley B.R., Gerstein M.B., Pennacchio L.A., Snyder M.P., Bernstein B.E., Wold B., Hardison R.C., Gingeras T.R., Stamatoyannopoulos J.A. & Weng Z. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699-710. 10.1038/s41586-020-2493-4

Creyghton M.P., Cheng A.W., Welstead G.G., Kooistra T., Carey B.W., Steine E.J., Hanna J., Lodato M.A., Frampton G.M., Sharp P.A., Boyer L.A., Young R.A. & Jaenisch R. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences 107, 21931-6. 10.1073/pnas.1016071107

de Souza M.M., Niciura S.C.M., Rocha M.I.P., Pan Z., Zhou H., Bruscadin J.J., da Silva Diniz W.J., Afonso J., de Oliveira P.S.N., Mourao G.B., Zerlotini A., Coutinho L.L., Koltes J.E. & de Almeida Regitano L.C. (2022) DNA methylation may affect beef tenderness through signal transduction in Bos indicus. Epigenetics Chromatin 15, 15. 10.1186/s13072-022-00449-4

De Vas M.G., Boulet F., Joshi S.S., Garstang M.G., Khan T.N., Atla G., Parry D., Moore D., Cebola I., Zhang S., Cui W., Lampe A.K., Lam W.W., Ferrer J., Pradeepa M.M. & Atanur S.S. (2023) Regulatory de novo mutations underlying intellectual disability. Life Sci Alliance 6. 10.26508/lsa.202201843

Deb R. & Sengar G.S. (2021) Comparative miRNA signatures among Sahiwal and Frieswal cattle breeds during summer stress. 3 Biotech 11, 79. 10.1007/s13205-020-02608-4

Del Corvo M., Lazzari B., Capra E., Zavarez L., Milanesi M., Utsunomiya Y.T., Utsunomiya A.T.H., Stella A., de Paula Nogueira G., Garcia J.F. & Ajmone-

Marsan P. (2021) Methylome Patterns of Cattle Adaptation to Heat Stress. Frontiers in Genetics 12. 10.3389/fgene.2021.633132

Dikmen S., Mateescu R.G., Elzo M.A. & Hansen P.J. (2018) Determination of the optimum contribution of Brahman genetics in an Angus-Brahman multibreed herd for regulation of body temperature during hot weather. Journal of Animal Science 96, 2175-83. 10.1093/jas/sky133

Dong Z., Ning Q., Liu Y., Wang S., Wang F., Luo X., Chen N. & Lei C. (2023) Comparative transcriptomics analysis of testicular miRNA from indicine and taurine cattle. Animal Biotechnology 34, 1436-46. 10.1080/10495398.2022.2029466

Du M., Tong J., Zhao J., Underwood K.R., Zhu M., Ford S.P. & Nathanielsz P.W. (2010) Fetal programming of skeletal muscle development in ruminant animals1. Journal of Animal Science 88, E51-E60. 10.2527/jas.2009-2311

Du M., Wang B., Fu X., Yang Q. & Zhu M.-J. (2015) Fetal programming in meat production. Meat Science 109, 40-7. https://doi.org/10.1016/j.meatsci.2015.04.010

Elzo M.A., Johnson D.D., Wasdin J.G. & Driver J.D. (2012) Carcass and meat palatability breed differences and heterosis effects in an Angus–Brahman multibreed population. Meat Science 90, 87-92. https://doi.org/10.1016/j.meatsci.2011.06.010

Eulalio A., Huntzinger E. & Izaurralde E. (2008) Getting to the root of miRNA-mediated gene silencing. Cell 132, 9-14.

Fabian M.R., Sonenberg N. & Filipowicz W. (2010) Regulation of mRNA translation and stability by microRNAs. Annual Review of Biochemistry 79, 351-79.

Firpi H.A., Ucar D. & Tan K. (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics 26, 1579-86. 10.1093/bioinformatics/btq248

Fishilevich S., Nudel R., Rappaport N., Hadar R., Plaschkes I., Iny Stein T., Rosen N., Kohn A., Twik M., Safran M., Lancet D. & Cohen D. (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford) 2017. 10.1093/database/bax028

Fletez-Brant C., Lee D., Mccallion A.S. & Beer M.A. (2013) kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic Acids Research 41, W544-W56. 10.1093/nar/gkt519

Forutan M., Vander Jagt C., Ross E., Chamberlain A., Mason B., Nguyen L., Moore S., Garner J., Xiang R. & Hayes B. (2021) Genome wide analysis of bovine enhancers and promoters across developmental stages in liver. In: *Proc. Assoc. Advmt. Anim. Breed. Genet*, pp. 126-30.

Gao T. & Qian J. (2019) EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Research 48, D58-D64. 10.1093/nar/gkz980

Gao Y., Wu F., Ren Y.X., Zhou Z.H., Chen N.B., Huang Y.Z., Lei C.Z., Chen H. & Dang R.H. (2020) MiRNAs Expression Profiling of Bovine (Bos taurus) Testes and Effect of bta-miR-146b on Proliferation and Apoptosis in Bovine Male Germline Stem Cells. International Journal of Molecular Sciences 21. 10.3390/ijms21113846

Geyer P.K., Green M.M. & Corces V.G. (1990) Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in Drosophila. The EMBO Journal 9, 2247-56. https://doi.org/10.1002/j.1460-2075.1990.tb07395.x

Ghandi M., Mohammad-Noori M., Ghareghani N., Lee D., Garraway L. & Beer M.A. (2016) gkmSVM: an R package for gapped-kmer SVM. Bioinformatics 32, 2205-7. 10.1093/bioinformatics/btw203

Goszczynski D.E., Corbi-Botto C.M., Durand H.M., Rogberg-Muñoz A., Munilla S., Peral-Garcia P., Cantet R.J.C. & Giovambattista G. (2018) Evidence of positive selection towards Zebuine haplotypes in the BoLA region of Brangus cattle. Animal 12, 215-23. https://doi.org/10.1017/S1751731117001380

Guo J.U., Su Y., Shin J.H., Shin J., Li H., Xie B., Zhong C., Hu S., Le T., Fan G., Zhu H., Chang Q., Gao Y., Ming G.-l. & Song H. (2014) Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nature Neuroscience 17, 215-22. 10.1038/nn.3607

Ha M. & Kim V.N. (2014) Regulation of microRNA biogenesis. Nature Reviews Molecular Cell Biology 15, 509-24. 10.1038/nrm3838

Handy D.E., Castro R. & Loscalzo J. (2011) Epigenetic modifications: basic mechanisms and role in cardiovascular disease. Circulation 123, 2145-56. 10.1161/circulationaha.110.956839

Heintzman N.D. & Ren B. (2009) Finding distal regulatory elements in the human genome. Current Opinion in Genetics & Development 19, 541-9. 10.1016/j.gde.2009.09.006

Heintzman N.D., Stuart R.K., Hon G., Fu Y., Ching C.W., Hawkins R.D., Barrera L.O., Van Calcar S., Qu C., Ching K.A., Wang W., Weng Z., Green R.D., Crawford G.E. & Ren B. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics 39, 311-8. 10.1038/ng1966

Hendrich B. & Bird A. (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. Molecular and Cellular Biology 18, 6538-47. 10.1128/mcb.18.11.6538

Hong J., Gao R. & Yang Y. (2021) CrepHAN: cross-species prediction of enhancers by using hierarchical attention networks. Bioinformatics. doi:10.1093/bioinformatics/btab349

Huang J.M., Ju Z.H., Li Q.L., Hou Q.L., Wang C.F., Li J.B., Li R.L., Wang L.L., Sun T., Hang S.Q., Gao Y.D., Hou M.H. & Zhong J.F. (2011) Solexa Sequencing of Novel and Differentially Expressed MicroRNAs in Testicular and Ovarian Tissues in Holstein Cattle. International Journal of Biological Sciences 7, 1016-26. 10.7150/ijbs.7.1016

Inayat N., Khan M., Iqbal N., Khan S., Raza M., Khan D.M., Khan A. & Wei D.Q. (2021) iEnhancer-DHF: Identification of Enhancers and Their Strengths Using Optimize Deep Neural Network With Multiple Features Extraction Methods. IEEE Access 9, 40783-96. 10.1109/access.2021.3062291

Ioannidis J., Sánchez-Molano E., Psifidi A., Donadeu F.X. & Banos G. (2018) Association of plasma microRNA expression with age, genetic background and functional traits in dairy cattle. Scientific Reports 8, 12955. 10.1038/s41598-018-31099-w

Jansz N. (2019) DNA methylation dynamics at transposable elements in mammals. Essays in Biochemistry 63, 677-89. 10.1042/ebc20190039

Jo Myung H., Shin S., Jung S.-R., Kim E., Song J.-J. & Hohng S. (2015) Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs. Molecular Cell 59, 117-24. 10.1016/j.molcel.2015.04.027

Jopling C. (2012) Liver-specific microRNA-122: Biogenesis and function. RNA Biol 9, 137-42. 10.4161/rna.18827

Kelley D.R. (2020) Cross-species regulatory sequence activity prediction. PLoS Computational Biology 16, e1008050. 10.1371/journal.pcbi.1008050

Kern C., Wang Y., Xu X., Pan Z., Halstead M., Chanthavixay G., Saelao P., Waters S., Xiang R., Chamberlain A., Korf I., Delany M.E., Cheng H.H., Medrano J.F., Van Eenennaam A.L., Tuggle C.K., Ernst C., Flicek P., Quon G., Ross P. & Zhou H. (2021) Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nature Communications 12, 1821. 10.1038/s41467-021-22100-8

Khambata-Ford S., Liu Y., Gleason C., Dickson M., Altman R.B., Batzoglou S. & Myers R.M. (2003) Identification of promoter regions in the human genome by using a retroviral plasmid library-based functional reporter gene assay. Genome Research 13, 1765-74. 10.1101/gr.529803

Kim D.H., Sætrom P., Snøve O. & Rossi J.J. (2008) MicroRNA-directed transcriptional gene silencing in mammalian cells. Proceedings of the National Academy of Sciences 105, 16230-5. doi:10.1073/pnas.0808830105

Kiserud T., Piaggio G., Carroli G., Widmer M., Carvalho J., Neerup Jensen L., Giordano D., Cecatti J.G., Abdel Aleem H., Talegawkar S.A., Benachi A., Diemert A., Tshefu Kitoto A., Thinkhamrop J., Lumbiganon P., Tabor A., Kriplani A., Gonzalez Perez R., Hecher K., Hanson M.A., Gülmezoglu A.M. & Platt L.D. (2017) The World Health Organization Fetal Growth Charts: A Multinational Longitudinal Study of Ultrasound Biometric Measurements and Estimated Fetal Weight. PLoS Med 14, e1002220. 10.1371/journal.pmed.1002220

Koren S., Rhie A., Walenz B.P., Dilthey A.T., Bickhart D.M., Kingan S.B., Hiendleder S., Williams J.L., Smith T.P.L. & Phillippy A.M. (2018) De novo assembly of haplotype-resolved genomes with trio binning. Nature Biotechnology 36, 1174-82. 10.1038/nbt.4277

Krog C.H., Agerholm J.S. & Nielsen S.S. (2018) Fetal age assessment for Holstein cattle. PLOS ONE 13, e0207682. 10.1371/journal.pone.0207682

Kumar M., Noyonika, Aggarwal A. & Kaul G. (2021) Novel and known miRNAs in zebu (Tharparkar) and crossbred (Karan-Fries) cattle under heat stress. Functional & Integrative Genomics 21, 405-19. 10.1007/s10142-021-00785-w

Lawson H.A., Cheverud J.M. & Wolf J.B. (2013) Genomic imprinting and parent-of-origin effects on complex traits. Nature Reviews: Genetics 14, 609-17. 10.1038/nrg3543

Lee Y.-L., Takeda H., Costa Monteiro Moreira G., Karim L., Mullaart E., Coppieters W., The Gplus E.c., Appeltant R., Veerkamp R.F., Groenen M.A.M., Georges M., Bosse M., Druet T., Bouwman A.C. & Charlier C. (2021) A 12 kb multi-allelic copy number variation encompassing a GC gene enhancer is associated with mastitis resistance in dairy cattle. PLoS Genetics 17, e1009331. 10.1371/journal.pgen.1009331

Lettice L.A., Heaney S.J.H., Purdie L.A., Li L., de Beer P., Oostra B.A., Goode D., Elgar G., Hill R.E. & de Graaff E. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics 12, 1725-35.

Li E. & Zhang Y. (2014) DNA methylation in mammals. Cold Spring Harb Perspect Biol 6, a019133. 10.1101/cshperspect.a019133

Li Z.X., Wang H.L., Chen L., Wang L.J., Liu X.L., Ru C.X. & Song A.L. (2014) Identification and characterization of novel and differentially expressed microRNAs in peripheral blood from healthy and mastitis Holstein cattle by deep sequencing. Animal Genetics 45, 20-7. 10.1111/age.12096

Liu B., Fang L.Y., Long R., Lan X. & Chou K.C. (2016) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32, 362-9. 10.1093/bioinformatics/btv604

Liu B., Li K., Huang D.S. & Chou K.C. (2018) iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinformatics 34, 3835-42. 10.1093/bioinformatics/bty458

Liu G., Liao Y., Sun B., Guo Y., Deng M., Li Y. & Liu D. (2020) Effects of chronic heat stress on mRNA and miRNA expressions in dairy cows. Gene 742, 144550. https://doi.org/10.1016/j.gene.2020.144550

Liu R., Tearle R., Low W.Y., Chen T., Thomsen D., Smith T.P.L., Hiendleder S. & Williams J.L. (2021) Distinctive gene expression patterns and imprinting signatures revealed in reciprocal crosses between cattle sub-species. BMC Genomics 22. 10.1186/s12864-021-07667-2

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 91, 2757-61. doi:10.1073/pnas.91.7.2757

Lomvardas S., Barnea G., Pisapia D.J., Mendelsohn M., Kirkland J. & Axel R. (2006) Interchromosomal Interactions and Olfactory Receptor Choice. Cell 126, 403-13. 10.1016/j.cell.2006.06.035

Low W.Y., Tearle R., Liu R., Koren S., Rhie A., Bickhart D.M., Rosen B.D., Kronenberg Z.N., Kingan S.B. & Tseng E. (2020) Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. Nature Communications 11, 1-14.

Lupiáñez D.G., Kraft K., Heinrich V., Krawitz P., Brancati F., Klopocki E., Horn D., Kayserili H., Opitz J.M. & Laxova R. (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161, 1012-25.

Macfarlane L.A. & Murphy P.R. (2010) MicroRNA: Biogenesis, Function and Role in Cancer. Current Genomics 11, 537-61. 10.2174/138920210793175895

MacHugh D.E., Larson G. & Orlando L. (2017) Taming the Past: Ancient DNA and the Study of Animal Domestication. Annual Review of Animal Biosciences 5, 329-51. 10.1146/annurev-animal-022516-022747

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 110, E1398-E406. doi:10.1073/pnas.1303367110

Minnoye L., Taskiran, II, Mauduit D., Fazio M., Van Aerschot L., Hulselmans G., Christiaens V., Makhzami S., Seltenhammer M., Karras P., Primot A., Cadieu E., van Rooijen E., Marine J.C., Egidy G., Ghanem G.E., Zon L., Wouters J. & Aerts S. (2020) Cross-species analysis of enhancer logic using deep learning. Genome Research 30, 1815-34. 10.1101/gr.260844.120

MLA (2022) State of the Industry Report: The Australian red meat and livestock industry. pp. 1-38. Meat & Livestock Australia.

MLA (2023) Industry Projections 2023: Australian cattle. pp. 1-9. Meat & Livestock Australia.

Moore L.D., Le T. & Fan G. (2013) DNA Methylation and Its Basic Function. Neuropsychopharmacology 38, 23-38. 10.1038/npp.2012.112

Muroya S., Taniguchi M., Shibata M., Oe M., Ojima K., Nakajima I. & Chikuni K. (2013) Profiling of differentially expressed microRNA and the bioinformatic target gene analyses in bovine fast- and slow-type muscles by massively parallel sequencing. Journal of Animal Science 91, 90-103. 10.2527/jas.2012-5371

Nguyen Q.H., Nguyen-Vo T.H., Le N.Q.K., Do T.T.T., Rahardja S. & Nguyen B.P. (2019) iEnhancer-ECNN: identifying enhancers and their strength using ensembles of convolutional neural networks. BMC Genomics 20, 951. 10.1186/s12864-019-6336-3

O'Brien J., Hayder H., Zayed Y. & Peng C. (2018) Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Frontiers in Endocrinology 9. 10.3389/fendo.2018.00402

Pacífico C., Ricci S., Sajovitz F., Castillo-Lopez E., Rivera-Chacon R., Petri R.M., Zebeli Q., Reisinger N. & Kreuzer-Redmer S. (2022) Bovine rumen epithelial miRNA-mRNA dynamics reveals post-transcriptional regulation of gene expression upon transition to high-grain feeding and phytogenic supplementation. Genomics 114, 110333. https://doi.org/10.1016/j.ygeno.2022.110333

Panigrahi A. & O'Malley B.W. (2021) Mechanisms of enhancer action: the known and the unknown. Genome Biology 22, 108. 10.1186/s13059-021-02322-1

Peterson S., Thompson J., Ufkin M., Sathyanarayana P., Liaw L. & Congdon C.B. (2014) Common features of microRNA target prediction tools. Frontiers in Genetics 5. 10.3389/fgene.2014.00023

Prowse-Wilkins C.P., Lopdell T.J., Xiang R., Vander Jagt C.J., Littlejohn M.D., Chamberlain A.J. & Goddard M.E. (2022) Genetic variation in histone modifications and gene expression identifies regulatory variants in the mammary gland of cattle. BMC Genomics 23, 815. 10.1186/s12864-022-09002-9

Prowse-Wilkins C.P., Wang J.H., Xiang R.D., Garner J.B., Goddard M.E. & Chamberlain A.J. (2021) Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues. Frontiers in Genetics 12. 10.3389/fgene.2021.664379

PwC (2011) The Australian Beef Industry: From family farm to international markets. PricewaterhouseCoopers.

Rada-Iglesias A., Bajpai R., Swigut T., Brugmann S.A., Flynn R.A. & Wysocka J. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279-83. 10.1038/nature09692

Ramsahoye B.H., Biniszkiewicz D., Lyko F., Clark V., Bird A.P. & Jaenisch R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proceedings of the National Academy of Sciences 97, 5237-42. doi:10.1073/pnas.97.10.5237

Razin A. & Cedar H. (1991) DNA methylation and gene expression. Microbiol Rev 55, 451-8. 10.1128/mr.55.3.451-458.1991

Reynolds L., Millaway D., Kirsch J., Infeld J. & Redmer D. (1990) Growth and in-vitro metabolism of placental tissues of cows from day 100 to day 250 of gestation. Reproduction 89, 213-22.

Rottiers V. & Näär A.M. (2012) MicroRNAs in metabolism and metabolic disorders. Nature Reviews: Molecular Cell Biology 13, 239-50. 10.1038/nrm3313

Sengar G.S., Deb R., Singh U., Junghare V., Hazra S., Raja T.V., Alex R., Kumar A., Alyethodi R.R., Kant R., Jakshara S. & Joshi C.G. (2018) Identification of differentially expressed microRNAs in Sahiwal (Bos indicus) breed of cattle during thermal stress. Cell Stress & Chaperones 23, 1019-32. 10.1007/s12192-018-0911-4

Stickland N. (1978) A quantitative study of muscle development in the bovine foetus (Bos indicus). Anatomia, histologia, embryologia 7, 193-205.

Sun G., Yan J., Noltner K., Feng J., Li H., Sarkis D.A., Sommer S.S. & Rossi J.J. (2009) SNPs in human miRNA genes affect biogenesis and function. RNA 15, 1640-51. 10.1261/rna.1560209

Sun J., Zhou Y., Cai H., Lan X., Lei C., Zhao X., Zhang C. & Chen H. (2014) Discovery of Novel and Differentially Expressed MicroRNAs between Fetal and Adult Backfat in Cattle. PLOS ONE 9, e90244. 10.1371/journal.pone.0090244

Visel A., Minovitsky S., Dubchak I. & Pennacchio L.A. (2007) VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Research 35, D88-D92.

Vrijens K., Bollati V. & Nawrot T.S. (2015) MicroRNAs as potential signatures of environmental exposure or effect: a systematic review. Environmental Health Perspectives 123, 399-411. 10.1289/ehp.1408459

Wang J., Dai X., Berry L.D., Cogan J.D., Liu Q. & Shyr Y. (2018) HACER: an atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Research 47, D106-D12. 10.1093/nar/gky864

Weedon M.N., Cebola I., Patch A.-M., Flanagan S.E., De Franco E., Caswell R., Rodríguez-Seguí S.A., Shaw-Smith C., Cho C.H.H., Allen H.L., Houghton J.A.L., Roth C.L., Chen R., Hussain K., Marsh P., Vallier L., Murray A., Ellard S., Ferrer J., Hattersley A.T. & International Pancreatic Agenesis C. (2014) Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nature Genetics 46, 61-4. 10.1038/ng.2826

Xiang R., Ghanipoor-Samami M., Johns W.H., Eindorf T., Rutley D.L., Kruk Z.A., Fitzsimmons C.J., Thomsen D.A., Roberts C.T., Burns B.M., Anderson G.I., Greenwood P.L. & Hiendleder S. (2013) Maternal and Paternal Genomes Differentially Affect Myofibre Characteristics and Muscle Weights of Bovine Fetuses at Midgestation. PLOS ONE 8, e53402. 10.1371/journal.pone.0053402

Xie W., Barr Cathy L., Kim A., Yue F., Lee Ah Y., Eubanks J., Dempster Emma L. & Ren B. (2012) Base-Resolution Analyses of Sequence and Parent-of-Origin Dependent DNA Methylation in the Mouse Genome. Cell 148, 816-31. 10.1016/j.cell.2011.12.035

Yang J., Horton J.R., Akdemir K.C., Li J., Huang Y., Kumar J., Blumenthal R.M., Zhang X. & Cheng X. (2021a) Preferential CEBP binding to T:G mismatches and increased C-to-T human somatic mutations. Nucleic Acids Research 49, 5084-94. 10.1093/nar/gkab276

Yang R., Wu F., Zhang C. & Zhang L. (2021b) iEnhancer-GAN: A Deep Learning Framework in Combination with Word Embedding and Sequence Generative Adversarial Net to Identify Enhancers and Their Strength. International Journal of Molecular Sciences 22, 3589. 10.3390/ijms22073589

Żemojtel T., Kiełbasa S.M., Arndt P.F., Behrens S., Bourque G. & Vingron M. (2011) CpG Deamination Creates Transcription Factor–Binding Sites with High Efficiency. Genome Biology and Evolution 3, 1304-11. 10.1093/gbe/evr107

Zeng W., Chen S., Cui X., Chen X., Gao Z. & Jiang R. (2021) SilencerDB: a comprehensive database of silencers. Nucleic Acids Research 49, D221-D8. 10.1093/nar/gkaa839

Zhu L., Marjani S.L. & Jiang Z. (2021) The Epigenetics of Gametes and Early Embryos and Potential Long-Range Consequences in Livestock Species—Filling in the Picture With Epigenomic Analyses. Frontiers in Genetics 12. 10.3389/fgene.2021.557934

Zhu M.J., Ford S.P., Means W.J., Hess B.W., Nathanielsz P.W. & Du M. (2006) Maternal nutrient restriction affects properties of skeletal muscle in offspring. J Physiol 575, 241-50. 10.1113/jphysiol.2006.112110

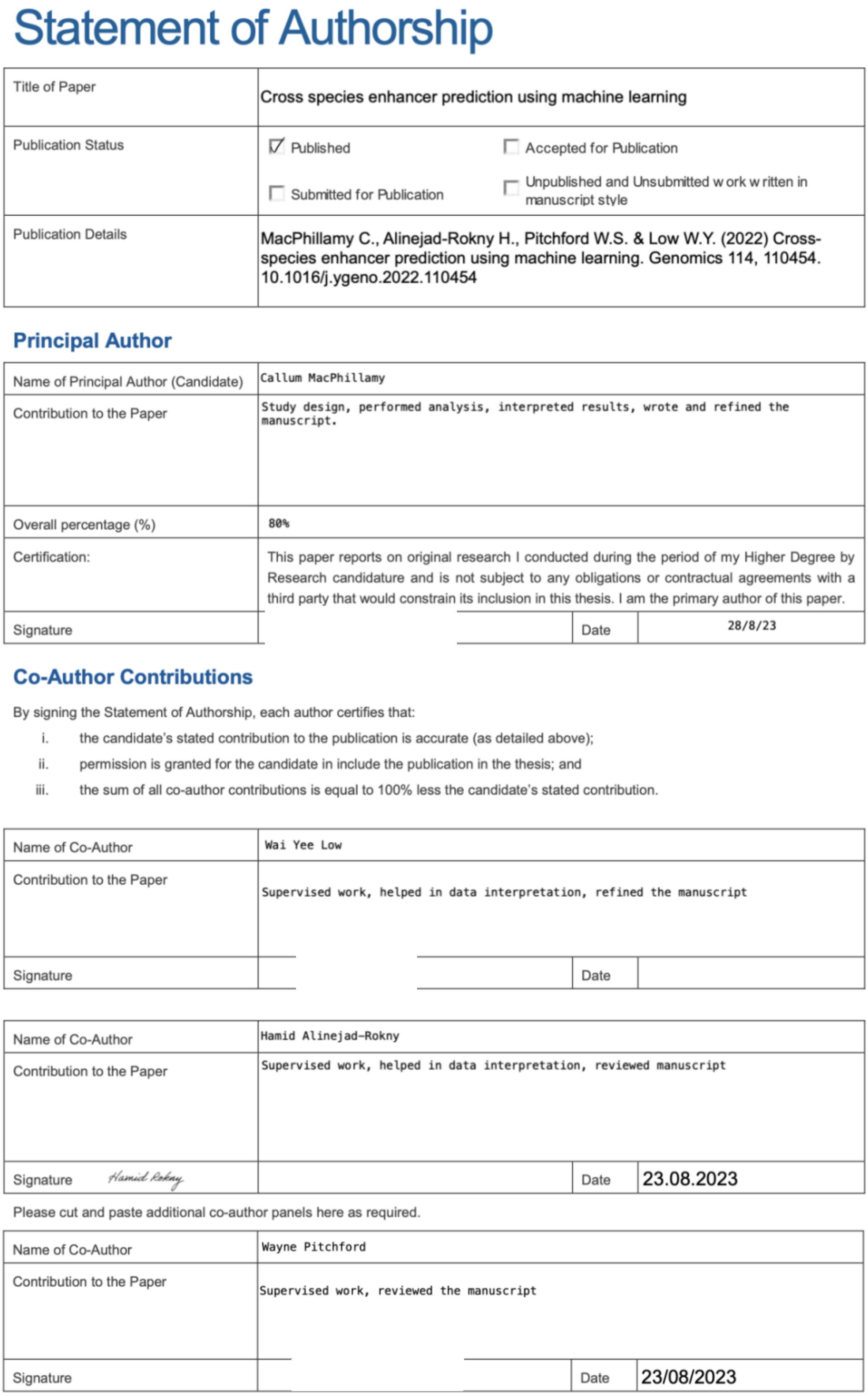
Ziller M.J., Müller F., Liao J., Zhang Y., Gu H., Bock C., Boyle P., Epstein C.B., Bernstein B.E., Lengauer T., Gnirke A. & Meissner A. (2011) Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. PLoS Genetics 7, e1002389. 10.1371/journal.pgen.1002389

Chapter 3: Cross-species prediction of enhancers using machine learning

Contextual Statement

Many machine learning tools have been developed to identify enhancers within the genome from DNA sequence alone. A subset of these have then been applied to the task of predicting enhancers across species. However, these comparisons have often focused on well-studied species like human and mouse. Few studies have investigated the performance of machine learning techniques in predicting enhancers across less well-studied species. Similarly, no previous study has evaluated different ways of representing the DNA sequence for machine-learning models in this context. Thus, this study aimed to evaluate both machine-learning models and DNA sequence representations for machine learning in the context of predicting enhancers in three relatively understudied species. This study compared nine machine learning models and four DNA representations for machine learning to determine which combination offered the best performance in predicting enhancers across species. The ultimate goal of this study was to determine which model would be best for predicting enhancers in the cattle genome. This chapter has been published in the journal *Genomics* and is available at: https://doi.org/10.1016/j.ygeno.2022.110454

Statement of authorship



Statement of Authorship

Title of Paper	Cross species enhancer prediction using machine learning
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	MacPhillamy C., Alinejad-Rokny H., Pitchford W.S. & Low W.Y. (2022) Cross-species enhancer prediction using machine learning. Genomics 114, 110454. 10.1016/j.ygeno.2022.110454

Principal Author

Name of Principal Author (Candidate)	Callum MacPhillamy		
Contribution to the Paper	Study design, performed analysis, interpreted results, wrote and refined the manuscript.		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	28/8/23

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Wai Yee Low		
Contribution to the Paper	Supervised work, helped in data interpretation, refined the manuscript		
Signature		Date	

Name of Co-Author	Hamid Alinejad-Rokny		
Contribution to the Paper	Supervised work, helped in data interpretation, reviewed manuscript		
Signature Hamid Rokny		Date	23.08.2023

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Wayne Pitchford		
Contribution to the Paper	Supervised work, reviewed the manuscript		
Signature		Date	23/08/2023

Cross-species enhancer prediction using machine learning

Authors

Callum MacPhillamy[1], Hamid Alinejad-Rokny[2,3], Wayne S Pitchford[1], Wai Yee Low[1]

[1] The Davies Livestock Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA, 5371, Australia

[2] BioMedical Machine Learning Lab, The Graduate School of Biomedical Engineering, UNSW, Sydney, NSW, 2052, Australia

[3] School of Computer Science and Engineering, University of New South Wales, Sydney, NSW, 2052, Australia

Abstract

Cis-regulatory elements (CREs) are non-coding parts of the genome that play a critical role in gene expression regulation. Enhancers, as an important example of CREs, interact with genes to influence complex traits like disease, heat tolerance and growth rate. Much of what is known about enhancers come from studies of humans and a few model organisms like mouse, with little known about other mammalian species. Previous studies have attempted to identify enhancers in less studied mammals using comparative genomics but with limited success. Recently, Machine Learning (ML) techniques have shown promising results to predict enhancer regions. Here, we investigated the ability of ML methods to identify enhancers in three non-model mammalian species (cattle, pig and dog) using human and mouse enhancer data from VISTA and publicly available ChIP-seq. We tested nine models, using four different representations of the DNA sequences in cross-species prediction using both the VISTA dataset and species-specific ChIP-seq data. We identified between 809,399 and 877,278 enhancer-like regions (ELRs) in the study species (11.6-13.7% of each genome). These predictions were close to the ~8% proportion of ELRs that covered the human genome. We propose that our ML methods have predictive ability for identifying enhancers in non-model mammalian species. We have provided a list of high confidence enhancers at https://github.com/DaviesCentreInformatics/Cross-species-enhancer-prediction and believe these enhancers will be of great use to the community.

Introduction

The genome is approximately 98% non-coding. These regions contain *cis*-regulatory elements (CREs) including enhancers, promoters, silencers, and insulators, which play crucial roles in gene expression and regulation. These effects occur through increasing the likelihood of transcription starting (enhancer), aiding the binding of transcription factors (promoter), the binding of repressors to suppress or stop gene transcription (silencer) and blocking of enhancers and acting as a barrier (insulators).

Complex interactions between enhancers and gene promoters have been shown to play a critical role in limb disorders like polydactyly (Lettice 2003; Lupiáñez *et al.* 2015), various cancers (Northcott *et al.* 2014; Weischenfeldt *et al.* 2017; Helmsauer *et al.* 2020), diabetes (Onengut-Gumuscu *et al.* 2015) and many other diseases, reviewed in 2021 (Claringbould & Zaugg 2021). There are many ways enhancers influence disease traits and at least three different scenarios can occur. A deletion in the boundary between two topologically associating domains (TADs) allows an enhancer to interact with developmental genes aberrantly, leading to polydactyly (Lupiáñez *et al.* 2015). Enhancer hijacking has been implicated in disease development (Ooi *et al.* 2020; Wang *et al.* 2021b). These enhancer hijacking events occur when a structural variation (SV) juxtaposes an enhancer next to cancer-associated genes, contributing to the abnormal expression of those genes (Wang *et al.* 2021b). Finally, single nucleotide polymorphisms (SNPs) associated with a complex disease trait like Type 1 diabetes have been found to be enriched in enhancer regions that are active in thymus, CD4+, CD8+, CD34+, B and T cells (Onengut-Gumuscu *et al.* 2015), suggesting enhancers play a crucial role in this autoimmune disease. These examples highlight some of the diverse functions enhancers have in complex traits.

Given this, it is likely that enhancers also have an essential role in complex production and fitness traits such as weight gain, milk yield, heat tolerance and disease in mammals.

Despite the importance of enhancers, relatively little is known about their identities and locations within mammalian genomes outside of human and model rodents. Mammalian livestock species like cattle and pig are important factors in both developing and developed nation economies for food security and international trade. As such, there is a need to know more about enhancers in mammalian livestock to better understand the genetic regulatory changes that are associated with productivity traits. Several studies have identified histone modifications often associated with enhancers in livestock species, such as cattle (Villar *et al.* 2015; Fang *et al.* 2019; Kern *et al.* 2021; Prowse-Wilkins *et al.* 2021) and pig (Kern *et al.* 2021; Pan *et al.* 2021; Zhao *et al.* 2021). These studies have used ChIP-seq to target H3K27ac and H3K4me1, two well-documented histone modifications associated with enhancers (Heintzman *et al.* 2007; Heintzman & Ren 2009; Creyghton *et al.* 2010; Rada-Iglesias *et al.* 2011). While these studies provide valuable insight into the possible location of enhancers in livestock, they are costly to perform and still require considerable effort and capital to determine whether those regions have actual enhancer function, i.e. proven to have a regulatory effect as assessed in transgenic mice (Visel *et al.* 2007). Identifying such enhancer regions would be immensely valuable to the livestock community as linking enhancers to genes, SNPs, SVs, or other regions of interest would provide valuable insight into how complex production and fitness traits are controlled. As no mammalian livestock species currently has a highly curated set of enhancers, there is an opportunity to leverage the high-quality data from human and mouse to develop tools for cross-species prediction. Success has been seen in adopting this approach to predict enhancers genome-wide in the human genome (Ghandi *et al.* 2014; Min *et al.* 2017; Li *et al.* 2018). Some human and mouse enhancer data, such as that available in VISTA (Visel *et al.* 2007), have been curated to the level that transgenic mice were used to evaluate the enhancer regulatory effect on gene expression. The use of human and mouse enhancers for cross-species prediction of the same feature should improve our understanding of this important regulatory element in livestock species.

There are three main ways to identify enhancers across species; comparative genomics, motif searching and examining the statistical features of the sequence. A straightforward method that uses a comparative genomics approach is to take a sequence known to be an enhancer in one genome and align it to the genome of interest (Erives & Levine 2004; Hare *et al.* 2008). While this approach would likely work for highly conserved enhancers, it is unlikely to be sensitive enough to capture those that are less conserved. Motif searching is another method that can be used when the sequence has conserved motifs (reviewed in Boeva (2016)), such as promoters that often have the highly conserved 'TATA' box. While this will likely work for identifying gene promoters, enhancers are more divergent between species (Villar *et al.* 2015; Yang *et al.* 2015), and so this method is unlikely to capture many of these regions.

Examining the statistical features of the sequence and applying machine learning (ML) techniques to the problem has shown great promise in identifying previously unknown CREs within a species, with numerous models being developed in pursuit of reliable CRE prediction (Firpi *et al.* 2010; Fletez-Brant *et al.* 2013; Ghandi *et al.* 2016; Liu *et al.* 2016; Liu *et al.* 2018; Nguyen *et al.* 2019; Cai *et al.* 2021; Inayat *et al.* 2021; Yang *et al.* 2021; Zeng *et al.* 2021). For example, Zeng *et al.* (2021) identified experimentally validated human (~8800) and mouse (~24000) silencer sequences from ~2300 published research articles. They then used these sequences with ML models to predict ~3.5 million and ~1.5 million silencers in human and mouse, respectively. These models used a combination of the sequence itself as a one-hot encoded matrix (discussed below) and the gapped and ungapped *k*-mer composition of the sequence as input (Zeng *et al.* 2021). Parallel to this, several tools have been developed to better extract meaningful features from the sequence (Chen *et al.* 2018b; Dong *et al.* 2018; Muhammod *et al.* 2019; Chen *et al.* 2020; Bonidia *et al.* 2022), which can then be used in more easily interpreted models like logistic regression to provide better insight into what sequence features are predictive of enhancer function. Deep learning (DL) is a branch of ML that uses “deep” (multiple hidden layers) neural networks to learn features associated with a label. DL models will extract their own features from the data and so can identify patterns not easily detected by a human. This automatic feature extraction is of great interest for enhancer prediction as patterns in the DNA sequence that may indicate enhancer activity are not often readily observable by humans. Recent studies have shown promising results regarding DL methods applied to enhancer prediction (Min *et al.* 2017; Nguyen *et al.* 2019; Oubounyt *et al.* 2019; Amin *et al.* 2020; Shujaat *et al.* 2020; Yang *et al.* 2021). Despite this progress, few papers have thoroughly examined the challenge of predicting enhancers across species using human and mouse data. Chen *et al.* (2018a) used a Convolutional Neural Network (CNN) and a Support Vector Machine (SVM) for cross-species enhancer prediction and reported high levels of accuracy for both methods. However, their definition of enhancer was less stringent as it included any region that had an H3K27ac ChIP-seq peak. Huh *et al.* (2018) found that sequence determinants from one species could reliably predict promoters in another species with high accuracy, with an area under the receiver operating curve (auROC) ~90%. However, model performance diminished when predicting enhancers across species (auROC ~65%). Hong *et al.* (2021) reported greater accuracy than Chen *et al.* (2018a) when predicting across species with a model based on the hierarchical attention network (HAN) (Yang *et al.* 2016) used in document classification. However, their cross-species prediction performance was still considerably lower than within-species performance.

While these studies provide good information on different approaches we may use to tackle the cross-species enhancer prediction challenge, they often focus on the computer science aspects of cross-species enhancer prediction. Less consideration appears to have been given to how the community may use the predictions and models, such as creating and sharing CRE annotations in newly assembled genomes.

There have been attempts to identify enhancers in livestock using the above-mentioned approaches with human and mouse data. For example, Wang *et al.* (2017) used a combination of cattle candidate enhancer regions identified by ChIP-seq and sequence homology of human and mouse enhancers to try and identify functional variants associated with milk production traits. However, they found that species-specific ChIP-seq data was the best option for identifying potential enhancers. This result is expected as they used sequence homology (least flexible method) as the primary search method for identifying cattle enhancers using human and mouse data.

Nguyen *et al.* (2018) developed a complex pipeline that first mapped human regulatory elements (e.g., enhancers) to the genome of a species of interest. It then used several species-specific data and gkmSVM (Ghandi *et al.* 2014; Ghandi *et al.* 2016) as part of a filtering process to identify high confidence regulatory elements in the species of interest, e.g., cattle. However, it is unclear from their results how well they were able to identify CREs in these species. Given that this pipeline also requires numerous genetic data to be available as input for the species of interest, it may prove unfeasible to use in newly assembled genomes, i.e., species that have never had their genome assembled before. The authors also acknowledge that confidence in the predictions is difficult to determine, given that H3K27ac peaks, while associated with enhancers, do not always guarantee a true enhancer. It is also worth noting that, like Wang *et al.* (2017), the identification of enhancers was reliant on sequence homology rather than learning any of the statistical features of the sequence (except for the gkmSVM step).

Motivated by these results and a desire to address the shortcomings of how the community can use these tools, we sought to evaluate a range of ML models designed to predict enhancers and validate their performance in cross-species prediction. We aimed to develop a model that had robust generalisation ability in predicting enhancers across different species. To achieve this, we trained several ML models across human, mouse and a combination of both datasets, using four different DNA representations. Finally, in order to evaluate how well the models identify potential enhancers in livestock species, we re-processed H3K27ac and H3K4me1 ChIP-seq data for cattle (Villar *et al.* 2015; Fang *et al.* 2019; Kern *et al.* 2021; Prowse-Wilkins *et al.* 2021), pig (Kern *et al.* 2021; Pan *et al.* 2021; Zhao *et al.* 2021) and dog (Villar *et al.* 2015) with the latest reference genomes ARS-UCD1.2 (GCF_002263795.1), Sscrofa11.1 (GCF_000003025.6), and Dog10K_Boxer_Tasha (GCF_000002285.5), respectively. While dog is not generally considered as a livestock species, it is a representative of another mammalian order different from those that contain human, mouse, cattle or pig. We then performed permutations to assess how well each model identified potential enhancers, or enhancer-like regions (ELRs). Once that was determined, we used the best performing model to predict ELRs genome-wide among the cattle, pig and dog genomes. We anticipate this will become a valuable resource to the research community working on mammals, especially livestock species, and expect model performance to improve as highly curated enhancers become available.

Methods

Data preparation

The sequence and genomic locations of human and mouse samples were retrieved from VISTA (Visel *et al.* 2007) (Figure 1A), which were all tissue samples. While enhancers are cell-type specific, our goal was to identify general enhancer signatures. This approach has been used in the CrepHAN model to predict enhancers across species (Hong *et al.* 2021). We first removed enhancers that were found on sex chromosomes. This gave us a total of 985 human and 653 mouse enhancer sequences (S. Table 1). The reference genomes used for human and mouse were hg19 and mm9, respectively. We then extracted 200 bp, 1 kb and 2 kb windows around the enhancer. If an enhancer region was smaller than the desired window size (*k*), we extracted a region equal to the window size centred around the midpoint of the original enhancer (Figure 1B). If the desired window size was smaller than the enhancer region (*L*), we took a sliding window equal to the window size with a step of one between the start and end of the enhancer region ($L - k + 1$), Figure 2B); this is similar to the data augmentation step used in Min *et al.* (2017). We chose 200 bp as this is a common window size used by several enhancer prediction methods (Liu *et al.* 2016; Liu *et al.* 2018; Nguyen *et al.* 2019; Cai *et al.* 2021; Inayat *et al.* 2021; Yang *et al.* 2021) that reported accuracy scores >70% when predicting enhancers within a single species. We also used 1 kb and 2 kb window sizes to assess the impact of the surrounding sequence had on model performance.

Next, we created the negative examples, non-enhancers, using a similar strategy as described in (Min *et al.* 2017; Hong *et al.* 2021). We created three bed formatted files (one for each input window size). These were a concatenation of our enhancer windows (either 200 bp, 1 kb or 2 kb), the annotation (e.g. genes, long non-coding RNAs) for the genome (hg19, mm9), the promoter regions (defined as a 2 kb region centred around the start site of gene transcription for protein-coding genes) and the blacklisted regions defined by ENCODE (Amemiya *et al.* 2019). The genomic coordinate complement of these concatenated regions was extracted using bedtools complement from BedTools (Quinlan & Hall 2010) v2.30.0 with parameter “-L" to create the non-enhancer regions. We then generated three new bed files using the non-enhancer regions as input and extracted 200 bp, 1 kb and 2 kb windows. If the non-enhancer region was smaller than our desired window size, it was ignored to ensure no overlap between positive and negative examples (Figure 1C). We then used bedtools getfasta with default parameters to extract the non-enhancer sequences. We filtered the resulting fasta records to ensure all negative examples contained no ambiguous bases and had a similar GC content distribution (mean ± std dev) as the original enhancers.

DNA representations

To give us the best chance of identifying a model that has a robust ability to generalise between species, we sought to test four different representations of DNA for machine learning. Our first representation was simply the proportional *k*-mer counts for the sequence. Here, we used $k = 1$, 2 and 3, counting all occurrences of A, C, G, T, AA, AC, …, TT and AAA, …, TTT in the sequence (Figure 1D). This allowed each sequence to be represented as a vector with a length of 84. Each value was the proportion that the given *k*-mer contributed to the sequence. E.g., a 200 bp sequence that contains 50 A's will have the value $50 \div 200 = 0.25$ for the explanatory variable A; a 200 bp sequence with 20 AAA's will have the value $20 \div 198 \approx 0.10$ for the explanatory variable AAA. We only used this representation in the logistic regression and SVM as these models can only take 1D vectors as input.

Our second method was one-hot encoding. Here, the sequence can be thought of as a rank three tensor with a shape *N*,1,4, where *N* is equal to the length of the sequence, 1 is the width of the matrix, and 4 is the number of channels which are equal to the number of nucleotides (Figure 1E).

Our third representation was an image. We used the Hilbert Curve to transform a 1D DNA sequence into a 2D image. We first created a mapping dictionary to map each *k*-mer to a value. We chose $k = 4$ as our experiments (S. table 12) and work by Yin *et al.* (2018) found this to be the best value for model performance in identifying histone modifications. As four nucleotides make up DNA and we are using $k = 4$, each 4-mer can be represented by a value between 1 and 256 ($4^4 = 256$), e.g., 'AAAA' = 1

… ‘TTTT’ = 256. We then mapped each 4-mer to the Hilbert Curve to generate an image where each point on the curve can be thought of as a pixel, and the value of the pixel is equal to the 4-mer’s mapping dictionary value, e.g., if the first 4-mer is ‘AAAA’ then ‘pixel’ (0, 0) on the curve will have the value 1 (Figure 1F). We chose the order of the Hilbert Curve such that it was the smallest possible order to contain the entire DNA sequence. For example, a sequence 200 bp long contains 200 – 4 + 1 = 197 windows of 4mers that need to be mapped to the curve. Therefore, we chose $p = 4$ as $(2^4)^2 = 256$ points on the Hilbert Curve as $197 \leq 256$. We used the formula $(2^4)^2$ to determine the smallest order (p) Hilbert Curve to contain the sequence. We used the HilbertCurve python package (v2.0.5) to generate the curves (https://github.com/galtay/hilbertcurve).

Our final representation was the word-vector; we used FastText (v0.9.2) (Bojanowski *et al.* 2017) to generate word vectors. We first split the genome into chunks at every ‘N’ and removed the ambiguous bases. We then split the genome into ten bases ‘words’ using a sliding window with a step = 1; this became our corpus (Hong *et al.* 2021). Next, we followed the word representation steps for FastText using default parameters, except for “-dim”, which we set to 30 (https://fasttext.cc/docs/en/unsupervised-tutorial.html). We chose FastText as it can generate vector representations for words not found in the original corpus. Once we had our vector representations for each ‘word’, we split the DNA sequence into non-overlapping ten base words and stacked their corresponding vector representation together (Figure 1G). As we set “-dim 30”, each word is represented by a 30 length vector, and a DNA sequence of L length can be represented as a matrix of shape $30,L\div10$, e.g., an $L = 200$bp sequence will become a 30,20 matrix.

Model Evaluation

Our primary goal was to identify the best model and sequence representation for cross-species enhancer prediction. With this in mind, we tested nine different models, specifically, logistic regression, SVM, our modified CNN based on (Min *et al.* 2017), SplitCNN (SCNN), our small implementation of the VGG model developed by Simonyan and Zisserman (2014)**,** two recurrent neural network (RNN) with an attention (Att) layer and CNN (RNN+Att+CNN) models based on work by Yu *et al.* (2020) (RNN+Att+CNN one-hot and RNN+Att+CNN word-vector) and two implementations of CrepHAN (Hong *et al.* 2021), modified for smaller input window sizes (CrepHAN one-hot, CrepHAN word-vector). The train and test datasets for logistic regression and SVM models were *k*-mer counts (Figure 1A,D). The train, validation and test datasets for CNN, SCNN, RNN+Att+CNN (one-hot) and CrepHAN (one-hot) were one-hot encoded data (Figure 1A,E). VGG was trained on DNA as an image data (Figure 1A,F), and RNN+Att+CNN (word-vector) and CrepHAN (word-vector) were trained independently on word-vector data (Figure 1A,G). We generated balanced, i.e., 1:1 positive and negative example datasets and trained and evaluated each model on the VISTA human and mouse datasets using a train-validation-test split of 70-20-10. We used four metrics to evaluate the models: Accuracy, equation 1, where TP is the true positive (an enhancer), TN is the true negative (not an enhancer), FP is false positive, non-enhancers classified as enhancers and FN is false negative, enhancers classified as not enhancers. F1-score, equation 2, where precision is equation 3 and recall is equation 4. The area under the receiver operating characteristic curve (auROC) and the area under the precision-recall curve (auPRC). We then chose the best performing model to continue through for cross-species prediction of enhancers. While we evaluated the models across three different input window sizes (S. Table 2), we chose 200 bp for the remainder of the analysis as this would give the greatest resolution for predictions among the three input sizes. Logistic regression and SVM were implemented with Scikit-Learn (Pedregosa *et al.* 2011) (v1.0). Deep learning models were implemented with Keras (v2.6) and Tensorflow (v2.6), with hyperparamter tuning being performed with KerasTuner (v1.0.4).

equation 1:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

equation 2:

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

equation 3:

$$Precision = \frac{TP}{(TP + FP)}$$

equation 4:

$$Recall = \frac{TP}{(TP + FN)}$$

Validation with ChIP-seq, permutations and final predictions

In order to assess how well the model predicts across species, we re-processed ChIP-seq data from (Villar *et al.* 2015; Fang *et al.* 2019; Kern *et al.* 2021; Pan *et al.* 2021; Prowse-Wilkins *et al.* 2021; Zhao *et al.* 2021) using the nf-core (Ewels *et al.* 2020) ChIP-seq pipeline (v1.2.2) with default parameters (https://github.com/nf-core/chipseq) and aligned all reads to the latest cattle reference (ARS-UCD1.2) (Rosen *et al.* 2020), pig reference (Sscrofa11.1) (Warr *et al.* 2020), and dog reference (Dog10K_Boxer_Tasha) (Jagannathan *et al.* 2021). The species' repeat masked regions were used in the "–blacklist" option. Effective genome size was calculated using the unique-kmers.py script from khmer (v.3.0.0a3) (Crusoe *et al.* 2015) with the length of the sequencing reads being the value for the "-k" argument. Once H3K27ac and H3K4me1 ChIP-seq mapping was completed, we performed two filtering steps to identify consensus ChIP-seq peaks. First, we determined which peaks had ≥ 50% reciprocal overlap among replicates within a single dataset for a given tissue and histone marker (Villar *et al.* 2015; Zhao *et al.* 2021). For example, if a set of H3K27ac ChIP-seq peaks reciprocally overlap by at least 50% between two or more biological replicates from a given dataset, we considered all those peaks to be true. We repeated this for all ChIP-seq datasets. Next, we concatenated all the peak files from the previous step into one file and then used bedtools merge with default parameters. We chose H3K27ac and H3K4me1 as they are both known to markers of enhancers (Creyghton *et al.* 2010; Rada-Iglesias *et al.* 2011; Zentner *et al.* 2011). We then identified candidate regions by subtracting the positions of exons, repetitive elements, and gaps from each genome with bedtools subtract. We repeated this step and included the ChIP-seq peaks for each species to generate a negative candidate region bed file for the permutation test. We took all regions that were ≥ 200 bp and extracted 200 bp windows with a step of 20 bp using a sliding window so that each 200 bp window overlapped the previous window by 180 bp. This step was similar to what was described in Figure 1B. We also filtered each 200 bp sequence at this point so that any sequences with ambiguous bases were removed. This step was repeated for the candidate regions, and the consensus ChIP peaks as these would be the negative and positive examples for the permutation tests, respectively.

Once we had our positive and negative examples, we performed permutations to determine 1) how the models compared against a naïve classifier. We defined a naïve classifier as one that randomly allocated a label to a given input, and 2) which model predicted enhancers with the most agreement to the H3K27ac and H3K4me1 ChIP-seq peaks. We took 500 random samples from the candidate regions (minus ChIP-seq peak regions) and 500 random samples from the consensus ChIP-seq peaks and obtained the prediction for this region for all models. We randomly sampled 1000 times for each model. We treated each region that overlapped a ChIP-seq peak by at least 50% as a true positive and each region that did not overlap with a ChIP-seq peak as a true negative. These regions would then form the ground truth, y. We then recorded each prediction the model made, $\hat{y}$, and computed the accuracy and F1 score. We then determined the best model based on its average accuracy and F1 score for the 1000 permutations. Once we had identified the best cross-species prediction method, we performed predictions genome-wide across all candidate regions in cattle, pig, and dog. As we trained and evaluated all models using balanced data but enhancer distribution across the genome is not balanced i.e., do not occupy 50% of the genome, we chose to filter our predictions so that only the high confidence (probability ≥ 0.99) were included in the final list of predicted enhancers.

Results

Shallow learning outperformed deep learning

While much of the recent literature has been dominated by the merits of deep learning for within species enhancer prediction, our results showed that the SVM was a reliable and performant method for cross-species enhancer prediction. Despite performing relatively poorly in within species prediction tasks (Figure 2A; Table 1; S. Table 11), we found that the SVM had a more robust performance when predicting on a new, unseen species (Figure 2B; Table 1; S. Table 11). The superior performance of the SVM in this dataset across species was surprising as the deep learning methods performed far better during within species enhancer prediction (Figure 2A; Table 1; S. Table 11).

Input sequence length influenced enhancer prediction

Next, we wanted to determine whether input size influenced the performance of deep learning models. Increasing the input size from 200 bp to 1000 bp generally resulted in better performing models except for the RNN+Att+CNN with one-hot encoded data (Figure 2A; S. Table 2). We also observed this increased performance carried over into the cross-species prediction tasks. However, these increases were far more modest than the within-species improvements. Interestingly, increasing the input window size from 200 bp to 1000 bp generally resulted in better model performance in within and cross-species prediction tasks. This trend appeared much more variable when increasing from 1000 bp to 2000 bp, with the most substantial of these fluctuations seen in the CNN model, which had an accuracy of ~97% at 1000 bp but only ~57% at 2000 bp (Figure 2A; S. Table 2). Similarly, in the cross-species performance, we observed a more variable impact on performance when increasing from 1000 bp to 2000 bp (Figure 2B; S. Table 2); however, the magnitude of these fluctuations was lesser than the within species performance. Despite the trend of model performance increasing as input window size increased, we chose the 200 bp input window size for the permutation tests and final whole-genome prediction as this would give the finest resolution for the enhancer predictions across the three input window sizes.

Training on one species was better than two

As we reviewed the literature (Chen *et al.* 2018a; Hong *et al.* 2021; Kamran *et al.* 2022) on enhancer prediction, we noticed a trend among relevant papers that they would only train on one species. We hypothesised that by using more than one species to train the models, they would learn a better function to identify enhancers than a model trained on a single species. As VISTA only has human and mouse data and given our primary goal was cross-species prediction in mammalian species, we used ChIP-seq data as the truth set for the permutations, i.e., we assumed all ChIP peaks were enhancers. We found that models trained exclusively on human data had significantly greater accuracy than those trained on human and mouse data for all target species i.e., cattle, pig, and dog (Figure 3; S. Table 3-6). This observation held for all models when comparing F1 scores, except for CrepHAN trained on one-hot encoded data in pig and dog (S. Table 3,5-6).

Despite only using one dataset for dog, which covered ~2.3% of the genome (see below), we still observed all models performed significantly better than a naïve classifier in identifying ELRs regardless of whether trained on human or human and mouse (Figure 3, S. Table 3-6).

Deep learning performed best when predicting enhancers genome-wide

It was clear that the SVM was the best model for cross-species prediction when using the highly curated dataset from VISTA (Visel *et al.* 2007). However, when performing the permutation tests, some deep learning models, specifically the CNN and SCNN models had significantly greater accuracy than the SVM (Figure 3; S. Table 3); this trend was also observed in pig and dog but did not hold for the F1 scores (S. Table 3-6). As CNN performed the best among the permutations, we included it in the final genome-wide prediction step so that we had SVM, and CNN models to make predictions on all three species.

Machine learning identified new ELRs not captured by ChIP-seq

We used consensus (see methods) H3K27ac and H3K4me1 ChIP-seq peaks as they are both markers of enhancers (Heintzman *et al.* 2007; Heintzman & Ren 2009; Creyghton *et al.* 2010; Rada-Iglesias *et al.* 2011; Kang *et al.* 2021). We identified 288,526, 155,671 and 38,578 consensus H3K27ac and H3K4me1 ChIP-peaks within the cattle, pig, and dog genomes, respectively (Figure 4A, S. Table 7). These accounted for 20.5%, 8.78% and 2.36% of the total genome lengths of our respective study species.

After filtering predicted enhancers for high confidence regions (see Methods), we identified 858,156, 841,617 and 835,403 ELRs in the cattle, pig and dog genomes using the SVM model, respectively (Figure 4A, S. Table 8). The CNN model predicted more ELRs in cattle and pig, with 877,278 and 856,105, respectively. However, in dog, we predicted only 809,399 ELRs with CNN (Figure 4A, S. Table 8). Despite identifying more ELRs than ChIP-seq, the combined length of each of the ML methods' predictions is reasonably close to the combined length of ELRs identified by ENCODE (Consortium *et al.* 2020) (S. Table 9). Both models predicted ELRs to occupy around ~11.50 – 13.71% of our study species' genomes (S. Table 8).

Looking more closely at the predicted ELRs, 364,885 SVM predictions and 394,416 CNN predictions completely overlapped with the consensus ChIP peaks in cattle (Figure 4B, S. Table 10). There were 493,271 SVM predictions and 482,862 CNN predictions not completely overlapping, i.e., were not completely inside a ChIP-seq peak region, in cattle. As the percentage of overlap required to score concordance (stringency) between ML models and ChIP peaks were decreased, we observed a general increase in the number of predicted ELRs overlapping with ChIP-seq. This was observed across all three study species. Additionally, the predictions of overlaps between the two ML models (i.e. CNN and SVM) were 76.3%, 76.8% and 73.4% concordant for cattle, pig and dog, respectively (Figure 4C). Interestingly, despite having fewer positive predictions made on dog, the CNN model had more predictions that overlap with ChIP-seq peaks than the SVM (Figure 4B, S. Table 10).

Finally, many of the ELRs predicted by both ML methods have no ChIP-seq support suggesting these may be enhancers missed by ChIP-seq; however, these would need experimental validation. The results of ELR annotations and the two best machine learning models used for their prediction are made publicly available here https://github.com/DaviesCentreInformatics/Cross-species-enhancer-prediction. We have provided a BED file of these high-confidence ELRs in the three study species presented here.

Discussion

This work set out to provide a high confidence list of enhancer-like regions (ELRs) to aid in analysing and interpreting various experiments, such as linking Genome-Wide Association Study (GWAS) SNPs to possible regulatory elements. We evaluated a variety of ML methods, from a shallow, linear model (logistic regression) to highly non-linear models like CNNs, as well as multiple ways to use DNA as inputs for ML (e.g., one-hot encoding, *k*-mer counts and word vector). To our surprise, the best performing model on the VISTA dataset when predicting across species was the SVM using the 1-, 2- and 3-mer counts of the sequence, regardless of input window size. This result was unexpected as SVMs require some prior knowledge from the user as to what features may be useful in the classification task, a term known as feature engineering. Additionally, much of the literature has focused on the merits of DL methods for DNA function prediction (Kelley *et al.* 2016; Liu *et al.* 2016; Min *et al.* 2017; Kelley *et al.* 2018; Liu *et al.* 2018; Nguyen *et al.* 2019; Kelley 2020; Hong *et al.* 2021; Inayat *et al.* 2021; Yang *et al.* 2021; Kamran *et al.* 2022). One possible reason we observed superior performance by the SVM is that there is some underlying bias within the VISTA dataset that the DL methods captured to a higher degree during their training than the SVM. DL methods rely on data that is high quality and voluminous. If biases exist within the dataset, the highly non-linear nature of many DL models will learn these biases, which in turn can negatively impact performance (Kim *et al.* 2019; Wich *et al.* 2020; Howard *et al.* 2021). There is precedent for SVMs outperforming CNNs in tasks typically believed to be ones that CNNs excel in. Wang *et al.* (2021a) investigated the performance differences between a representative CNN model and SVM for image classification. They observed superior performance from the SVM when small datasets were used and greater performance from the CNN when large datasets were used. What constitutes a small and large dataset for enhancer prediction is likely to remain debatable for some time. However, given that the SVM consistently performed well in cross-species prediction across multiple window sizes within the VISTA dataset, it is reasonable to infer that the VISTA dataset contained too few enhancer sequences to make full use of DL for cross-species prediction. This problem was only exacerbated when input window size was increased as it reduced the number of training examples. Similarly, it is possible that the DL methods were too flexible for the data available and essentially began to memorise the data (and biases) rather than learning a robust function for enhancer prediction resulting in poor cross-species accuracy. Again, this would be made worse by increasing input window size as the number of training examples consequently decreases.

One of the questions we set out to answer in this work was whether using enhancer data from multiple species could improve a models' ability to identify enhancers across species. Contrary to our expectation that training on multiple species would be more predictive, we found that training exclusively on human data resulted in more accurate models than training on a combination. To our knowledge, no other papers have looked at this aspect of enhancer prediction specifically and so comparing how our results fit with the literature is difficult. However, it may simply be a case of the mouse enhancer dataset being inferior compared to human. It is possible that sequences that are similar between human and mouse have only been identified as an enhancer in human. Therefore, the similar sequence in mouse would be labelled as "non-enhancer" in our dataset. This conflict i.e. similarity in DNA sequence as input but difference in label would then likely reduce the models' ability to learn features associated with a given label, i.e., "enhancer" or "non-enhancer", thus leading to the reduced performance compared to models trained on a single species. Alternatively, it may be that the ELRs in our study species are evolutionarily closer to human than to mouse.

Recently, Ghorbani and Zou (2019) proposed a framework for identifying the predictive value of data points to a prediction task. It would be interesting to apply this to enhancer prediction to better identify what features of the sequence are useful for prediction and what features of the sequence are confounding the model. Regardless, both human and human plus mouse trained models all perform significantly better than a naïve classifier (Figure 3, S. Table 4-6), indicating that both have merit for enhancer prediction. However, performance is generally going to be better when trained with either human or mouse data.

Given how well the SVM handled enhancer prediction across species, we expected the SVM to outperform the DL methods when testing in our study species. Interestingly, several DL models and specifically the CNN outperformed SVM when testing on our study species and using ChIP-seq as the ground truth. It may be that what appeared as overfitting or bias when testing within the VISTA enhancer dataset was the DL methods learning sequence features associated with the ChIP-seq signal. Recall that an enhancer must meet strict criteria in VISTA, one of which is strong ChIP-seq support. Given that the CNN model captured more ELRs than the SVM in cattle and pig and that it had greater ChIP-seq agreement in dog despite having fewer positive predictions, it is possible that the CNN learnt features about the sequence that are predictive of whether H3K27ac or H3K4me1 histone ChIP-seq is likely to map to that region. DL methods learning features associated with ChIP-seq signal is not unlikely, as work by Yin *et al.* (2019) developed a deep learning model that predicted histone modifications based on a combination of DNA and DNase-seq signal. This model achieved an auROC of 0.869 and 0.883 when predicting H3K4me1 and H3K27ac histone modifications, respectively, from DNA sequence alone. These results suggest that while enhancer prediction from sequence alone may still be a challenge given our limited understanding of enhancers, predicting more conserved features like CTCF boundaries (Henderson *et al.* 2019) or histone modifications (Yin *et al.* 2019) is possible.

The most recent effort to predict enhancers in cattle using human data used a HAN and word-vectors to learn the features associated with enhancers (Hong *et al.* 2021). In this work, each enhancer was represented as an 8 kb sequence centred around the midpoint of the enhancer. This model showed excellent performance (auROC = 0.960) within species (i.e., from human to human) but performance decreased considerably when predicting on cattle (auROC = 0.703) and dog (auROC = 0.695). Again, it is worth noting that the model was evaluated using cattle H3K27ac ChIP-seq data from one study (Villar *et al.* 2015), so not all peaks may be true enhancers, and not all true enhancers may have been captured by the ChIP-seq experiment. Despite the lack of validated livestock enhancers, the work by Hong *et al.* (2021) demonstrates exciting prospects for cross-species enhancer prediction as they were able to achieve predictive value using just the sequence.

In this work, we have extended the work by Hong *et al.* (2021) to include more models and ChIP-seq datasets that have only been made available recently. We observed the greatest ChIP-seq agreement with our ML models when predicting ELRs in cattle. The most likely reason for this is that we had the most diverse range of tissues, timepoints and histone markers available for cattle, unlike pig and dog that had more limited ChIP-seq datasets. This meant that the true enhancers would likely be better represented in cattle than in pig and dog, which also meant the chances of overlapping ChIP-seq positive example data with ML enhancer prediction were higher. We see the degree of ChIP overlap decrease in pig and again in dog. This decreased ChIP-seq agreement is likely a result of H3K4me1 ChIP-seq failing to pass QC, resulting in only one histone modification being available to identify candidate enhancers in pig. Additionally, the tissues and timepoints available in pig were less varied than those in cattle. Lastly, we used only one dataset (Villar *et al.* 2015) in dog, which only represented a single time point from a single tissue with a single histone modification. As a result, we would expect fewer predictions to overlap with the ChIP-seq peaks.

It is reasonable to infer that the predicted enhancers that overlap with ChIP-seq peaks are likely to have enhancer function. In contrast, the predicted ELRs with no ChIP-seq agreement are potentially novel enhancers we uncovered or false positives. We know that not all H3K27ac and H3K4me1 ChIP-seq peaks are enhancers. Interestingly, the ML predicted ELRs cover 11.59-13.71% of the genome (S. Table 8), which is similar to the ~8% covered by candidate enhancers in the human genome (S. Table 9) (Consortium *et al.* 2020). This similarity leads us to believe that most of the ML predicted ELRs represent true enhancers, although there is the possibility of false positives. Unfortunately, previous papers have not reported how their models performed when predicting enhancers genome-wide. Therefore, we cannot compare how our predicted ELRs compare with other models' predictions in terms of what proportion of the genome they predict as an enhancer. Until we can generate an ENCODE style candidate enhancer list for livestock species, we will never be sure of the actual proportion of the genome that are enhancers in our most important livestock species. We believe the list provided in this work will be the first step toward creating an ENCODE style candidate enhancers list for livestock species.

Conclusions

In this work, we sought to evaluate whether ML could be utilised to improve our understanding of livestock genomes by enabling us to identify enhancers from the sequence alone using human and mouse data. We tested a variety of DNA representations and models and found they all provide predictive value above a naïve classifier. However, the SVM was the most robust in our evaluation using the VISTA data, and CNN was the most robust in the genome-wide prediction evaluation using ChIP-seq data. We found that while the DNA sequence could be predictive of whether an enhancer was present, it is still limited in its predictive value for cross-species prediction, this is likely in part, due to the propensity for enhancers to diverge at the sequence level but not function level among species (Yang *et al.* 2015). We propose that future work be aimed at integrating other data types into these two enhancer prediction models to assess their predictive value. For example, using a combination of the DNA sequence and ATAC-seq signal to identify enhancers; similar to work by Yin *et al.* (2019) in histone modification prediction. We also recommend the continued use of ChIP-seq and other epigenomic assays such as ATAC-seq, WGBS-seq, FAIRE-seq, massively parallel reporter assay and Hi-ChIP to improve our understanding of where enhancers may be within the non-model organism. These epigenomic data will ultimately give us the best insight into where these features are. We believe this work and the candidate enhancers list provided in this study will be of great interest to the livestock community and help bring our understanding of livestock genomes closer to the level of human.

References

Amemiya H.M., Kundaje A. & Boyle A.P. (2019) The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Scientific Reports 9. 10.1038/s41598-019-45839-z

Amin R., Rahman C.R., Ahmed S., Sifat M.H.R., Liton M.N.K., Rahman M.M., Khan M.Z.H. & Shatabda S. (2020) iPromoter-BnCNN: a novel branched CNN-based predictor for identifying and classifying sigma promoters. Bioinformatics 36, 4869-75. 10.1093/bioinformatics/btaa609

Boeva V. (2016) Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. Frontiers in Genetics 7. 10.3389/fgene.2016.00024

Bojanowski P., Grave E., Joulin A. & Mikolov T. (2017) Enriching word vectors with subword information. Transactions of the association for computational linguistics 5, 135-46.

Bonidia R.P., Domingues D.S., Sanches D.S. & De Carvalho A.C.P.L.F. (2022) MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors. Briefings in Bioinformatics 23. 10.1093/bib/bbab434

Cai L.J., Ren X.B., Fu X.Z., Peng L., Gao M.Y. & Zeng X.X. (2021) iEnhancer-XG: interpretable sequence-based enhancers and their strength predictor. Bioinformatics 37, 1060-7. 10.1093/bioinformatics/btaa914

Chen L., Fish A.E. & Capra J.A. (2018a) Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. PLoS Computational Biology 14, e1006484. 10.1371/journal.pcbi.1006484

Chen Z., Zhao P., Li F., Leier A., Marquez-Lago T.T., Wang Y., Webb G.I., Smith A.I., Daly R.J., Chou K.-C. & Song J. (2018b) iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 34, 2499-502. 10.1093/bioinformatics/bty140

Chen Z., Zhao P., Li F., Marquez-Lago T.T., Leier A., Revote J., Zhu Y., Powell D.R., Akutsu T., Webb G.I., Chou K.-C., Smith A.I., Daly R.J., Li J. & Song J. (2020) iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Briefings in Bioinformatics 21, 1047-57. 10.1093/bib/bbz041

Claringbould A. & Zaugg J.B. (2021) Enhancers in disease: molecular basis and emerging treatment strategies. Trends in Molecular Medicine 27, 1060-73. 10.1016/j.molmed.2021.07.012

Consortium E.P., Moore J.E., Purcaro M.J., Pratt H.E., Epstein C.B., Shoresh N., Adrian J., Kawli T., Davis C.A., Dobin A., Kaul R., Halow J., Van Nostrand E.L., Freese P., Gorkin D.U., Shen Y., He Y., Mackiewicz M., Pauli-Behn F., Williams B.A., Mortazavi A., Keller C.A., Zhang X.O., Elhajjajy S.I., Huey J., Dickel D.E., Snetkova V., Wei X., Wang X., Rivera-Mulia J.C., Rozowsky J., Zhang J., Chhetri S.B., Zhang J., Victorsen A., White K.P., Visel A., Yeo G.W., Burge C.B., Lecuyer E., Gilbert D.M., Dekker J., Rinn J., Mendenhall E.M., Ecker J.R., Kellis M., Klein R.J., Noble W.S., Kundaje A., Guigo R., Farnham P.J., Cherry J.M., Myers R.M., Ren B., Graveley B.R., Gerstein M.B., Pennacchio L.A., Snyder M.P., Bernstein B.E., Wold B., Hardison R.C., Gingeras T.R., Stamatoyannopoulos J.A. & Weng Z. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699-710. 10.1038/s41586-020-2493-4

Creyghton M.P., Cheng A.W., Welstead G.G., Kooistra T., Carey B.W., Steine E.J., Hanna J., Lodato M.A., Frampton G.M., Sharp P.A., Boyer L.A., Young R.A. & Jaenisch R. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences 107, 21931-6. 10.1073/pnas.1016071107

Crusoe M.R., Alameldin H.F., Awad S., Boucher E., Caldwell A., Cartwright R., Charbonneau A., Constantinides B., Edvenson G., Fay S., Fenton J., Fenzl T., Fish J., Garcia-Gutierrez L., Garland P., Gluck J., González I., Guermond S., Guo J., Gupta A., Herr J.R., Howe A., Hyer A., Härpfer A., Irber L., Kidd R., Lin D., Lippi J., Mansour T., Mca'Nulty P., Mcdonald E., Mizzi J., Murray K.D., Nahum J.R., Nanlohy K., Nederbragt A.J., Ortiz-Zuazaga H., Ory J., Pell J., Pepe-Ranney C., Russ Z.N., Schwarz E., Scott C., Seaman J., Sievert S., Simpson J., Skennerton C.T., Spencer J., Srinivasan R., Standage D., Stapleton J.A., Steinman S.R., Stein J., Taylor B., Trimble W., Wiencko H.L., Wright M., Wyss B., Zhang Q., Zyme E. & Brown C.T. (2015) The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research 4, 900. 10.12688/f1000research.6924.1

Dong J., Yao Z.-J., Zhang L., Luo F., Lin Q., Lu A.-P., Chen A.F. & Cao D.-S. (2018) PyBioMed: a python library for various molecular representations of chemicals, proteins and DNAs and their interactions. Journal of Cheminformatics 10. 10.1186/s13321-018-0270-2

Erives A. & Levine M. (2004) Coordinate enhancers share common organizational features in the *Drosophila* genome. Proceedings of the National Academy of Sciences 101, 3851-6. 10.1073/pnas.0400611101

Ewels P.A., Peltzer A., Fillinger S., Patel H., Alneberg J., Wilm A., Garcia M.U., Di Tommaso P. & Nahnsen S. (2020) The nf-core framework for community-curated bioinformatics pipelines. Nature Biotechnology 38, 276-8. 10.1038/s41587-020-0439-x

Fang L., Liu S., Liu M., Kang X., Lin S., Li B., Connor E.E., Baldwin R.L., Tenesa A., Ma L., Liu G.E. & Li C.-J. (2019) Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. BMC Biology 17. 10.1186/s12915-019-0687-8

Firpi H.A., Ucar D. & Tan K. (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics 26, 1579-86. 10.1093/bioinformatics/btq248

Fletez-Brant C., Lee D., Mccallion A.S. & Beer M.A. (2013) kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic Acids Research 41, W544-W56. 10.1093/nar/gkt519

Ghandi M., Lee D., Mohammad-Noori M. & Beer M.A. (2014) Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Computational Biology 10, e1003711. 10.1371/journal.pcbi.1003711

Ghandi M., Mohammad-Noori M., Ghareghani N., Lee D., Garraway L. & Beer M.A. (2016) gkmSVM: an R package for gapped-kmer SVM. Bioinformatics 32, 2205-7. 10.1093/bioinformatics/btw203

Ghorbani A. & Zou J. (2019) Data shapley: Equitable valuation of data for machine learning. In: *International Conference on Machine Learning*, pp. 2242-51. PMLR.

Hare E.E., Peterson B.K., Iyer V.N., Meier R. & Eisen M.B. (2008) Sepsid even-skipped Enhancers Are Functionally Conserved in Drosophila Despite Lack of

Sequence Conservation. PLoS Genetics 4, e1000106. 10.1371/journal.pgen.1000106

Heintzman N.D. & Ren B. (2009) Finding distal regulatory elements in the human genome. Current Opinion in Genetics & Development 19, 541-9. 10.1016/j.gde.2009.09.006

Heintzman N.D., Stuart R.K., Hon G., Fu Y., Ching C.W., Hawkins R.D., Barrera L.O., Van Calcar S., Qu C., Ching K.A., Wang W., Weng Z., Green R.D., Crawford G.E. & Ren B. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics 39, 311-8. 10.1038/ng1966

Helmsauer K., Valieva M.E., Ali S., Chamorro González R., Schöpflin R., Röefzaad C., Bei Y., Dorado Garcia H., Rodriguez-Fos E., Puiggròs M., Kasack K., Haase K., Keskeny C., Chen C.Y., Kuschel L.P., Euskirchen P., Heinrich V., Robson M.I., Rosswog C., Toedling J., Szymansky A., Hertwig F., Fischer M., Torrents D., Eggert A., Schulte J.H., Mundlos S., Henssen A.G. & Koche R.P. (2020) Enhancer hijacking determines extrachromosomal circular MYCN amplicon architecture in neuroblastoma. Nature Communications 11. 10.1038/s41467-020-19452-y

Henderson J., Ly V., Olichwier S., Chainani P., Liu Y. & Soibam B. (2019) Accurate prediction of boundaries of high resolution topologically associated domains (TADs) in fruit flies using deep learning. Nucleic Acids Research 47, e78-e.

Hong J., Gao R. & Yang Y. (2021) CrepHAN: cross-species prediction of enhancers by using hierarchical attention networks. Bioinformatics. doi:10.1093/bioinformatics/btab349

Howard F.M., Dolezal J., Kochanny S., Schulte J., Chen H., Heij L., Huo D., Nanda R., Olopade O.I., Kather J.N., Cipriani N., Grossman R.L. & Pearson A.T. (2021) The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature Communications 12. 10.1038/s41467-021-24698-1

Huh I., Mendizabal I., Park T. & Yi S.V. (2018) Functional conservation of sequence determinants at rapidly evolving regulatory regions across mammals. PLoS Computational Biology 14, e1006451. 10.1371/journal.pcbi.1006451

Inayat N., Khan M., Iqbal N., Khan S., Raza M., Khan D.M., Khan A. & Wei D.Q. (2021) iEnhancer-DHF: Identification of Enhancers and Their Strengths Using Optimize Deep Neural Network With Multiple Features Extraction Methods. IEEE Access 9, 40783-96. 10.1109/access.2021.3062291

Jagannathan V., Hitte C., Kidd J.M., Masterson P., Murphy T.D., Emery S., Davis B., Buckley R.M., Liu Y.-H., Zhang X.-Q., Leeb T., Zhang Y.-P., Ostrander E.A. & Wang G.-D. (2021) Dog10K_Boxer_Tasha_1.0: A Long-Read Assembly of the Dog Reference Genome. Genes 12, 847. 10.3390/genes12060847

Kamran H., Tahir M., Tayara H. & Chong K.T. (2022) iEnhancer-Deep: A Computational Predictor for Enhancer Sites and Their Strength Using Deep Learning. Applied Sciences 12, 2120. 10.3390/app12042120

Kang Y., Kim Y.W., Kang J. & Kim A. (2021) Histone H3K4me1 and H3K27ac play roles in nucleosome eviction and eRNA transcription, respectively, at enhancers. The FASEB Journal 35. 10.1096/fj.202100488r

Kelley D.R. (2020) Cross-species regulatory sequence activity prediction. PLoS Computational Biology 16, e1008050. 10.1371/journal.pcbi.1008050

Kelley D.R., Reshef Y.A., Bileschi M., Belanger D., McLean C.Y. & Snoek J. (2018) Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Research 28, 739-50. 10.1101/gr.227819.117

Kelley D.R., Snoek J. & Rinn J.L. (2016) Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research 26, 990-9. 10.1101/gr.200535.115

Kern C., Wang Y., Xu X., Pan Z., Halstead M., Chanthavixay G., Saelao P., Waters S., Xiang R., Chamberlain A., Korf I., Delany M.E., Cheng H.H., Medrano J.F., Van Eenennaam A.L., Tuggle C.K., Ernst C., Flicek P., Quon G., Ross P. & Zhou H. (2021) Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. Nature Communications 12, 1821. 10.1038/s41467-021-22100-8

Kim B., Kim H., Kim K., Kim S. & Kim J. (2019) Learning not to learn: Training deep neural networks with biased data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9012-20.

Lettice L.A. (2003) A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human Molecular Genetics 12, 1725-35. 10.1093/hmg/ddg180

Li Y.F., Shi W.Q. & Wasserman W.W. (2018) Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics 19. 10.1186/s12859-018-2187-1

Liu B., Fang L.Y., Long R., Lan X. & Chou K.C. (2016) iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics 32, 362-9. 10.1093/bioinformatics/btv604

Liu B., Li K., Huang D.S. & Chou K.C. (2018) iEnhancer-EL: identifying enhancers and their strength with ensemble learning approach. Bioinformatics 34, 3835-42. 10.1093/bioinformatics/bty458

Lupiáñez D.G., Kraft K., Heinrich V., Krawitz P., Brancati F., Klopocki E., Horn D., Kayserili H., Opitz J.M. & Laxova R. (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161, 1012-25.

Min X., Zeng W., Chen S., Chen N., Chen T. & Jiang R. (2017) Predicting enhancers with deep convolutional neural networks. BMC Bioinformatics 18, 478. 10.1186/s12859-017-1878-3

Muhammod R., Ahmed S., Md Farid D., Shatabda S., Sharma A. & Dehzangi A. (2019) PyFeat: a Python-based effective feature generation tool for DNA, RNA and protein sequences. Bioinformatics 35, 3831-3. 10.1093/bioinformatics/btz165

Nguyen Q.H., Nguyen-Vo T.H., Le N.Q.K., Do T.T.T., Rahardja S. & Nguyen B.P. (2019) iEnhancer-ECNN: identifying enhancers and their strength using ensembles of convolutional neural networks. BMC Genomics 20, 951. 10.1186/s12864-019-6336-3

Nguyen Q.H., Tellam R.L., Naval-Sanchez M., Porto-Neto L.R., Barendse W., Reverter A., Hayes B., Kijas J. & Dalrymple B.P. (2018) Mammalian genomic regulatory regions predicted by utilizing human genomics, transcriptomics, and epigenetics data. Gigascience 7, 1-17. 10.1093/gigascience/gix136

Northcott P.A., Lee C., Zichner T., Stütz A.M., Erkek S., Kawauchi D., Shih D.J.H., Hovestadt V., Zapatka M., Sturm D., Jones D.T.W., Kool M., Remke M., Cavalli F.M.G., Zuyderduyn S., Bader G.D., Vandenberg S., Esparza L.A., Ryzhova M., Wang W., Wittmann A., Stark S., Sieber L., Seker-Cin H., Linke

L., Kratochwil F., Jäger N., Buchhalter I., Imbusch C.D., Zipprich G., Raeder B., Schmidt S., Diessl N., Wolf S., Wiemann S., Brors B., Lawerenz C., Eils J., Warnatz H.-J., Risch T., Yaspo M.-L., Weber U.D., Bartholomae C.C., Von Kalle C., Turányi E., Hauser P., Sanden E., Darabi A., Siesjö P., Sterba J., Zitterbart K., Sumerauer D., Van Sluis P., Versteeg R., Volckmann R., Koster J., Schuhmann M.U., Ebinger M., Grimes H.L., Robinson G.W., Gajjar A., Mynarek M., Von Hoff K., Rutkowski S., Pietsch T., Scheurlen W., Felsberg J., Reifenberger G., Kulozik A.E., Von Deimling A., Witt O., Eils R., Gilbertson R.J., Korshunov A., Taylor M.D., Lichter P., Korbel J.O., Wechsler-Reya R.J. & Pfister S.M. (2014) Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. Nature 511, 428-34. 10.1038/nature13379

Onengut-Gumuscu S., Chen W.-M., Burren O., Cooper N.J., Quinlan A.R., Mychaleckyj J.C., Farber E., Bonnie J.K., Szpak M., Schofield E., Achuthan P., Guo H., Fortune M.D., Stevens H., Walker N.M., Ward L.D., Kundaje A., Kellis M., Daly M.J., Barrett J.C., Cooper J.D., Deloukas P., Todd J.A., Wallace C., Concannon P. & Rich S.S. (2015) Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. Nature Genetics 47, 381-6. 10.1038/ng.3245

Ooi W.F., Nargund A.M., Lim K.J., Zhang S., Xing M., Mandoli A., Lim J.Q., Ho S.W.T., Guo Y., Yao X., Lin S.J., Nandi T., Xu C., Ong X., Lee M., Tan A.L.-K., Lam Y.N., Teo J.X., Kaneda A., White K.P., Lim W.K., Rozen S.G., Teh B.T., Li S., Skanderup A.J. & Tan P. (2020) Integrated paired-end enhancer profiling and whole-genome sequencing reveals recurrent <em>CCNE1</em> and <em>IGF2</em> enhancer hijacking in primary gastric adenocarcinoma. Gut 69, 1039-52. 10.1136/gutjnl-2018-317612

Oubounyt M., Louadi Z., Tayara H. & Chong K.T. (2019) DeePromoter: Robust Promoter Predictor Using Deep Learning. Frontiers in Genetics 10, 286. 10.3389/fgene.2019.00286

Pan Z., Yao Y., Yin H., Cai Z., Wang Y., Bai L., Kern C., Halstead M., Chanthavixay G., Trakooljul N., Wimmers K., Sahana G., Su G., Lund M.S., Fredholm M., Karlskov-Mortensen P., Ernst C.W., Ross P., Tuggle C.K., Fang L. & Zhou H. (2021) Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. Nature Communications 12. 10.1038/s41467-021-26153-7

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R. & Dubourg V. (2011) Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12, 2825-30.

Prowse-Wilkins C.P., Wang J.H., Xiang R.D., Garner J.B., Goddard M.E. & Chamberlain A.J. (2021) Putative Causal Variants Are Enriched in Annotated Functional Regions From Six Bovine Tissues. Frontiers in Genetics 12. 10.3389/fgene.2021.664379

Quinlan A.R. & Hall I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-2. 10.1093/bioinformatics/btq033

Rada-Iglesias A., Bajpai R., Swigut T., Brugmann S.A., Flynn R.A. & Wysocka J. (2011) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470, 279-83. 10.1038/nature09692

Rosen B.D., Bickhart D.M., Schnabel R.D., Koren S., Elsik C.G., Tseng E., Rowan T.N., Low W.Y., Zimin A., Couldrey C., Hall R., Li W., Rhie A., Ghurye J., McKay S.D., Thibaud-Nissen F., Hoffman J., Murdoch B.M., Snelling W.M., McDaneld T.G., Hammond J.A., Schwartz J.C., Nandolo W., Hagen D.E., Dreischer C., Schultheiss S.J., Schroeder S.G., Phillippy A.M., Cole J.B., Van Tassell C.P., Liu G., Smith T.P.L. & Medrano J.F. (2020) De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9, giaa021-giaa. 10.1093/gigascience/giaa021

Shujaat M., Wahab A., Tayara H. & Chong K.T. (2020) pcPromoter-CNN: A CNN-Based Prediction and Classification of Promoters. Genes (Basel) 11. 10.3390/genes11121529

Simonyan K. & Zisserman A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Villar D., Berthelot C., Aldridge S., Rayner T.F., Lukk M., Pignatelli M., Park T.J., Deaville R., Erichsen J.T., Jasinska A.J., Turner J.M., Bertelsen M.F., Murchison E.P., Flicek P. & Odom D.T. (2015) Enhancer evolution across 20 mammalian species. Cell 160, 554-66. 10.1016/j.cell.2015.01.006

Visel A., Minovitsky S., Dubchak I. & Pennacchio L.A. (2007) VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Research 35, D88-D92.

Wang M., Hancock T.P., MacLeod I.M., Pryce J.E., Cocks B.G. & Hayes B.J. (2017) Putative enhancer sites in the bovine genome are enriched with variants affecting complex traits. Genetics Selection Evolution 49, 56. 10.1186/s12711-017-0331-4

Wang P., Fan E. & Wang P. (2021a) Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognition Letters 141, 61-7.

Wang X., Xu J., Zhang B., Hou Y., Song F., Lyu H. & Yue F. (2021b) Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. Nature Methods 18, 661-8. 10.1038/s41592-021-01164-w

Warr A., Affara N., Aken B., Beiki H., Bickhart D.M., Billis K., Chow W., Eory L., Finlayson H.A., Flicek P., Giron C.G., Griffin D.K., Hall R., Hannum G., Hourlier T., Howe K., Hume D.A., Izuogu O., Kim K., Koren S., Liu H., Manchanda N., Martin F.J., Nonneman D.J., O'Connor R.E., Phillippy A.M., Rohrer G.A., Rosen B.D., Rund L.A., Sargent C.A., Schook L.B., Schroeder S.G., Schwartz A.S., Skinner B.M., Talbot R., Tseng E., Tuggle C.K., Watson M., Smith T.P.L. & Archibald A.L. (2020) An improved pig reference genome sequence to enable pig genetics and genomics research. Gigascience 9, 1-14. 10.1093/gigascience/giaa051

Weischenfeldt J., Dubash T., Drainas A.P., Mardin B.R., Chen Y., Stütz A.M., Waszak S.M., Bosco G., Halvorsen A.R., Raeder B., Efthymiopoulos T., Erkek S., Siegl C., Brenner H., Brustugun O.T., Dieter S.M., Northcott P.A., Petersen I., Pfister S.M., Schneider M., Solberg S.K., Thunissen E., Weichert W., Zichner T., Thomas R., Peifer M., Helland A., Ball C.R., Jechlinger M., Sotillo R., Glimm H. & Korbel J.O. (2017) Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nature Genetics 49, 65-74. 10.1038/ng.3722

Wich M., Bauer J. & Groh G. (2020) Impact of politically biased data on hate speech classification. In: *Proceedings of the fourth workshop on online abuse and harms*, pp. 54-64.

Yang R., Wu F., Zhang C. & Zhang L. (2021) iEnhancer-GAN: A Deep Learning Framework in Combination with Word Embedding and Sequence Generative

Adversarial Net to Identify Enhancers and Their Strength. International Journal of Molecular Sciences 22, 3589. 10.3390/ijms22073589

Yang S., Oksenberg N., Takayama S., Heo S.-J., Poliakov A., Ahituv N., Dubchak I. & Boffelli D. (2015) Functionally conserved enhancers with divergent sequences in distant vertebrates. BMC Genomics 16. 10.1186/s12864-015-2070-7

Yang Z., Yang D., Dyer C., He X., Smola A. & Hovy E. (2016) Hierarchical attention networks for document classification. In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 1480-9.

Yin B., Balvert M., Zambrano D., Schönhuth A. & Bohte S. (2018) An image representation based convolutional network for DNA classification. arXiv preprint arXiv:1806.04931.

Yin Q., Wu M., Liu Q., Lv H. & Jiang R. (2019) DeepHistone: a deep learning approach to predicting histone modifications. BMC Genomics 20, 193. 10.1186/s12864-019-5489-4

Yu S., Liu D., Zhu W., Zhang Y. & Zhao S. (2020) Attention-based LSTM, GRU and CNN for short text classification. Journal of Intelligent & Fuzzy Systems 39, 333-40. 10.3233/JIFS-191171

Zeng W., Chen S., Cui X., Chen X., Gao Z. & Jiang R. (2021) SilencerDB: a comprehensive database of silencers. Nucleic Acids Research 49, D221-D8. 10.1093/nar/gkaa839

Zentner G.E., Tesar P.J. & Scacheri P.C. (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Research 21, 1273-83. 10.1101/gr.122382.111

Zhao Y., Hou Y., Xu Y., Luan Y., Zhou H., Qi X., Hu M., Wang D., Wang Z., Fu Y., Li J., Zhang S., Chen J., Han J., Li X. & Zhao S. (2021) A compendium and comparative epigenomics analysis of cis-regulatory elements in the pig genome. Nature Communications 12. 10.1038/s41467-021-22448-x

Tables

Table 1. Comparison of within and between species performance of nine ML models using 200 bp sequence input.

	Human to Human				Human to Mouse				Mouse to Mouse				Mouse to Human			
	Acc [1]	F1 [2]	ROC [3]	PR [4]	Acc [1]	F1 [2]	ROC [3]	PR [4]	Acc [1]	F1 [2]	ROC [3]	PR [4]	Acc [1] (%)	F1 [2]	ROC [3]	PR [4]
CNN	0.954	0.95	0.95	0.93	0.641	0.58	0.64	0.60	0.936	0.93	0.93	0.97	0.572	0.62	0.57	0.54
Cr-HAN	0.877	0.88	0.88	0.83	0.642	0.62	0.64	0.59	0.859	0.86	0.86	0.80	0.572	0.64	0.57	0.54
Cr-HAN*	0.754	0.75	0.75	0.69	0.623	0.64	0.62	0.58	0.747	0.74	0.75	0.69	0.595	0.66	0.59	0.55
RNN	0.982	0.98	0.98	0.97	0.640	0.56	0.64	0.60	0.992	0.99	0.99	0.99	0.538	0.59	0.54	0.52
RNN*	0.804	0.80	0.80	0.75	0.638	0.63	0.64	0.59	0.817	0.81	0.82	0.76	0.593	0.65	0.59	0.55
SCNN	0.976	0.98	0.98	0.96	0.646	0.57	0.65	0.61	0.980	0.98	0.98	0.97	0.580	0.59	0.58	0.55
VGG	0.730	0.70	0.73	0.68	0.630	0.61	0.63	0.58	0.735	0.64	0.74	0.70	0.571	0.63	0.57	0.54
SVM	0.684	0.67	0.68	0.63	0.634	0.61	0.63	0.59	0.648	0.63	0.65	0.60	0.669	0.70	0.67	0.61
LR	0.767	0.73	0.77	0.72	0.650	0.63	0.65	0.60	0.770	0.72	0.77	0.74	0.615	0.66	0.62	0.57

The four columns on the first row identify that training and testing split that that group of metrics represents for each of the nine models. ‘Human to Human’ denotes the metrics for each of the models that were trained and tested on human data. ‘Human to Mouse’ denotes the metrics for each model trained on human and tested on mouse. The metrics presented here are for models trained on 200 bp input sequences.

* denotes an identical model trained on word-vector data instead of one-hot encoded.

[1] represents the accuracy of the model as described in Methods

[2] represents the F1 of the model, as described in Methods

[3] represents the area under the Receiver Operating Characteristic curve.

[4] represents the area under the Precision-Recall curve.

Figures

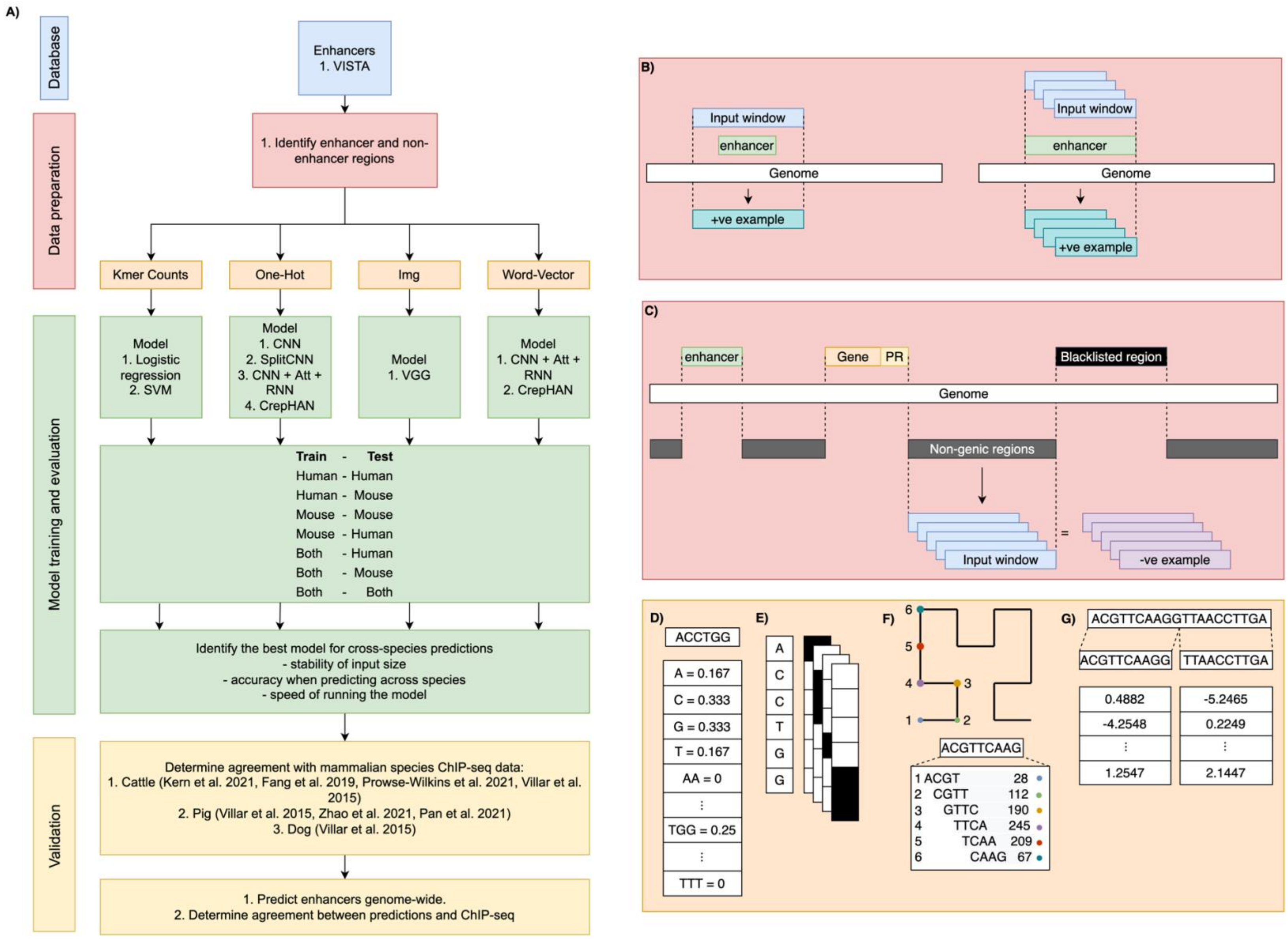

A)
Database
Data preparation
Model training and evaluation
Validation
Enhancers
1. VISTA
1. Identify enhancer and non-enhancer regions
Kmer Counts
One-Hot
Img
Word-Vector
Model
1. Logistic regression
2. SVM
Model
1. CNN
2. SplitCNN
3. CNN + Att + RNN
4. CrepHAN
Model
1. VGG
Model
1. CNN + Att + RNN
2. CrepHAN
Train - Test
Human - Human
Human - Mouse
Mouse - Mouse
Mouse - Human
Both - Human
Both - Mouse
Both - Both
Identify the best model for cross-species predictions
- stability of input size
- accuracy when predicting across species
- speed of running the model
Determine agreement with mammalian species ChIP-seq data:
1. Cattle (Kern et al. 2021, Fang et al. 2019, Prowse-Wilkins et al. 2021, Villar et al. 2015)
2. Pig (Villar et al. 2015, Zhao et al. 2021, Pan et al. 2021)
3. Dog (Villar et al. 2015)
1. Predict enhancers genome-wide.
2. Determine agreement between predictions and ChIP-seq
B)
Input window
enhancer
Genome
+ve example
Input window
enhancer
Genome
+ve example
C)
enhancer
Gene
PR
Blacklisted region
Genome
Non-genic regions
Input window
=
-ve example
D)
ACCTGG
A = 0.167
C = 0.333
G = 0.333
T = 0.167
AA = 0
⋮
TGG = 0.25
⋮
TTT = 0
E)
A
C
C
T
G
G
F)
6
5
4
3
1
2
ACGTTCAAG
1 ACGT 28
2 CGTT 112
3 GTTC 190
4 TTCA 245
5 TCAA 209
6 CAAG 67
G)
ACGTTCAAGGTTAACCTTGA
ACGTTCAAGG
TTAACCTTGA
0.4882
-4.2548
⋮
1.2547
-5.2465
0.2249
⋮
2.1447

Figure 1. Overview of experimental methodology, models tested, and data inputs used. (**A**) Flowchart of the overall experimental design. We took all autosomal enhancer sequences for human and mouse from the VISTA enhancer database and used these to construct our training, validation, and testing datasets. We evaluated nine models with various training and testing splits, e.g., train on human, test on human and train on human, test on mouse. We then identified the best model based on the cross-species performance using the VISTA dataset. We tested all models again using species-specific ChIP-seq as the positive set and regions that did not have any ChIP-seq map to them as negative. We selected the best model based on the accuracy on the study species and used the best model from each evaluation step to predict enhancers genome-wide in each of our study species. (**B**) We generated positive training examples from the enhancers by taking a sliding window within the enhancer range if the enhancer region was larger than our window size and centred around the midpoint of the enhancer if our window size was larger. (**C**) We generated negative training examples by subtracting the regions generated in **B**, the genome annotation, inferred promoter regions and blacklisted regions from the genome. We then took a sliding window from all regions ≥ to the length of our window. (**D**) graphical representation of *k*-mer count vector that represents a DNA sequence. (**E**) a simple example of a one-hot encoded matrix used to represent the DNA sequence. (**F**) graphical representation of how the Hilbert Curve maps 4mers to points. (**G**) A simple example of how we represented the sequence using word-vectors. Each 10mer was represented by a word-vector; we then took non-overlapping

windows from the sequence and stacked the corresponding word vectors together.

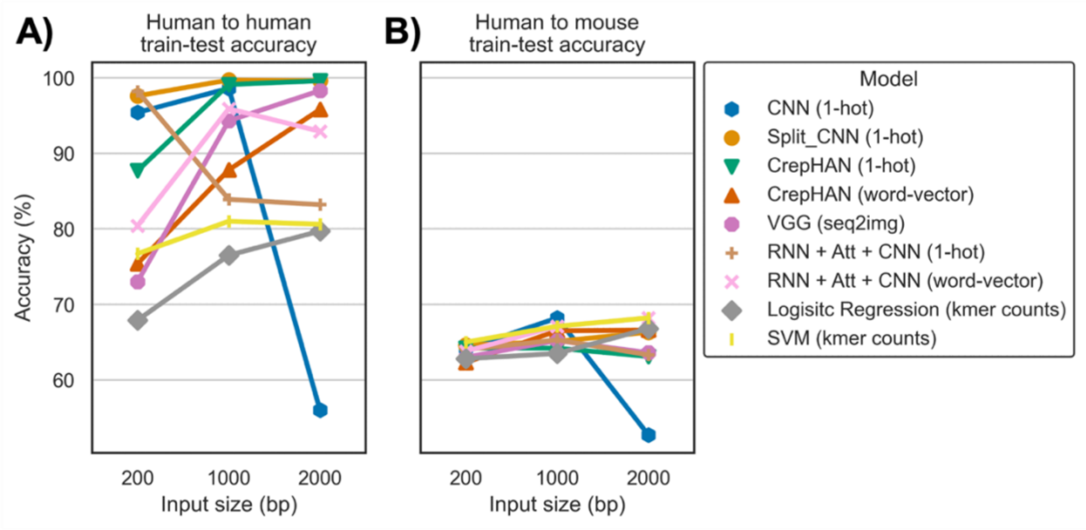


Figure 2. Comparison of model performance across each three input window sizes and prediction scenarios. The left panel shows the human to human (train-test) accuracy, and the right panel shows the human to mouse (train-test) accuracy for each of the nine models. The y axis shows model accuracy on the test data, while the x-axis shows the model's input size. Each model is colour-coded with blue representing the CNN; orange, SCNN; green, CrepHAN (one-hot); red, CrepHAN (word-vector); purple, VGG; brown, RNN+Att+CNN (one-hot); pink, RNN+Att+CNN (word-vector); grey, logistic regression and gold, SVM.

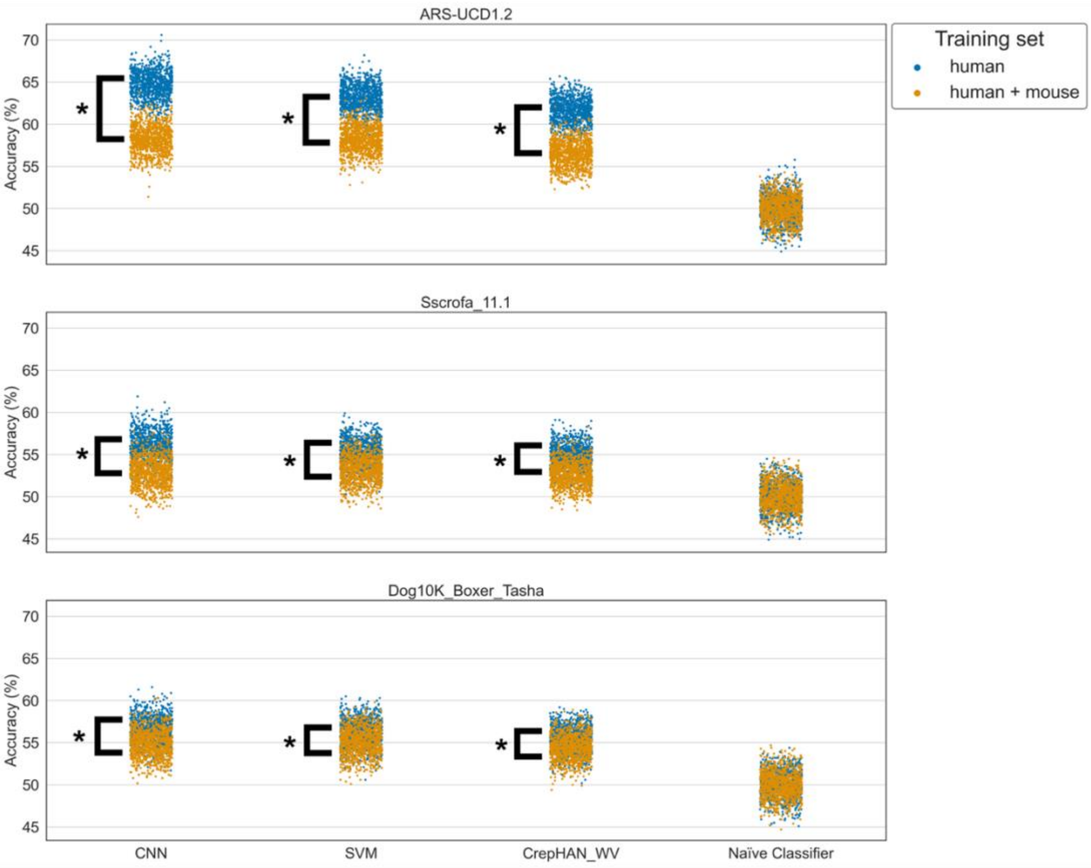


Figure 3. Comparison of model performance predicting enhancers-like regions within ARS-UCD1.2, Sscrofa_11.1 and Dog10K_Boxer_Tasha genomes. The top panel shows the accuracy of CNN, SVM, CrepHAN trained on word-vectors (CrepHAN_WV), and a naïve classifier, i.e., has not learnt any features. Blue highlights models trained exclusively on human VISTA enhancer data, and orange denotes models trained on both human and mouse VISTA enhancer data. Each dot represents the accuracy of enhancer prediction from a randomly chosen set of sequences. There are 1000 dots to represent using human sequence as the training input, and another 1000 dots to represent using human plus mouse sequence as the training input. Results are presented for each species in different panels * denotes a p-value <0.0001 (Mann-Whitney U test).

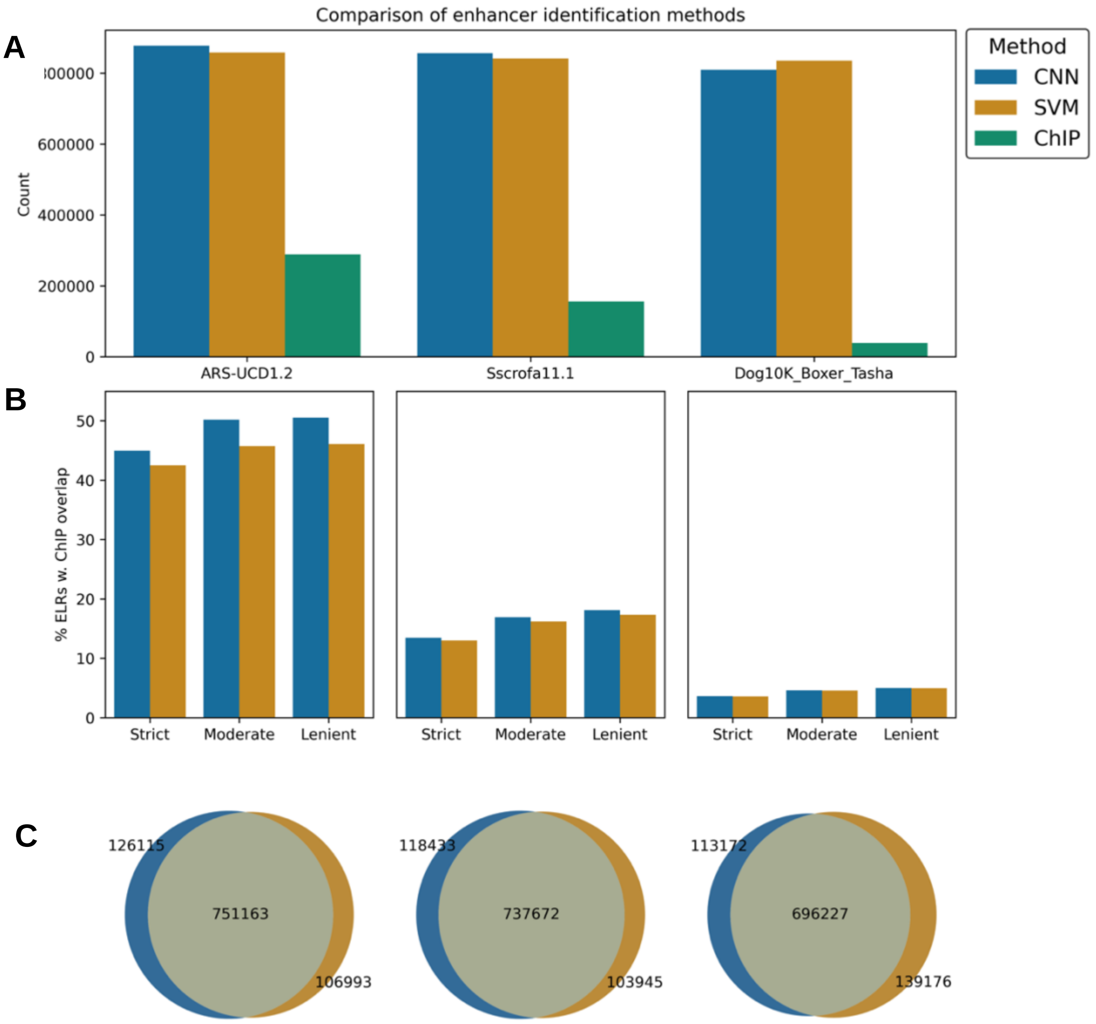


Figure 4. Comparison of enhancer-like regions (ELRs) predicted by three different methods and degree of agreement between predicted ELRs and ChIP-seq. (**A**) shows a comparison between the number of ELRs predicted by CNN (blue), SVM (orange) and ChIP (green). These colours are used throughout the figure. (**B**) Each panel represents one of our study species. The Y-axis represents the percentage of the predicted ELRs that overlap with a ChIP peak region. The X-axis represents the three categories we used to determine a positive hit, strict, moderate and lenient. A positive hit in the 'strict' category had to have its complete length covered by a ChIP-seq peak, 'moderate' needed 50% of its length covered by a peak and 'lenient' needed

25% of its length covered. (**C**) Venn diagrams showing the intersection between ELRs predicted by CNN and ELRS predicted by SVM for each of our study species.

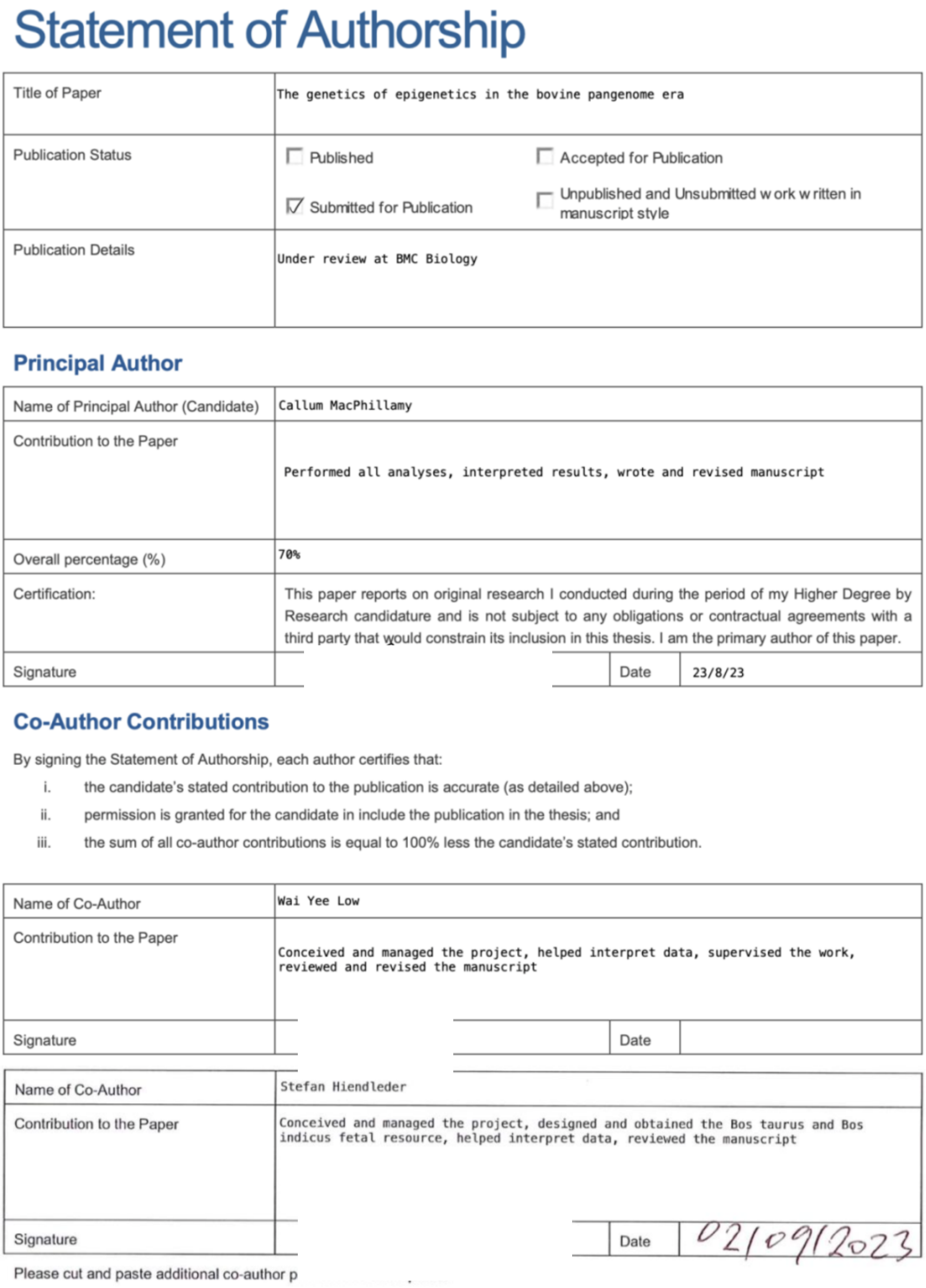
Supplementary materials for Chapter 3 can be in Appendix I.

Chapter 4: The genetics of epigenetics in the bovine pangenome era

Contextual Statement

The preceding chapter evaluated the cross-species enhancer prediction accuracy of several machine-learning models and DNA representations. The best model identified was then used to predict enhancers in the cattle genome. Chapter 4 builds on Chapter 3 by investigating an epigenetic modification that influences gene expression, DNA methylation. No studies have investigated the DNA methylation between Brahman and Angus with the ability to disentangle breed and parent-of-origin effects. Furthermore, with the creation of bovine pangenomes, no previous studies have investigated the impact of reference genome choice in the analysis of DNA methylation data. To that end, this study investigated the breed and parent-of-origin effects of DNA methylation and what role the choice of reference genome plays in the analysis. Standard tools for DNA methylation analysis and conventional statistical methods were used to compare DNA methylation between breeds at genomic regions like enhancers (Chapter 3) and promoters and the impact of reference genome choice on the analysis.

Statement of authorship



Statement of Authorship

Title of Paper	The genetics of epigenetics in the bovine pangenome era	
Publication Status	□ Published ☑ Submitted for Publication	□ Accepted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Under review at BMC Biology	

Principal Author

Name of Principal Author (Candidate)	Callum MacPhillamy		
Contribution to the Paper	Performed all analyses, interpreted results, wrote and revised manuscript		
Overall percentage (%)	70%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	23/8/23

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

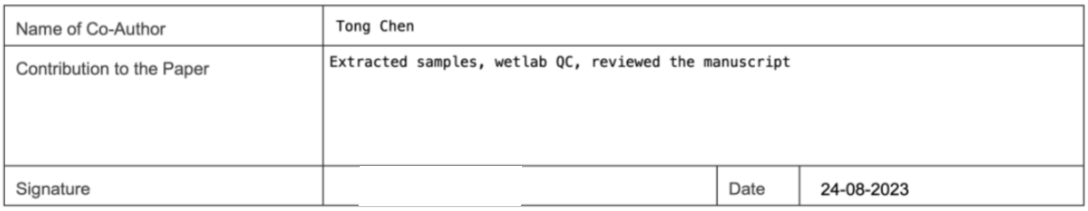
ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

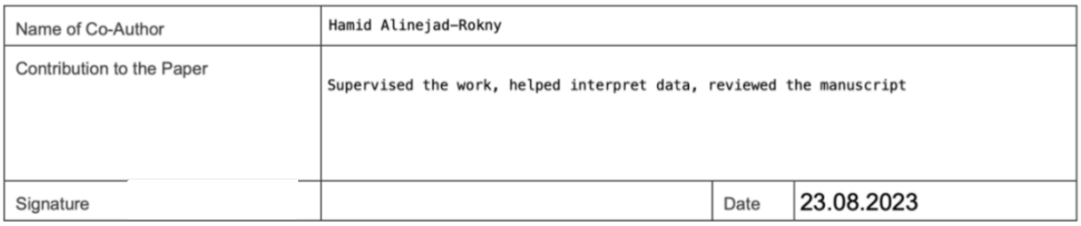
Name of Co-Author	Wai Yee Low		
Contribution to the Paper	Conceived and managed the project, helped interpret data, supervised the work, reviewed and revised the manuscript		
Signature		Date	

Name of Co-Author	Stefan Hiendleder		
Contribution to the Paper	Conceived and managed the project, designed and obtained the Bos taurus and Bos indicus fetal resource, helped interpret data, reviewed the manuscript		
Signature		Date	02/09/2023

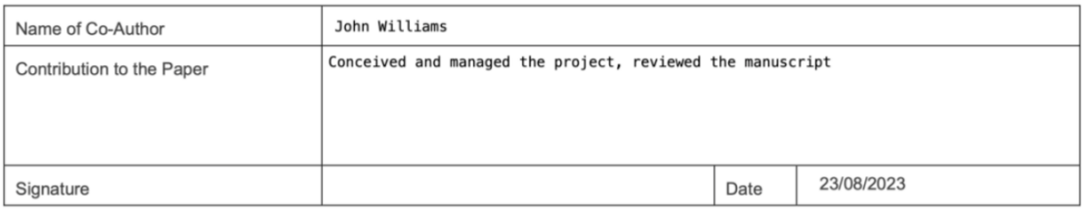
Please cut and paste additional co-author p



Name of Co-Author	Tong Chen		
Contribution to the Paper	Extracted samples, wetlab QC, reviewed the manuscript		
Signature		Date	24-08-2023



Name of Co-Author	Hamid Alinejad-Rokny		
Contribution to the Paper	Supervised the work, helped interpret data, reviewed the manuscript		
Signature		Date	23.08.2023



Name of Co-Author	John Williams		
Contribution to the Paper	Conceived and managed the project, reviewed the manuscript		
Signature		Date	23/08/2023

The genetics of epigenetics in the bovine pangenome era

Authors

Callum MacPhillamy[1], Tong Chen[1], Stefan Hiendleder[1,2], John L. Williams[1,3], Hamid Alinejad-Rokny[4], Wai Yee Low[1]

[1]The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA 5371, Australia

[2]Robinson Research Institute, The University of Adelaide, North Adelaide, SA 5006, Australia

[3]Department of Animal Science, Food and Nutrition, Università Cattolica del Sacro Cuore, 29122 Piacenza, Italy

[4]BioMedical Machine Learning Lab, The Graduate School of Biomedical Engineering, UNSW, Sydney, NSW 2052, Australia

Abstract

Most DNA methylation studies have used a single reference genome with little attention paid to the bias introduced due to the reference chosen. Genetic variation, including single nucleotide polymorphism (SNPs) and structural variants (SVs), can lead to differences in methylation sites (CpGs) between individuals of the same species. We analysed whole genome bisulfite sequencing (WGBS) data from the fetal liver of Angus (*Bos taurus taurus*), Brahman (*Bos taurus indicus*) and reciprocally crossed samples. Using reference genomes for each breed from the Bovine Pangenome Consortium, we investigated the influence of reference genome choice on the breed- and parent-of-origin effects in methylome analyses. Our findings revealed that about 75% of CpG sites were shared between Angus and Brahman, ~5% were breed-specific, and ~20% were unresolved. We demonstrated up to 2% quantification bias in global methylation when an incorrect reference genome was used. Furthermore, we found that SNPs and SVs were 8-fold (p-value < 5×10^{-324}) and 1.13-fold (p-value < 5×10^{-324}) higher in CpGs, respectively, compared to the rest of the genome. We found only 0.8% of differentially methylated regions (DMRs) overlapped with differentially expressed genes (DEGs) and suggest that DMRs may be impacting enhancers that target these DEGs. DMRs overlapped with imprinted genes, of which one, *Dgat1*, which is important for fat metabolism and weight gain, was found in the breed-specific and sire-of-origin comparisons. This work demonstrates the need to consider reference genome effects to explore genetic and epigenetic differences accurately and identify DMRs involved in controlling certain genes.

Introduction

DNA methylation is a key epigenetic modification that plays a vital role in regulating gene expression, repression of transposable elements, and parental chromosome specific regulation through genomic imprinting and X-chromosome inactivation (Li & Zhang 2014; Jansz 2019). In mammals, DNA methylation primarily occurs at C-phosphate-G dinucleotides (CpGs) (Ramsahoye *et al.* 2000; Ziller *et al.* 2011)**.** DNA methylation influences gene expression either by recruiting proteins involved in gene repression or by blocking transcription factor binding sites (TFBSs) within promoter regions (reviewed by Moore *et al.* 2013). Hypomethylation of a promoter has been associated with the increased expression of the corresponding gene (Kass *et al.* 1997). However, recent work has shown that promoter hypermethylation can also lead to gene expression (Smith *et al.* 2020). The relationship between DNA methylation and gene expression is complicated by the role of enhancer methylation in regulating gene expression (Spainhour *et al.* 2019; Cho *et al.* 2022). In the presence of high DNA methylation, enhancers have been observed to be associated with high levels of the histone modification H3K27ac (Charlet *et al.* 2016), which is often associated with active gene transcription (Creyghton *et al.* 2010; Wang *et al.* 2017; Kang *et al.* 2021; Zhu *et al.* 2022).

Most DNA methylation studies have used a single reference genome with little or no knowledge of the impact of reference genome choice on the interpretation of methylome differences. The choice of reference genome has been shown to have an impact on DNA methylation analyses, with up to a nine per cent bias reported when the incorrect reference is used (Wulfridge *et al.* 2019). Using a single reference genome has been shown to bias read mapping in favour of reads with high similarity to the reference (Degner *et al.* 2009; Brandt *et al.* 2015; Salavati *et al.* 2019; Groza *et al.* 2020; Chen *et al.* 2021). This bias occurs because reads containing non-reference alleles or regions that are divergent from the reference either align poorly, align to the wrong genomic region, or fail to align. This reference bias has been shown to affect analyses of cattle breeds (Crysnanto & Pausch 2020; Lloret-Villas *et al.* 2021), humans (Degner *et al.* 2009; Günther & Nettelblad 2019), and sheep (Salavati *et al.* 2019).

The majority of mammalian methylation occurs in the CpG context. Consequently, a single nucleotide polymorphism (SNP) can remove a methylation site, thus introducing a reference bias if the individuals being studied do not possess the same SNPs as the individual used to generate the reference. In addition to SNPs, structural variations (SVs) among individuals may remove or introduce CpG sites. The disparity between CpG sites can confound analyses by identifying a methylated CpG in one individual when another individual has no CpG at that position. As a result, SNPs and SVs can both introduce bias, individuals with different SNPs or SVs relative to the reference will have a poorer alignment accuracy than those without. As a result, entire genomic regions may lose sufficient coverage for analysis, despite likely having important biological function. Moreover, if individuals have insertion SVs that carry CpG sites, reads that originate from the SV can only be mapped if the complete pangenome for the population is available, or at the very least, that SV is present within the reference genome. We consider SNPs and SVs that alter CpGs as genetic changes with potential effects on epigenetic regulation. We use the term ‘genetics of epigenetics’ to describe this phenomenon.

As more genomes for a given species become available, the research community is gradually shifting toward using pangenomes to account for genetic variation within a population more accurately. A pangenome is a collection of the genomes of multiple individuals, representing all genetic variation within that population and is thus a more accurate way to represent genetic diversity than a single reference genome (Wang *et al.* 2022). Current pangenome projects include human (Wang *et al.* 2022; Liao *et al.* 2023), cattle (Smith *et al.* 2023), and maize (Woodhouse *et al.* 2021). As genetic differences within a population can result in CpG differences, these pangenomes provide a valuable resource to study DNA methylation changes between diverse groups of individuals of the same species. However, where a pangenome is unavailable but breed-specific genomes exist, it is still possible to gain insight into DNA methylation changes and how they may contribute to phenotypic differences between individuals within the same species.

The two main lineages of modern cattle breeds are generally accepted to have been derived from two separate domestication events of the wild auroch (*Bos primigenius*) (McTavish *et al.* 2013). The first domestication event occurred in the Fertile Crescent around 10,000 years ago and gave rise to *Bos taurus taurus* from the wild auroch, *B. p. primigenius* (Bruford *et al.* 2003; Ajmone-Marsan *et al.* 2010; MacHugh *et al.* 2017). A second domestication event occurred in the Indus Valley, ~1,500 years later, from *B. p. nomadicus*, which separated from the *B. p. primigenius* around 250-330,000 years ago (Loftus *et al.* 1994) and gave rise to *Bos taurus indicus*. The subspecies are referred to here as taurine and indicine cattle, respectively (McTavish *et al.* 2013), where the Angus breed represents taurine cattle, and Brahman is representative of indicine cattle.

Angus and Brahman have contrasting phenotypes, e.g., Angus have been bred for meat production traits (Elzo *et al.* 2012), whereas Brahman have superior heat and disease tolerance traits (Dikmen *et al.* 2018; Goszczynski *et al.* 2018). DNA methylation differences may partly be responsible for the phenotypic differences between these two breeds.

As expected from their domestication history, Angus and Brahman cattle represent genetically highly diverged subspecies (Decker *et al.* 2014; Koren *et al.* 2018). However, as they produce fertile offspring when mated (Hiendleder *et al.* 2008), they are an appropriate model to investigate the impact of using a single reference genome on methylome analysis of two genetically diverse populations. We have previously produced high-quality haplotype-resolved reference genomes for Angus and Brahman (Low *et al.* 2020), enabling us to thoroughly evaluate the impact of reference genome choice in WGBS analysis.

Breed-specific differences in CpGs may occur due to a SNP, such as those caused by spontaneous deamination (Żemojtel *et al.* 2011; Yang *et al.* 2021), or may result from SVs. A single SNP affecting a CpG site has been shown to drastically alter the methylation state of the *Igf2* gene in pigs leading to changes in muscle development (Van Laere *et al.* 2003). SVs have been associated with decreased methylation in cancers (Zhang *et al.* 2019) and with changes in the methylation of the kappa opioid receptor *(kor)* promoter associated with KOR dysfunction and schizophrenia (Lutz *et al.* 2018).

Parent-of-origin effects (POEs) occur when only one allele is expressed, and the phenotype in the offspring may depend on which parent contributed the expressed allele (Lawson *et al.* 2013). Reciprocal crossing is necessary to elucidate how each parent contributes to a particular phenotype. POEs have been observed in hybrids of mice (Shi *et al.* 2005), cattle (Vaughn *et al.* 2022) and pigs (Pan *et al.* 2012), and there is increasing evidence that fetal development is influenced by POEs (Doria *et al.* 2010; Piedrahita 2011; Yuen *et al.* 2011; Moore *et al.* 2015; Eggermann *et al.* 2021).

To investigate the potential impact of reference genome choice on methylome analyses and to improve our understanding of the genetic and epigenetic factors driving the phenotypic differences between cattle subspecies, we used WGBS data from 24 fetal liver samples of purebred Brahman and Angus cattle and their reciprocal-crosses to perform a comprehensive assessment of the impact of reference genome choice on differential methylation and gene expression. This study serves as an example of how to investigate epigenetic differences between breeds, strains, and populations within species and informs about reference genome effects on the interpretation of methylome analyses.

Methods

Study Animals and Sample Collection

All animal experiments and procedures described in this study complied with Australian guidelines, approved by the University of Adelaide Animal Ethics Committee and followed the ARRIVE Guidelines (https://arriveguidelines.org/) (Approval No. S-094-2005). Liver tissue samples from concepti were the same as those described in Liu *et al.* (2021). Briefly, the parents were purebred Angus (*B. t.*

taurus) and purebred Brahman (*B. t. indicus*), herein denoted as BT and BI. Primiparous females and their fetuses were ethically sacrificed at day 153±1 of gestation. Concepti were dissected, and tissue samples snap-frozen in liquid nitrogen and stored at -80°C until further use. Liver samples from three female and three male individuals from each of the four genetic combinations: BT x BT, BT x BI, BI x BT, and BI x BI were used.

DNA extraction and sequencing

DNA was extracted from frozen fetal liver tissues using Qiagen® DNeasy® Blood & Tissue Kit following the manufacturer's instruction and sent to BGI Hong Kong, China, for WGBS library preparation and sequencing. Bisulfite conversion was performed using the Zymo Research™ EZ DNA Methylation™ - Gold Kit (D5005). All samples were sequenced in a single batch, and each sample was sequenced to ~30X coverage using the BGI DNB-seq.

RNA was extracted from frozen fetal liver tissues using Illumina® RiboZero Gold kits following the manufacturer's instruction and prepared for Illumina RNA-seq short-read sequencing. The RNA-seq protocol and data availability (GEO accession number: GSE148909) have been described in our previous work (Liu *et al.* 2021). The same tissue samples were used in both the RNA-seq and WGBS. Individual sample names and their corresponding genetic group are given in Supplementary Table 11.

WGBS mapping

WGBS reads were mapped using the MethylSeq Nextflow pipeline (v. 1.6.1) (Di Tommaso *et al.* 2017) with the '--zymo' trimming parameter. The reads were first checked for quality with FastQC (v. 0.11.9) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), then adapters were trimmed using Trim Galore (v. 0.6.6) (https://github.com/FelixKrueger/TrimGalore), and the reads were reassessed for quality post-trimming. Trimmed reads passing qvalue ≥ 20 were mapped to the Brahman (GCA_003369695.2) and

Angus (GCA_003369685.2) genomes (Low *et al.* 2020) using BWA-Meth (0.2.2) (arXiv:1401.1129). The non-pseudo autosomal region of the Angus Y chromosome was added to the Brahman reference. This step enabled us to include the Y chromosome sequence whilst avoiding duplication of the pseudoautosomal region. Both Brahman and Angus chromosome sequences were reorientated to match the orientation of ARS-UCD1.2 chromosomes (Rosen *et al.* 2020). After sorting the alignment files with SAMtools (v. 1.11) (Li *et al.* 2009), duplicates were marked with Picard (v. 2.25.4) (https://broadinstitute.github.io/picard/). Bam file quality control was performed with the bamqc function from qualimap (v. 2.2.2d) (Okonechnikov *et al.* 2016) by setting the '-gd' parameter to HUMAN. Methylation calls were extracted using MethylDackel (v. 0.5.2) (https://github.com/dpryan79/MethylDackel) extract with the parameter '—minDepth 10' and output in MethylKit (Akalin *et al.* 2012) format ('—methylKit') and a more generic cytosine report ('—cytosine_report'). In-house scripts were used to convert the MethylDackel output for use with DNMTools https://dnmtools.readthedocs.io/en/latest/ (see https://github.com/DaviesCentreInformatics/Brahman_Angus_WGBS). All samples had a bisulfite conversion efficiency of >99%. All downstream analyses only used CpGs from autosomes with ≥ 10X coverage.

Identification of shared and breed-specific CpG sites

For a given autosome, we extracted 1000bp around all CpGs that were not in the first 500bp or last 500bp of the chromosome; this yielded sequences that were 1002bp long. We then mapped the 1002bp CpG sequences from one subspecies to the reference of the other. We used minimap2 (v. 2.24) (Li 2018) with the 'map-hifi' preset to align CpGs from a given chromosome in one breed to the same chromosome in the other breed; alignments were sorted using SAMtools (v 1.11) (Li *et al.* 2009). Once the long sequences were aligned, we filtered the BAM file and considered all alignments where at least 900 bp were successfully aligned to the reference. We then used the Align package from BioPython (v 1.80) (Cock *et al.* 2009) to perform a local alignment between the 102bp sequences taken from the midpoint of the query and reference. We then recorded which CpG sites were shared between Brahman and Angus, which CpG sites differed, and which could not be aligned during the initial minimap2 alignment step (S. table 6). We performed subsequent analyses using all CpGs present on the autosomes for each reference and again using only the shared CpGs that passed the ≥ 10X coverage criteria. The CpG sites that could not be aligned in the initial alignment step with minimap2 (i.e. a genomic region that is present in one breed but missing in the other breed) or constituted a SNP were considered breed-specific CpG sites. All steps described in this section were performed for both Brahman and Angus reference genomes.

Identification of differentially methylated regions

The methylKit package (v. 1.22.0) (Akalin *et al.* 2012) was used to identify DMRs between breed and POE groups. We investigated breed effects by comparing

BIBI samples with BTBT samples. The reference group was always the breed that matched the reference genome. i.e., BIBI samples were the reference group when reads were aligned to the Brahman reference. Any DMRs identified were either hypo- or hypermethylated with respect to this reference group. To study POEs, we investigated the maternal effects by comparing samples with BIBI dams (BIBI; BTBI) and those with BTBT dams (BTBT; BIBT). Similarly, to study the paternal effects, we compared those with BIBI sires (BIBI; BIBT) and those with BTBT sires (BTBT; BTBI). For each comparison, we removed all CpG sites with less than 10X coverage and more than the 99.9th percentile of coverage. Reads with too high coverage (e.g. from PCR duplication bias) can impair the accurate determination of the methylation percentage at that site (Akalin *et al.* 2012). We then normalized the coverage using the default methylKit normalization strategy. We merged the CpG counts per group using the ‘unite’ function with ‘destrand = T’ and ‘min.per.group = 5L’ so that a given CpG site had to be covered by at least ten reads in five out of six samples per group. For the parent of origin DMR analysis, we set ‘min.per.group = 10L’.

We then identified differentially methylated cytosines between groups using the ‘calculateDiffMeth’ function, with sex as a covariate in the model. To determine differentially methylated regions, we used the ‘tileMethylCounts’ function with default parameters to divide the genome into regions for differential methylation analysis. This step allowed methylKit to divide the genome into non-overlapping regions based on the tiling windows. Briefly, methylKit models the methylation at a given cytosine or region by fitting a logistic regression:

$$\log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 * T_i$$

P_i denotes the methylation proportion for sample i in samples 1,…,*n*, where *n* is the number of samples across both groups in the comparison (Akalin *et al.* 2012). T_i represents the groups (0 for group 1, 1 for group 2). β_0 denotes the log odds of group 1 (fraction of reads reporting C / 1 – the fraction of reads reporting C). β_1 denotes the log odds ratio between the groups. For further details, refer to Akalin *et al.* (2012). We retained all DMRs with a difference in methylation of ≥ 10% and a qvalue of ≤ 0.01 for further analysis. An overview of the samples, reference genomes, types of CpGs and DMR analysis is given in Fig 1A-G.

Identification of SNPs and SVs between Angus and Brahman genomes

We used MUMmer (v. 4.0.0) (Marçais *et al.* 2018) to identify SNPs and generate input files for Assemblytics (v. 1.2.1) (Nattestad & Schatz 2016). Briefly, 'dnadiff' from MUMmer with default parameters was used to align the Brahman and Angus autosomes. We extracted SNPs for all autosomes from the '.snps' file generated by 'dnadiff'. The delta file generated in this step was used as input to Assemblytics. The output from Assemblytics was then used to identify SVs. We considered SVs as variants greater than or equal to 50 bp in length.

DMR coordinate conversion

To determine if a given DMR changed methylation direction between genomes, we had to convert the coordinates of DMRs identified by alignment with the Angus genome to Brahman coordinates and vice versa. We considered a DMR as changing methylation direction if, for example, it is hypomethylated in BIBI samples compared to BTBT samples when using the Brahman reference but becomes hypermethylated in BIBI using the Angus reference genome. To investigate this, we first converted the DMR bed files to GTF files and then used Liftoff (v.1.6.2) (Shumate & Salzberg 2021) to transfer coordinates from one reference genome to the other. We then identified DMRs reciprocally overlapping one another by at least 90% between the two genomes, with these DMRs being considered successfully lifted over. DMRs that did not overlap by 90% were not considered for the methylation direction change analysis.

RNA-seq mapping and pre-processing

RNA-seq reads were mapped to the Brahman and Angus genomes as in the WGBS mapping step. Briefly, reads were checked for quality using FastQC (v. 0.11.4) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) before being trimmed with Trim Galore (v. 0.4.2) (https://github.com/FelixKrueger/TrimGalore) with the parameters '--quality 10' and '--length 100'. Reads were mapped using HiSAT2 (v. 2.1.0) to both the Brahman and Angus reference genomes (Low *et al.* 2020); alignment files were sorted using SAMtools (v. 1.10) (Li *et al.* 2009). FeatureCount from the Rsubread package (v. 2.10.5) (Liao *et al.* 2019) was used to count how many reads mapped to genes.

Differential gene expression

Differential gene expression analysis was performed using an in-house R script similar to previously published work (Liu *et al.* 2021), with the edgeR (v. 3.38.4) (Robinson *et al.* 2009) and limma (v. 3.52.4) (Ritchie *et al.* 2015) R packages. The genome annotation was based on Ensembl v.104 for Brahman and Angus. The orientation of the genes was reversed where necessary to correspond with the orientation of the chromosomes of ARS-UCD1.2. Briefly, genes that had fewer than

0.5 counts per million (CPM) in fewer than three samples were removed from the dataset. Reads were normalized using the trimmed mean of M-values and then weighted using ‘voomWithQualityWeights’. As two samples were sequenced in a separate batch, we added the batch as a term to the model to account for any variability introduced by the separate sequencing run. We then compared differential gene expression between purebred Angus and Brahman, Angus dams and Brahman dams, and Angus sires and Brahman sires. Genes with significant differences in gene expression at an adjusted p-value ≤ 0.05 were retained for further analysis.

Identifying imprinted genes

We downloaded a list of genes with evidence of imprinting in human, mouse and cattle from Morison *et al.* (2005) and https://www.geneimprint.org. We then used OrthoFinder to identify human orthologs of both Brahman and Angus genes (Emms & Kelly 2019), allowing us to assign Human Genome Organisation Gene Nomenclature Committee (HGNC) symbols to genes in each breed. To do this, we first identified which Brahman proteins had orthologs in human. We then identified the genes that encoded these proteins and used this information to assign human and Brahman genes as orthologs. We repeated the process for the Angus genes. We then identified all genes that could be assigned an HGNC symbol from the Brahman Ensembl annotation version 104 that were also present in the imprinted gene list (S. table 12). This filtering gave us 80 imprinted genes for Brahman autosomes. We repeated the process for Angus using the Angus Ensembl annotation version 104 and identified 79 imprinted genes. The discrepancy is due to one imprinted gene for Angus occurring on an unplaced scaffold.

Linking DMRs to DEGs

For each DEG, we considered five different regions in and around the gene where DMRs might have an influence. These regions included putative enhancer regions, 5kb outside the gene body, and the gene body itself (S. figure 11). The upstream putative enhancer region started 130kb upstream of the gene and then stopped 5kb upstream of the gene body for a total length of 125kb. We repeated this for the downstream putative enhancer region, starting 5kb downstream of the gene body and extending out 125kb. This putative enhancer region length was based on the median distance between enhancers and their gene targets (Jin *et al.* 2013). The 5kb region was from upstream of the start of the gene body to the start of the gene body. Again, this was repeated for the downstream 5kb region. The gene body was the region annotated as "gene" in the Ensembl annotation file. We then found all DMRs that overlapped these regions by at least 90% of their length using 'bedtools intersect' with the '-f' and '-F' arguments, both set at 0.9 and the '-e' argument set to True.

Results

Mapping of WGBS data and calling CpG

Each of the 24 samples representing the four genetic groups (Fig 1A; S. table 1) was sequenced for WGBS analysis to at least 30X coverage and then mapped separately to the Brahman and Angus genomes (Fig 1B). An average mapping rate of ~95% was achieved when reads were mapped from each sample to the Brahman reference genome (Table 1; S. table 1). All samples had at least 10X coverage for 93% of the Brahman sequence. Using the Angus reference, all samples had 10X coverage for at least 90% of the sequence (Table 1; S. table 1).

We performed all analyses twice for each reference genome, first using all CpGs with ≥10X in each reference genome and again where we retained only CpG sites with ≥10X that we could confidently assign as being shared between both breeds. Between 85% and 88% of autosomal CpG sites had coverage ≥ 10X when considering all CpG sites on both the Brahman and Angus reference and shared CpG sites (Table 1; S. table 2). Median coverage of CpG sites across all samples ranged from 25-34X regardless of reference and CpG sites considered (i.e., shared or all) (Table 1; S. table 3).

Clustering of genetic groups

Comparing the methylation patterns between the genetic groups, we found that samples within a genetic group were more similar to each other than with samples from other groups. For example, samples from the BTBT group had higher correlations with other BTBT samples than BIBI samples. BTBT had the highest within-group Pearson correlations (*r* between 0.81 and 0.88) (S. figure 1). The samples that were least correlated with one another were those belonging to BTBT and BIBI, with correlations between 0.75 and 0.78. Samples from the reciprocal cross groups (BIBT; BTBI) had similar correlations with other samples within their own group (*r* between 0.81 and 0.83) as well as with samples from the alternative reciprocal cross (*r* between 0.80 and 0.83). Overall, correlations were high within each genetic group (*r* ≥ 0.8) (S. figure 1).

We performed a principal component analysis of the 24 samples using CpG sites covered by at least 10 reads in all samples (Fig 2A). BTBT and BIBI formed distinct clusters distant from one another, with the two hybrid genetic groups clustering much closer together and between the two parental genetic groups. Nevertheless, the hybrid groups were clearly separated on the PCA plot (Fig 2A). The separation of groups we observed from the methylation data was similar to that seen for the gene expression data (Fig 2B).

Overview of DNA methylation patterns

Samples had global mean CpG methylation of between 47-62%, with most samples ranging from 49-54% (Fig 3A; S. table 4). Mean exon CpG methylation was 48-58% for all samples, with most samples ranging from 48-54% methylation (S. Fig 2; S. table 5). The 5' UTRs and promoter regions had the lowest mean CpG methylation percentage across all samples, between 9-14% and 14-19%, respectively (S. Fig 3-4; S. table 5). The intergenic regions displayed mean methylation levels that ranged from 47 – 65% (S. Fig 5; S. table 5), with most samples ranging from 49-57%, similar to the global mean. The introns exhibited slightly higher methylation levels, with means ranging from 51-66% (S. Fig 6; S. table 5), and most samples in the 53-59% range. The 3' UTRs revealed the highest overall CpG methylation levels, 57-69% (S. Fig 7; S. table 5). Lastly, the predicted enhancers, according to MacPhillamy *et al.* (2022), exhibited CpG methylation levels ranging from 42-53%, with most samples within 43-48% methylated (S. Fig 8; S. table 5). Similar methylation patterns were observed using only shared CpGs in exons, 5'UTRs, intergenic, introns, promoters, predicted enhancers, and 3'UTRs, regardless of the reference genome used.

Shared and breed-specific CpGs

We were able to confidently identify 74-75% of CpGs in the Brahman and Angus genomes that were shared between the two breeds (Table 2; S. table 6). We found that around four per cent of CpG alignments contained a SNP between reference genomes (S. table 6; S. figure 9), i.e., were breed-specific. About 22% of CpGs could not be confidently assigned as shared or breed-specific and so were not considered in the shared CpG analysis. By definition, breed-specific regions with CpG sites did not align when the other breed genome was used as the reference. We found ~1% of such CpG sites. In total, the SNP change and breed-specific categories of CpG sites constituted 4.7% and 4.9% of CpGs between the Angus and Brahman reference genomes, respectively, and were considered breed-specific.

Enrichment of SNPs affecting CpG sites

Using the autosomal SNPs identified by MUMmer, we observed that Brahman and Angus autosomal sequences differ by an average of ~0.6% (S. table 7). SNPs in CpG sites were found to be enriched by ~8 times compared to the genome-wide average (binomial test, p-value < 5×10^{-324}), which means there is a higher level of divergence between Angus and Brahman at CpG sites than at other autosomal sites. Looking more closely at the CpG SNP changes, we found that most (~81%) of the CpG SNP were either C to T or G to A changes (S. figure 7). The remaining SNP changes individually comprised less than ~10% of the total observed mutations at CpG sites (S figure 9; S. table 6).

Increased number of CpGs within structural variants

Next, we tested whether CpGs were enriched in SVs compared to the rest of the genome. Using the Brahman genome as the reference, we observed 25,009 SVs between Brahman and Angus, making up ~27Mb of sequence. We observed 1.13-fold more CpGs within SVs than in non-SVs (binomial test, p-value < 5×10^{-324}). When considering CpGs affected by SNPs and SVs, the CpG mutation rate is approximately 4% between Brahman and Angus compared to the genome-wide mutation rate of around 1.7%.

Choice of reference genome influences methylome results

We observed a statistically significant difference in global CpG methylation in BTBT, BIBT and BIBI depending on whether the Angus or Brahman genome was used to quantify methylation at 10X coverage (paired Wilcoxon test, p-value < 0.05) (Figure 3A; S. figure 10). Although not statistically significant, the BTBI group (paired Wilcoxon test, p-value = 0.06) approaches significance. The reference genomes showed no difference in global methylation when only shared CpG sites were considered (Figure 3B). When comparing global CpG methylation differences between samples mapped to Brahman versus those mapped to Angus, the largest quantification bias was ~2% for BTBT. The other quantification biases were ~0.8%, ~0.7% and ~0.3% for BTBI, BIBT and BIBI samples, respectively (Figure 3A).

To further investigate the influence of the reference genome on downstream analyses, we compared DMRs identified by the two reference genomes to evaluate if the direction of methylation changed, i.e., hypermethylated vs hypomethylated and vice versa. Approximately 12% (28,922) of Angus DMRs overlapped with Brahman DMRs by at least 90% of their length. Of the DMRs that mapped to the Angus reference, 3,575 showed changes in methylation direction when mapped to the Brahman reference (S. table 8). We observed similar numbers (3,581) when lifting DMRs from Brahman to Angus (S. table 8). There were no methylation direction changes when we considered differentially methylated cytosines (DMCs).

Breed-specific CpGs show distinct methylation patterns

Looking more closely at the breed-specific CpGs, we first determined whether a given breed-specific CpG was methylated or not. We binned individual CpGs into "unmethylated" (CpG methylation ≤ 35%) and "methylated" (CpG methylation ≥ 65%). CpGs that were between 35% and 65% were considered "hemimethylated" and were excluded from this analysis. We only considered breed-specific CpGs with at least 10X coverage in all purebred samples mapped to their respective genome. We then randomly sampled 1000 CpG sites for 100 iterations, recording the number of "methylated" and "unmethylated" CpGs for each breed. We observed significantly more Brahman-specific CpGs as hypomethylated than hypermethylated (Mann-Whitney U-test, p $=2.5 \times 10^{-34}$). Interestingly, we observed the inverse when considering Angus-specific CpGs; significantly more Angus-specific CpGs were hypermethylated than hypomethylated (Mann-Whitney U-test, p $=2.5 \times 10^{-34}$) (Figure 3C).

Breed-specific DMRs show poor association with DEGs

As we observed a quantification bias when using all CpGs mapped against each reference genome, we restricted breed-specific and POE analyses to those CpGs identified as shared. Additionally, we examined the number of DMRs at the 25% and 50% difference thresholds, i.e., more stringent thresholds for calling DMRs which substantially reduced the numbers (S. table 9). Given that minor changes of less than 10-15% in methylation have been observed to influence gene expression and phenotype (Leenen *et al.* 2016; Thomson *et al.* 2022), we used a difference threshold of 10% to interpret the results.

Using Brahman as the reference, we identified 123,602 DMRs and 1,397 DEGs (S. table 9; S. table 10). Of the 123,602 DMRs observed, ~20% (25,367) overlapped with the surrounding region of the significant DEGs, with most (~69%) falling within the putative-enhancer region. Only 0.8% of the DMRs overlapped with promoters of DEGs. When the Angus reference was used, of the 125,544 DMRs identified, ~32% (40,787) of those overlapped with a DEG. Most (~64%) of these DMRs fell into the putative enhancer region, while only 0.8% of the DMRs overlapped with a DEG promoter, despite substantially more (2,151) DEGs observed (S. table 10).

We then examined the overlap of DMRs and imprinted genes, first using Brahman as the reference. Here, ~1% (1,282) of the DMRs identified between BIBI and BTBT overlapped 79 imprinted genes. Only one imprinted gene, Par-6 family cell polarity regulator gamma (*Pard6g*), did not overlap with any DMR. Most DMRs (~72% of the 1,282) that overlapped an imprinted gene fell into putative enhancer regions. Seven imprinted genes were significantly differentially expressed when comparing BIBI and BTBT (Table 3). These genes were SPARC related modular calcium-binding 1 (*Smoc1*), neuronatin (*Nnat*), ganglioside induced differentiation associated protein 1 like 1 (*Gdap1l1*), diacylglycerol O-acyltransferase 1 (*Dgat1*), solute-carrier family 22 member 18 (*Slc22a18*), potassium voltage-gated channel subfamily Q member 1 (*Kcnq1*) and protein phosphatase 1 regulatory subunit 9A (*Ppp1r9a*). *Nnat* and *Gdap1l1* had higher expression in BTBT, and the remaining five DEGs, *Smoc1*, *Dgat1*, *Slc22a18*, *Kcnq1* and *Ppp1r9a*, had higher expression in BIBI. We observed seven imprinted DEGs using the Angus reference; however, the genes identified differed: estrogen receptor 2 (*Esr2*), *Ppp1r9a,* and DLG-associated protein

2 (*Dlgap2*) had higher expression in BIBI (Table 4), while *Nnat* and zinc finger protein 90 (*Zfp90*) had higher expression in BTBT.

Dam-of-origin methylation

To investigate the dam-of-origin effects (DOEs), we compared samples with Brahman dams (BIBI and BTBI) with those with Angus dams (BTBT and BIBT). Using the Brahman genome as the reference, 243 DEGs were identified in the DOE comparison. Around 3% (497) of the DMRs overlapped with DEGs. Most DMRs (~357) fell into the putative enhancer region; 0.9% of DMRs overlapped with a DEG promoter (S. table 9; S. table 10). There were 52 imprinted genes that overlapped with ~1% (188) of the DMRs identified in the comparison. Of the 188 DMRs that overlapped with an imprinted gene, 125 DMRs overlapped with the putative enhancer region. Only two imprinted genes were significantly differentially expressed (Table 3). Zinc finger CCCH-type containing 12C (*Zc3h12c*) and *Ppp1r9a* had higher expression in samples with Brahman mothers. Using Angus as the reference, we observed 809 (~4%) DMRs overlap with the 260 DEGs. Most (592) of these DMRs fell into the putative enhancer regions. The same imprinted DEGs (*Zc3h12c* and *Ppp1r9a*) were identified using the Angus genome as the reference (Table 4).

Sire-of-origin methylation may be driving differential gene expression

To investigate the sire-of-origin effects (SOEs), we compared samples with Brahman sires (BIBI and BIBT) with those with Angus sires (BTBT and BTBI). Using the Brahman reference, we identified 62,056 DMRs and 1,364 DEGs in the sire group comparison using shared CpGs (S. table 9; S. table 10). There were 63,516 DMRs identified using the Angus reference, but substantially more DEGs (2,137) were identified (S. table 10). Around 15% (~9,300) of the DMRs overlap with a significant DEG; most (72% of the 9,300) of those overlaps were in a putative enhancer region. Using the Angus reference, ~27% (~17,100) of DMRs overlapped a DEG, with most (~70%) of the DMRs overlapping a putative enhancer region. One per cent and 0.7% of DMRs overlapped with a DEG promoter when using either the Brahman or Angus as the reference.

We observed 73 imprinted genes that overlapped DMRs identified between the two different sire groups. Less than 1% (586) of DMRs overlapped the 73 imprinted genes, with most (~73% of 586) occurring in the putative enhancer region. Seven imprinted genes were significantly differentially expressed and overlapped with a DMR (Table 3). These genes were DS cell adhesion molecule (*Dscam*), 5-hydroxytryptamine receptor 2A (*Htr2a*), *Nnat*, *Dgat1*, AXL receptor tyrosine kinase (*Axl*), necdin MAGE family member (*Ndn*), and tissue factor pathway inhibitor 2 (*Tfpi2*). *Dscam*, *Nnat*, *Ndn* and *Tfpi2* had higher expression in samples with Angus sires, with the other genes being more highly expressed in samples with Brahman sires. *Slc22a18* showed high expression in the Brahman sire group but did not overlap with any DMRs. More significantly differentially expressed imprinted genes were observed using the Angus than the Brahman reference (Table 4). In this case, the ten significantly differentially expressed imprinted genes with DMR overlap were *Dscam*, *Esr2*, *Htr2a*, *Nnat*, succinate dehydrogenase complex subunit D (*Sdhd*), *Axl*, makorin ring finger protein 3 (*Mkrn3*), *Ndn*, *Dlgap2* and *Slc22a18*. *Dscam*, *Nnat*, *Sdhd*, *Mkrn3* and *Ndn* had higher expression in samples with an Angus sire. The remaining five genes (*Esr2*, *Htr2a*, *Axl*, *Dlgap2*, *Slc22a18*) had higher expression in samples with a Brahman sire. *Dgat1* did not overlap any DMRs when using the Angus reference.

Discussion

In the present study, we observed genome-wide CpG methylation correlations among replicates that ranged from 75% to 82% between groups and from 81% to 87% within groups. These correlations were similar to a recent study in mice where genome-wide CpG methylation correlations among replicates ranged from 73% to greater than 80% (He *et al.* 2020). Moreover, we observed levels of liver global CpG methylation between 47-62% in the present study, which is similar to previous studies of human (Hama *et al.* 2018), mouse (He *et al.* 2020) and cattle (Zhou *et al.* 2020).

Mapping statistics can provide an insight into how the choice of reference genome will affect downstream analyses (Valiente-Mullor *et al.* 2021). However, we observed negligible differences in raw mapping statistics regardless of whether the Angus or Brahman reference genomes were used. Additionally, the global methylation quantification bias observed was less than 2%, depending on the reference genome used. This quantification bias is lower than the 7-9% quantification bias found in the mouse genome, depending on the reference genome used (Wulfridge *et al.* (2019). The extent of this bias is influenced by the divergence between reference genomes and whether the breed-specific CpGs tend to be hypo- or hypermethylated. Brahman and Angus have a CpG divergence of ~4%, whereas the mouse genomes analysed by Wulfridge *et al.* (2019) had a CpG divergence of 10.7%. The bias we observed was greatest in the BTBT samples when the Brahman genome was used as the reference, most likely because the Angus-specific CpG sites tended to be hypermethylated. Conversely, the quantification bias was lower in the other genetic groups, possibly due to the hypomethylation in Brahman-specific CpG sites.

Spontaneous deamination of methylated CpG to TpG is the most common dinucleotide mutation in the mammalian genome (Żemojtel *et al.* 2011; Yang *et al.* 2021). We observed around 800,000 C-T or G-A mutations between Brahman and Angus. A recent study observed 34,677 SNPs affecting CpG sites between indicine and taurine genomes (Capra *et al.* 2023). The difference in the number of SNPs between the two studies is likely due to Capra *et al.* (2023) having used reduced representation bisulfite sequencing with substantially lower coverage than the present study and that they only considered SNPs that affected CpG sites. When differential methylation analysis is performed, a breed that has lost the C (mutated to T) will be reported as having 0% methylation at that site when, in fact, there is no CpG present. This incorrect identification of an unmethylated site can then severely impact the interpretation of results.

SVs have been associated with various traits in humans, including HIV-1 susceptibility (Gonzalez *et al.* 2005), autism (Kumar *et al.* 2008; Marshall *et al.* 2008; Weiss *et al.* 2008) and carcinogen metabolism (Bell *et al.* 1993). In livestock, SVs have been implicated in diverse traits ranging from horn (polled) status (Rothammer *et al.* 2014; Lamb *et al.* 2020) to bulldog calf syndrome (Jacinto *et al.* 2020). The SVs between Brahman and Angus have significantly more CpGs than the background genome, potentially introducing CpGs with important regulatory effects. However, due to their presence in only one subspecies, a single reference genome will fail to account for these breed-specific CpG sites. Therefore, the phenotypic differences between the two breeds may be influenced by CpGs that cannot be compared accurately with a single reference genome if they are in breed-specific regions.

We observed relatively few changes in methylation direction, and these were likely to be an artifact of how the genome was tiled and possible erroneous alignments in the coordinate conversion. A common step of some DMR callers is to perform window tiling of the genome to identify DMRs or to enable analysis when coverage is low (Akalin *et al.* 2012; Park *et al.* 2014; Kishore *et al.* 2015). SNPs and SVs can potentially complicate analyses when genome tiling is used to identify DMRs, as a single reference genome cannot account for these variants. Although we observed no directional changes when considering DMCs, it is possible that some DMRs were specific to the reference genome, which affected the analysis. Researchers should be careful when using genome tiling in methylation analyses that compare breeds, strains or populations.

Most DMRs identified in this study were not associated with DEGs. However, of the DEGs that overlapped with a DMR, there was a tendency for the overlap to occur more frequently in the putative enhancer region than in promoters or DEG bodies. This trend suggests that differential methylation of enhancers may impact gene expression differences in bovine fetal liver. Indeed, a growing body of evidence suggests that enhancer methylation is important in embryonic and fetal development (Lee *et al.* 2015; Slieker *et al.* 2015; He *et al.* 2020; Alajem *et al.* 2021).

There were more significant DEGs when mapping to the Angus reference than the Brahman reference. Interestingly, genes that were DE using the Brahman reference were not always DE when using the Angus reference. The choice of reference genome has been shown to impact differential expression analysis in rice (Slabaugh *et al.*

2019), bacteria (Price & Gibas 2017) and human (Wu *et al.* 2013; Kaminow *et al.* 2022) when using short-read RNA-seq. When the reference genome better represents the individuals being studied, more reads can be uniquely aligned to the correct position, providing a more accurate estimate of gene expression.

We identified several interesting DEGs associated with DMRs, particularly imprinted genes. Among these was *Dgat1*, which is involved in fat metabolism in milk production (Khan *et al.* 2021), feed conversion and adipogenesis (Abeel *et al.* 2009; Khan *et al.* 2021). Several studies have investigated the role of *Dgat1* in weight gain (Zhang *et al.* 2010; Tsuda *et al.* 2014); expression of *Dgat1* is necessary for weight gain, especially when the caloric density of food is high (Zhang *et al.* 2010). We found differential expression of *Dgat1*, with higher expression in Brahman than Angus (BIBI vs BTBT) and when Brahman was the sire (BIBI, BIBT vs BTBT, BTBI). Taken together, the breed and sire of origin comparisons suggest that the breed of the sire may be an important determinant in the expression of this gene. Higher *Dgat1* expression may result from adaptation to poor feed quality, e.g. Elzo *et al.* (2009) observed better feed conversion efficiency in Brahman compared to Angus and Brahman x Angus cattle. Regulation of *Dgat1* expression may occur via DNA methylation, as there is a DMR ~42kb downstream of the transcription start site, which was identified in both the breed-specific and SOE comparisons.

Parent-of-origin DMRs may change how cis-regulatory elements interact with target genes and influence gene expression in the offspring (Giannoukakis *et al.* 1993; Lawson *et al.* 2013). It has been observed that parent-specific methylation can alter the cis-regulatory landscape around certain genes, such as *Igf*2 (Szabo *et al.* 2000;

Yang *et al.* 2003). DMRs may influence the DEGs and, ultimately, help drive the differences in phenotype. However, to confidently assign gene expression and DMRs to a particular parent, long-read sequencing (Ren *et al.* 2023) is needed to identify variations that link the sequences to the parent of origin. Additionally, the use of reciprocal crosses will enable one to investigate if a combination of breed and sex of the parent impacts which transcript is expressed.

SNPs and SVs have been shown to complicate and bias analyses in several studies (Wu *et al.* 2013; Price & Gibas 2017; Slabaugh *et al.* 2019; Wulfridge *et al.* 2019; Kaminow *et al.* 2022). In our analysis, we observed an enrichment of SNPs affecting CpGs between Brahman and Angus. Capra *et al.* (2023) also reported a higher frequency of breed-specific SNPs around DMCs in a study of indicine and taurine cattle. This finding suggests that genetic differences between the two breeds may contribute to epigenetic variations. Using individual animal genomes in the study to account for genetic variations, as Wulfridge *et al.* (2019) suggested, would enhance the accuracy for each individual. However, despite decreasing sequencing costs, the cost will likely be prohibitive in most livestock contexts. A possible solution was explored in a recent study comparing methylation in taurine and indicine cattle (Capra *et al.* 2023). Here the authors used genotyping by sequencing to exclude SNPs affecting CpG sites from the analysis (Capra *et al.* 2023). While this simplifies downstream analysis, it may also remove CpGs involved in the phenotypic differences between the two breeds, representing a limitation of the present study and that of Capra *et al.* (2023). An alternative approach to using single reference genomes is the utilisation of pan-genomes, which encompass the majority of variations within the population (Paten *et al.* 2017; Groza *et al.* 2020)**.** This feature is particularly important in the context of DNA methylation studies where, demonstrated in our study, SNPs at CpG sites can exert substantial effects on local methylation information.

Conclusions

This study generated a substantial WGBS dataset derived from two phenotypically diverse cattle breeds which are representative of the two cattle subspecies and highlighted the importance of reference genome choice in methylation analyses. Our findings suggest that the DMRs may primarily exert their influence on enhancer elements rather than promoters. We also identified 11 genes that might be under DMR control. The results underscore the advantages of using the appropriate reference genome for the data set and, through highlighting the limitations of linear reference genomes when comparing DNA methylation between diverse populations, provide additional evidence supporting the incorporation of genome graphs to improve methylation analyses of populations with high genetic divergence.

References

Abeel T., Van de Peer Y. & Saeys Y. (2009) Toward a gold standard for promoter prediction evaluation. Bioinformatics 25, I313-I20. 10.1093/bioinformatics/btp191

Ajmone-Marsan P., Garcia J.F. & Lenstra J.A. (2010) On the origin of cattle: How aurochs became cattle and colonized the world. Evolutionary Anthropology: Issues, News, and Reviews 19, 148-57. https://doi.org/10.1002/evan.20267

Akalin A., Kormaksson M., Li S., Garrett-Bakelman F.E., Figueroa M.E., Melnick A. & Mason C.E. (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biology 13, R87. 10.1186/gb-2012-13-10-r87

Alajem A., Roth H., Ratgauzer S., Bavli D., Motzik A., Lahav S., Peled I. & Ram O. (2021) DNA methylation patterns expose variations in enhancer-chromatin modifications during embryonic stem cell differentiation. PLoS Genetics 17, e1009498. 10.1371/journal.pgen.1009498

Bell D.A., Taylor J.A., Paulson D.F., Robertson C.N., Mohler J.L. & Lucier G.W. (1993) Genetic risk and carcinogen exposure: a common inherited defect of the carcinogen-metabolism gene glutathione S-transferase M1 (GSTM1) that increases susceptibility to bladder cancer. Journal of the National Cancer Institute 85, 1159-64. 10.1093/jnci/85.14.1159

Brandt D.Y.C., Aguiar V.R.C., Bitarello B.D., Nunes K., Goudet J. & Meyer D. (2015) Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. G3 Genes|Genomes|Genetics 5, 931-41. 10.1534/g3.114.015784

Bruford M.W., Bradley D.G. & Luikart G. (2003) DNA markers reveal the complexity of livestock domestication. Nature Reviews Genetics 4, 900-10. 10.1038/nrg1203

Capra E., Lazzari B., Milanesi M., Nogueira G.P., Garcia J.f., Utsunomiya Y.T., Ajmone-Marsan P. & Stella A. (2023) Comparison between indicine and taurine cattle DNA methylation reveals epigenetic variation associated to differences in morphological adaptive traits. Epigenetics 18, 2163363. 10.1080/15592294.2022.2163363

Charlet J., Duymich Christopher E., Lay Fides D., Mundbjerg K., Dalsgaard Sørensen K., Liang G. & Jones Peter A. (2016) Bivalent Regions of Cytosine Methylation and H3K27 Acetylation Suggest an Active Role for DNA Methylation at Enhancers. Molecular Cell 62, 422-31. 10.1016/j.molcel.2016.03.033

Chen N.-C., Solomon B., Mun T., Iyer S. & Langmead B. (2021) Reference flow: reducing reference bias using multiple population genomes. Genome Biology 22, 8. 10.1186/s13059-020-02229-3

Cho J.-W., Shim H.S., Lee C.Y., Park S.Y., Hong M.H., Lee I. & Kim H.R. (2022) The importance of enhancer methylation for epigenetic regulation of tumorigenesis in squamous lung cancer. Experimental & Molecular Medicine 54, 12-22. 10.1038/s12276-021-00718-4

Cock P.J.A., Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B. & de Hoon M.J.L. (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422-3. 10.1093/bioinformatics/btp163

Creyghton M.P., Cheng A.W., Welstead G.G., Kooistra T., Carey B.W., Steine E.J., Hanna J., Lodato M.A., Frampton G.M., Sharp P.A., Boyer L.A., Young R.A. & Jaenisch R. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences 107, 21931-6. 10.1073/pnas.1016071107

Crysnanto D. & Pausch H. (2020) Bovine breed-specific augmented reference graphs facilitate accurate sequence read mapping and unbiased variant discovery. Genome Biology 21, 184. 10.1186/s13059-020-02105-0

Decker J.E., Mckay S.D., Rolf M.M., Kim J., Molina Alcalá A., Sonstegard T.S., Hanotte O., Götherström A., Seabury C.M., Praharani L., Babar M.E., Correia De Almeida Regitano L., Yildiz M.A., Heaton M.P., Liu W.-S., Lei C.-Z., Reecy J.M., Saif-Ur-Rehman M., Schnabel R.D. & Taylor J.F. (2014) Worldwide Patterns of Ancestry, Divergence, and Admixture in Domesticated Cattle. PLoS Genetics 10, e1004254. 10.1371/journal.pgen.1004254

Degner J.F., Marioni J.C., Pai A.A., Pickrell J.K., Nkadori E., Gilad Y. & Pritchard J.K. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25, 3207-12. 10.1093/bioinformatics/btp579

Di Tommaso P., Chatzou M., Floden E.W., Barja P.P., Palumbo E. & Notredame C. (2017) Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316-9. 10.1038/nbt.3820

Dikmen S., Mateescu R.G., Elzo M.A. & Hansen P.J. (2018) Determination of the optimum contribution of Brahman genetics in an Angus-Brahman multibreed herd for regulation of body temperature during hot weather. Journal of Animal Science 96, 2175-83. 10.1093/jas/sky133

Doria S., Sousa M., Fernandes S., Ramalho C., Brandao O., Matias A., Barros A. & Carvalho F. (2010) Gene expression pattern of IGF2, PHLDA2, PEG10 and CDKN1C imprinted genes in spontaneous miscarriages or fetal deaths. Epigenetics 5, 444-50. 10.4161/epi.5.5.12118

Eggermann T., Davies J.H., Tauber M., van den Akker E., Hokken-Koelega A., Johansson G. & Netchine I. (2021) Growth Restriction and Genomic Imprinting-Overlapping Phenotypes Support the Concept of an Imprinting Network. Genes 12. 10.3390/genes12040585

Elzo M.A., Johnson D.D., Wasdin J.G. & Driver J.D. (2012) Carcass and meat palatability breed differences and heterosis effects in an Angus–Brahman multibreed population. Meat Science 90, 87-92. https://doi.org/10.1016/j.meatsci.2011.06.010

Elzo M.A., Riley D.G., Hansen G.R., Johnson D.D., Myer R.O., Coleman S.W., Chase C.C., Wasdin J.G. & Driver J.D. (2009) Effect of breed composition on phenotypic residual feed intake and growth in Angus, Brahman, and Angus x Brahman crossbred cattle. Journal of Animal Science 87, 3877-86. 10.2527/jas.2008-1553

Emms D.M. & Kelly S. (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20, 238. 10.1186/s13059-019-1832-y

Giannoukakis N., Deal C., Paquette J., Goodyer C.G. & Polychronakos C. (1993) Parental genomic imprinting of the human IGF2 gene. Nature Genetics 4, 98-101.

Gonzalez E., Kulkarni H., Bolivar H., Mangano A., Sanchez R., Catano G., Nibbs R.J., Freedman B.I., Quinones M.P., Bamshad M.J., Murthy K.K., Rovin B.H., Bradley W., Clark R.A., Anderson S.A., O'Connell R J., Agan B.K., Ahuja S.S., Bologna R., Sen L., Dolan M.J. & Ahuja S.K. (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307, 1434-40. 10.1126/science.1101160

Goszczynski D.E., Corbi-Botto C.M., Durand H.M., Rogberg-Muñoz A., Munilla S., Peral-Garcia P., Cantet R.J.C. & Giovambattista G. (2018) Evidence of positive selection towards Zebuine haplotypes in the BoLA region of Brangus cattle. Animal 12, 215-23. https://doi.org/10.1017/S1751731117001380

Groza C., Kwan T., Soranzo N., Pastinen T. & Bourque G. (2020) Personalized and graph genomes reveal missing signal in epigenomic data. Genome Biology 21, 124. 10.1186/s13059-020-02038-8

Günther T. & Nettelblad C. (2019) The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLoS Genetics 15, e1008302. 10.1371/journal.pgen.1008302

Hama N., Totoki Y., Miura F., Tatsuno K., Saito-Adachi M., Nakamura H., Arai Y., Hosoda F., Urushidate T., Ohashi S., Mukai W., Hiraoka N., Aburatani H., Ito T. & Shibata T. (2018) Epigenetic landscape influences the liver cancer genome architecture. Nature Communications 9. 10.1038/s41467-018-03999-y

He Y., Hariharan M., Gorkin D.U., Dickel D.E., Luo C., Castanon R.G., Nery J.R., Lee A.Y., Zhao Y., Huang H., Williams B.A., Trout D., Amrhein H., Fang R., Chen H., Li B., Visel A., Pennacchio L.A., Ren B. & Ecker J.R. (2020) Spatiotemporal DNA methylome dynamics of the developing mouse fetus. Nature 583, 752-9. 10.1038/s41586-020-2119-x

Hiendleder S., Lewalski H. & Janke A. (2008) Complete mitochondrial genomes of Bos taurus and Bos indicus provide new insights into intra-species variation, taxonomy and domestication. Cytogenetic and Genome Research 120, 150-6. 10.1159/000118756

Jacinto J.G.P., Häfliger I.M., Letko A., Drögemüller C. & Agerholm J.S. (2020) A large deletion in the COL2A1 gene expands the spectrum of pathogenic variants causing bulldog calf syndrome in cattle. Acta Vet Scand 62, 49. 10.1186/s13028-020-00548-w

Jansz N. (2019) DNA methylation dynamics at transposable elements in mammals. Essays in Biochemistry 63, 677-89. 10.1042/ebc20190039

Jin F., Li Y., Dixon J.R., Selvaraj S., Ye Z., Lee A.Y., Yen C.A., Schmitt A.D., Espinoza C.A. & Ren B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503, 290-4. 10.1038/nature12644

Kaminow B., Ballouz S., Gillis J. & Dobin A. (2022) Pan-human consensus genome significantly improves the accuracy of RNA-seq analyses. Genome Research 32, 738-49. 10.1101/gr.275613.121

Kang Y., Kim Y.W., Kang J. & Kim A. (2021) Histone H3K4me1 and H3K27ac play roles in nucleosome eviction and eRNA transcription, respectively, at enhancers. The FASEB Journal 35. 10.1096/fj.202100488r

Kass S.U., Landsberger N. & Wolffe A.P. (1997) DNA methylation directs a time-dependent repression of transcription initiation. Current Biology 7, 157-65. 10.1016/s0960-9822(97)70086-1

Khan M.Z., Ma Y., Ma J., Xiao J., Liu Y., Liu S., Khan A., Khan I.M. & Cao Z. (2021) Association of DGAT1 With Cattle, Buffalo, Goat, and Sheep Milk and Meat

Production Traits. Frontiers in Veterinary Science 8. 10.3389/fvets.2021.712470

Kishore K., de Pretis S., Lister R., Morelli M.J., Bianchi V., Amati B., Ecker J.R. & Pelizzola M. (2015) methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. BMC Bioinformatics 16, 313. 10.1186/s12859-015-0742-6

Koren S., Rhie A., Walenz B.P., Dilthey A.T., Bickhart D.M., Kingan S.B., Hiendleder S., Williams J.L., Smith T.P.L. & Phillippy A.M. (2018) De novo assembly of haplotype-resolved genomes with trio binning. Nature Biotechnology 36, 1174-82. 10.1038/nbt.4277

Kumar R.A., KaraMohamed S., Sudi J., Conrad D.F., Brune C., Badner J.A., Gilliam T.C., Nowak N.J., Cook E.H., Jr., Dobyns W.B. & Christian S.L. (2008) Recurrent 16p11.2 microdeletions in autism. Human Molecular Genetics 17, 628-38. 10.1093/hmg/ddm376

Lamb H.J., Ross E.M., Nguyen L.T., Lyons R.E., Moore S.S. & Hayes B.J. (2020) Characterization of the poll allele in Brahman cattle using long-read Oxford Nanopore sequencing. Journal of Animal Science 98. 10.1093/jas/skaa127

Lawson H.A., Cheverud J.M. & Wolf J.B. (2013) Genomic imprinting and parent-of-origin effects on complex traits. Nature Reviews: Genetics 14, 609-17. 10.1038/nrg3543

Lee H.J., Lowdon R.F., Maricque B., Zhang B., Stevens M., Li D., Johnson S.L. & Wang T. (2015) Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. Nature Communications 6, 6315. 10.1038/ncomms7315

Leenen F.A.D., Muller C.P. & Turner J.D. (2016) DNA methylation: conducting the orchestra from exposure to phenotype? Clinical Epigenetics 8, 92. 10.1186/s13148-016-0256-8

Li E. & Zhang Y. (2014) DNA methylation in mammals. Cold Spring Harb Perspect Biol 6, a019133. 10.1101/cshperspect.a019133

Li H. (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094-100.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. & Subgroup G.P.D.P. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-9. 10.1093/bioinformatics/btp352

Liao W.-W., Asri M., Ebler J., Doerr D., Haukness M., Hickey G., Lu S., Lucas J.K., Monlong J., Abel H.J., Buonaiuto S., Chang X.H., Cheng H., Chu J., Colonna V., Eizenga J.M., Feng X., Fischer C., Fulton R.S., Garg S., Groza C., Guarracino A., Harvey W.T., Heumos S., Howe K., Jain M., Lu T.-Y., Markello C., Martin F.J., Mitchell M.W., Munson K.M., Mwaniki M.N., Novak A.M., Olsen H.E., Pesout T., Porubsky D., Prins P., Sibbesen J.A., Sirén J., Tomlinson C., Villani F., Vollger M.R., Antonacci-Fulton L.L., Baid G., Baker C.A., Belyaeva A., Billis K., Carroll A., Chang P.-C., Cody S., Cook D.E., Cook-Deegan R.M., Cornejo O.E., Diekhans M., Ebert P., Fairley S., Fedrigo O., Felsenfeld A.L., Formenti G., Frankish A., Gao Y., Garrison N.A., Giron C.G., Green R.E., Haggerty L., Hoekzema K., Hourlier T., Ji H.P., Kenny E.E., Koenig B.A., Kolesnikov A., Korbel J.O., Kordosky J., Koren S., Lee H., Lewis A.P., Magalhães H., Marco-Sola S., Marijon P., McCartney A., McDaniel J., Mountcastle J., Nattestad M., Nurk S., Olson N.D., Popejoy A.B., Puiu D., Rautiainen M., Regier A.A., Rhie A., Sacco S., Sanders A.D., Schneider V.A., Schultz B.I., Shafin K., Smith M.W., Sofia H.J., Abou Tayoun A.N., Thibaud-Nissen F., Tricomi F.F., Wagner J., Walenz B., Wood J.M.D., Zimin A.V., Bourque G., Chaisson M.J.P., Flicek P., Phillippy A.M., Zook J.M., Eichler E.E., Haussler D., Wang T., Jarvis E.D., Miga K.H., Garrison E., Marschall T., Hall I.M., Li H. & Paten B. (2023) A draft human pangenome reference. Nature 617, 312-24. 10.1038/s41586-023-05896-x

Liao Y., Smyth G.K. & Shi W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research 47, e47-e. 10.1093/nar/gkz114

Liu R., Tearle R., Low W.Y., Chen T., Thomsen D., Smith T.P.L., Hiendleder S. & Williams J.L. (2021) Distinctive gene expression patterns and imprinting signatures revealed in reciprocal crosses between cattle sub-species. BMC Genomics 22. 10.1186/s12864-021-07667-2

Lloret-Villas A., Bhati M., Kadri N.K., Fries R. & Pausch H. (2021) Investigating the impact of reference assembly choice on genomic analyses in a cattle breed. BMC Genomics 22. 10.1186/s12864-021-07554-w

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 91, 2757-61. doi:10.1073/pnas.91.7.2757

Low W.Y., Tearle R., Liu R., Koren S., Rhie A., Bickhart D.M., Rosen B.D., Kronenberg Z.N., Kingan S.B. & Tseng E. (2020) Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. Nature Communications 11, 1-14.

Lutz P.E., Almeid D., Belzeaux R., Yalcin I. & Turecki G. (2018) Epigenetic regulation of the kappa opioid receptor gene by an insertion-deletion in the promoter region. European Neuropsychopharmacology 28, 334-40. 10.1016/j.euroneuro.2017.12.013

MacHugh D.E., Larson G. & Orlando L. (2017) Taming the Past: Ancient DNA and the Study of Animal Domestication. Annual Review of Animal Biosciences 5, 329-51. 10.1146/annurev-animal-022516-022747

MacPhillamy C., Alinejad-Rokny H., Pitchford W.S. & Low W.Y. (2022) Cross-species enhancer prediction using machine learning. Genomics 114, 110454. 10.1016/j.ygeno.2022.110454

Marçais G., Delcher A.L., Phillippy A.M., Coston R., Salzberg S.L. & Zimin A. (2018) MUMmer4: A fast and versatile genome alignment system. PLoS Computational Biology 14, e1005944. 10.1371/journal.pcbi.1005944

Marshall C.R., Noor A., Vincent J.B., Lionel A.C., Feuk L., Skaug J., Shago M., Moessner R., Pinto D., Ren Y., Thiruvahindrapduram B., Fiebig A., Schreiber S., Friedman J., Ketelaars C.E., Vos Y.J., Ficicioglu C., Kirkpatrick S., Nicolson R., Sloman L., Summers A., Gibbons C.A., Teebi A., Chitayat D., Weksberg R., Thompson A., Vardy C., Crosbie V., Luscombe S., Baatjes R., Zwaigenbaum L., Roberts W., Fernandez B., Szatmari P. & Scherer S.W. (2008) Structural variation of chromosomes in autism spectrum disorder. American Journal of Human Genetics 82, 477-88. 10.1016/j.ajhg.2007.12.009

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 110, E1398-E406. doi:10.1073/pnas.1303367110

Moore G.E., Ishida M., Demetriou C., Al-Olabi L., Leon L.J., Thomas A.C., Abu-Amero S., Frost J.M., Stafford J.L., Chaoqun Y., Duncan A.J., Baigel R., Brimioulle M., Iglesias-Platas I., Apostolidou S., Aggarwal R., Whittaker J.C., Syngelaki A., Nicolaides K.H., Regan L., Monk D. & Stanier P. (2015) The role and interaction of imprinted genes in human fetal growth. Philos Trans R Soc Lond B Biol Sci 370, 20140074. 10.1098/rstb.2014.0074

Moore L.D., Le T. & Fan G. (2013) DNA Methylation and Its Basic Function. Neuropsychopharmacology 38, 23-38. 10.1038/npp.2012.112

Morison I.M., Ramsay J.P. & Spencer H.G. (2005) A census of mammalian imprinting. Trends in Genetics 21, 457-65. 10.1016/j.tig.2005.06.008

Nattestad M. & Schatz M.C. (2016) Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics 32, 3021-3. 10.1093/bioinformatics/btw369

Okonechnikov K., Conesa A. & García-Alcalde F. (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32, 292-4. 10.1093/bioinformatics/btv566

Pan Z.X., Zhang J.L., Zhang J.B., Zhou B., Chen J., Jiang Z.H. & Liu H.L. (2012) Expression Profiles of the Insulin-like Growth Factor System Components in Liver Tissue during Embryonic and Postnatal Growth of Erhualian and Yorkshire Reciprocal Cross F-1 Pigs. Asian-Australasian Journal of Animal Sciences 25, 903-12. 10.5713/ajas.2011.11385

Park Y., Figueroa M.E., Rozek L.S. & Sartor M.A. (2014) MethylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics 30, 2414-22. 10.1093/bioinformatics/btu339

Paten B., Novak A.M., Eizenga J.M. & Garrison E. (2017) Genome graphs and the evolution of genome inference. Genome Research 27, 665-76. 10.1101/gr.214155.116

Piedrahita J.A. (2011) The Role of Imprinted Genes in Fetal Growth Abnormalities. Birth Defects Research Part a-Clinical and Molecular Teratology 91, 682-92. 10.1002/bdra.20795

Price A. & Gibas C. (2017) The quantitative impact of read mapping to non-native reference genomes in comparative RNA-Seq studies. PLOS ONE 12, e0180904. 10.1371/journal.pone.0180904

Ramsahoye B.H., Biniszkiewicz D., Lyko F., Clark V., Bird A.P. & Jaenisch R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proceedings of the National Academy of Sciences 97, 5237-42. doi:10.1073/pnas.97.10.5237

Ren Y., Tseng E., Smith T.P.L., Hiendleder S., Williams J.L. & Low W.Y. (2023) Long read isoform sequencing reveals hidden transcriptional complexity between cattle subspecies. BMC Genomics 24, 108. 10.1186/s12864-023-09212-9

Ritchie M.E., Phipson B., Wu D., Hu Y., Law C.W., Shi W. & Smyth G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, e47-e. 10.1093/nar/gkv007

Robinson M.D., McCarthy D.J. & Smyth G.K. (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139-40. 10.1093/bioinformatics/btp616

Rosen B.D., Bickhart D.M., Schnabel R.D., Koren S., Elsik C.G., Tseng E., Rowan T.N., Low W.Y., Zimin A., Couldrey C., Hall R., Li W., Rhie A., Ghurye J., McKay S.D., Thibaud-Nissen F., Hoffman J., Murdoch B.M., Snelling W.M., McDaneld T.G., Hammond J.A., Schwartz J.C., Nandolo W., Hagen D.E., Dreischer C., Schultheiss S.J., Schroeder S.G., Phillippy A.M., Cole J.B., Van

Tassell C.P., Liu G., Smith T.P.L. & Medrano J.F. (2020) De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9, giaa021-giaa. 10.1093/gigascience/giaa021

Rothammer S., Capitan A., Mullaart E., Seichter D., Russ I. & Medugorac I. (2014) The 80-kb DNA duplication on BTA1 is the only remaining candidate mutation for the polled phenotype of Friesian origin. Genetics Selection Evolution 46, 44. 10.1186/1297-9686-46-44

Salavati M., Bush S.J., Palma-Vera S., McCulloch M.E.B., Hume D.A. & Clark E.L. (2019) Elimination of Reference Mapping Bias Reveals Robust Immune Related Allele-Specific Expression in Crossbred Sheep. Frontiers in Genetics 10. 10.3389/fgene.2019.00863

Shi W., Krella A., Orth A., Yu Y. & Fundele R. (2005) Widespread disruption of genomic imprinting in adult interspecies mouse (Mus) hybrids. Genesis 43, 100-8. 10.1002/gene.20161

Shumate A. & Salzberg S.L. (2021) Liftoff: accurate mapping of gene annotations. Bioinformatics 37, 1639-43. 10.1093/bioinformatics/btaa1016

Slabaugh E., Desai J.S., Sartor R.C., Lawas L.M.F., Jagadish S.V.K. & Doherty C.J. (2019) Analysis of differential gene expression and alternative splicing is significantly influenced by choice of reference genome. RNA 25, 669-84. 10.1261/rna.070227.118

Slieker R.C., Roost M.S., van Iperen L., Suchiman H.E., Tobi E.W., Carlotti F., de Koning E.J., Slagboom P.E., Heijmans B.T. & Chuva de Sousa Lopes S.M. (2015) DNA Methylation Landscapes of Human Fetal Development. PLoS Genetics 11, e1005583. 10.1371/journal.pgen.1005583

Smith J., Sen S., Weeks R.J., Eccles M.R. & Chatterjee A. (2020) Promoter DNA Hypermethylation and Paradoxical Gene Activation. Trends in Cancer 6, 392-406. https://doi.org/10.1016/j.trecan.2020.02.007

Smith T.P.L., Bickhart D.M., Boichard D., Chamberlain A.J., Djikeng A., Jiang Y., Low W.Y., Pausch H., Demyda-Peyrás S., Prendergast J., Schnabel R.D., Rosen B.D. & Bovine Pangenome C. (2023) The Bovine Pangenome Consortium: democratizing production and accessibility of genome assemblies for global cattle breeds and other bovine species. Genome Biology 24, 139. 10.1186/s13059-023-02975-0

Spainhour J.C.G., Lim H.S., Yi S.V. & Qiu P. (2019) Correlation Patterns Between DNA Methylation and Gene Expression in The Cancer Genome Atlas. Cancer Informatics 18. 10.1177/1176935119828776

Szabo P.E., Tang S.H.E., Rentsendorj A., Pfeifer G.P. & Mann J.R. (2000) Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function. Current Biology 10, 607-10. 10.1016/s0960-9822(00)00489-9

Thomson K., Game J., Karouta C., Morgan I.G. & Ashby R. (2022) Correlation between small-scale methylation changes and gene expression during the development of myopia. The FASEB Journal 36, e22129. https://doi.org/10.1096/fj.202101487R

Tsuda N., Kumadaki S., Higashi C., Ozawa M., Shinozaki M., Kato Y., Hoshida K., Kikuchi S., Nakano Y., Ogawa Y. & Furusako S. (2014) Intestine-Targeted DGAT1 Inhibition Improves Obesity and Insulin Resistance without Skin Aberrations in Mice. PLOS ONE 9, e112027. 10.1371/journal.pone.0112027

Valiente-Mullor C., Beamud B., Ansari I., Francés-Cuesta C., García-González N., Mejía L., Ruiz-Hueso P. & González-Candelas F. (2021) One is not enough:

On the effects of reference genome for the mapping and subsequent analyses of short-reads. PLoS Computational Biology 17, e1008678. 10.1371/journal.pcbi.1008678

Van Laere A.-S., Nguyen M., Braunschweig M., Nezer C., Collette C., Moreau L., Archibald A.L., Haley C.S., Buys N., Tally M., Andersson G., Georges M. & Andersson L. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425, 832-6. 10.1038/nature02064

Vaughn R.N., Kochan K.J., Torres A.K., Du M., Riley D.G., Gill C.A., Herring A.D., Sanders J.O. & Riggs P.K. (2022) Skeletal Muscle Expression of Actinin-3 (ACTN3) in Relation to Feed Efficiency Phenotype of F-2 Bos indicus-Bos taurus Steers. Frontiers in Genetics 13. 10.3389/fgene.2022.796038

Wang M., Hancock T.P., MacLeod I.M., Pryce J.E., Cocks B.G. & Hayes B.J. (2017) Putative enhancer sites in the bovine genome are enriched with variants affecting complex traits. Genetics Selection Evolution 49, 56. 10.1186/s12711-017-0331-4

Wang T., Antonacci-Fulton L., Howe K., Lawson H.A., Lucas J.K., Phillippy A.M., Popejoy A.B., Asri M., Carson C., Chaisson M.J.P., Chang X., Cook-Deegan R., Felsenfeld A.L., Fulton R.S., Garrison E.P., Garrison N.A., Graves-Lindsay T.A., Ji H., Kenny E.E., Koenig B.A., Li D., Marschall T., McMichael J.F., Novak A.M., Purushotham D., Schneider V.A., Schultz B.I., Smith M.W., Sofia H.J., Weissman T., Flicek P., Li H., Miga K.H., Paten B., Jarvis E.D., Hall I.M., Eichler E.E., Haussler D. & the Human Pangenome Reference C. (2022) The Human Pangenome Project: a global resource to map genomic diversity. Nature 604, 437-46. 10.1038/s41586-022-04601-8

Weiss L.A., Shen Y., Korn J.M., Arking D.E., Miller D.T., Fossdal R., Saemundsen E., Stefansson H., Ferreira M.A., Green T., Platt O.S., Ruderfer D.M., Walsh C.A., Altshuler D., Chakravarti A., Tanzi R.E., Stefansson K., Santangelo S.L., Gusella J.F., Sklar P., Wu B.L. & Daly M.J. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358, 667-75. 10.1056/NEJMoa075974

Woodhouse M.R., Cannon E.K., Portwood J.L., Harper L.C., Gardiner J.M., Schaeffer M.L. & Andorf C.M. (2021) A pan-genomic approach to genome databases using maize as a model system. BMC Plant Biology 21, 385. 10.1186/s12870-021-03173-5

Wu P.-Y., Phan J.H. & Wang M.D. (2013) Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinformatics 14, S8. 10.1186/1471-2105-14-S11-S8

Wulfridge P., Langmead B., Feinberg A.P. & Hansen K.D. (2019) Analyzing whole genome bisulfite sequencing data from highly divergent genotypes. Nucleic Acids Research 47, e117-e. 10.1093/nar/gkz674

Yang J., Horton J.R., Akdemir K.C., Li J., Huang Y., Kumar J., Blumenthal R.M., Zhang X. & Cheng X. (2021) Preferential CEBP binding to T:G mismatches and increased C-to-T human somatic mutations. Nucleic Acids Research 49, 5084-94. 10.1093/nar/gkab276

Yang Y.W., Hu J.F., Ulaner G.A., Li T., Yao X.M., Vu T.H. & Hoffman A.R. (2003) Epigenetic regulation of Igf2/H19 imprinting at CTCF insulator binding sites. Journal of Cellular Biochemistry 90, 1038-55. 10.1002/jcb.10684

Yuen R.K.C., Jiang R., Penaherrera M.S., McFadden D.E. & Robinson W.P. (2011) Genome-wide mapping of imprinted differentially methylated regions by DNA methylation profiling of human placentas from triploidies. Epigenetics & Chromatin 4. 10.1186/1756-8935-4-10

Żemojtel T., Kiełbasa S.M., Arndt P.F., Behrens S., Bourque G. & Vingron M. (2011) CpG Deamination Creates Transcription Factor–Binding Sites with High Efficiency. Genome Biology and Evolution 3, 1304-11. 10.1093/gbe/evr107

Zhang X.D., Yan J.W., Yan G.R., Sun X.Y., Ji J., Li Y.M., Hu Y.H. & Wang H.Y. (2010) Pharmacological inhibition of diacylglycerol acyltransferase 1 reduces body weight gain, hyperlipidemia, and hepatic steatosis in db/db mice. Acta Pharmacol Sin 31, 1470-7. 10.1038/aps.2010.104

Zhang Y.Q., Yang L.X., Kucherlapati M., Hadjipanayis A., Pantazi A., Bristow C.A., Lee E.A., Mahadeshwar H.S., Tang J.B., Zhang J.H., Seth S., Lee S., Ren X.J., Song X.Z., Sun H.D., Seidman J., Luquette L.J., Xi R.B., Chin L., Protopopov A., Park P.J., Kucherlapati R. & Creighton C.J. (2019) Global impact of somatic structural variation on the DNA methylome of human cancers. Genome Biology 20. 10.1186/s13059-019-1818-9

Zhou Y., Liu S., Hu Y., Fang L., Gao Y., Xia H., Schroeder S.G., Rosen B.D., Connor E.E., Li C.-j., Baldwin R.L., Cole J.B., Van Tassell C.P., Yang L., Ma L. & Liu G.E. (2020) Comparative whole genome DNA methylation profiling across cattle tissues reveals global and tissue-specific methylation patterns. BMC Biology 18, 85. 10.1186/s12915-020-00793-5

Zhu Y., Zhou Z., Huang T., Zhang Z., Li W., Ling Z., Jiang T., Yang J., Yang S., Xiao Y., Charlier C., Georges M., Yang B. & Huang L. (2022) Mapping and analysis of a spatiotemporal H3K27ac and gene expression spectrum in pigs. Sci China Life Sci 65, 1517-34. 10.1007/s11427-021-2034-5

Ziller M.J., Müller F., Liao J., Zhang Y., Gu H., Bock C., Boyle P., Epstein C.B., Bernstein B.E., Lengauer T., Gnirke A. & Meissner A. (2011) Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. PLoS Genetics 7, e1002389. 10.1371/journal.pgen.1002389

Tables

Table 1. Mapping statistics of Angus and Brahman reference genomes.

	Angus	Brahman
Mapped reads*	1,455,481,398	1,457,794,807
Duplication rate (%)*	10	14
CpGs with ≥ 10X coverage in all samples	22,116,287	21,962,589
CpG coverage*	30	30

* Mean of all samples.

Table 2. Number of CpGs in the Angus and Brahman reference genomes.

	Angus	Brahman
Total CpGs[A]	25,712,300	25,799,151
CpGs aligned to other reference[B]	25,209,966	25,228,509
CpGs shared in other genome[C]	18,813,726	18,781,688
CpGs affected by SNP[D]	993,318	1,003,167
Unresolved CpGs[E]	5,402,922	5,443,654

[A] Total number of CpGs present within the genome.

[B] Number of CpGs that could be aligned from one genome to the other using Minimap2.

[C] Number of CpGs in B that were CpGs in both species.

[D] Number of CpGs in B that were a CpG in one species but are no longer CpGs in the other.

[E] Number of CpGs in B that could not be confidently assigned as either shared or a SNP.

Table 3. Significant imprinted DEGs and their overlap with DMRs when using the Brahman reference genome.

Gene ID	Gene name	Protein name	Increased expression in Brahman*	Number of hypo-DMRs in Brahman	Number of hyper-DMRs in Brahman
Breed comparison					
ENSBIXG00005001873	S*moc1*	SPARC-related modular calcium-binding 1	Yes	16	1
ENSBIXG00005012203	*Nnat*	Neuronatin	No	37	1
ENSBIXG00005014559	*Gdap1l1*	Ganglioside-induced differentiation-associated protein 1 like 1	No	30	2
ENSBIXG00005009822	*Dgat1*	Diacylglycerol O-acyltransferase 1	Yes	3	2
ENSBIXG00005024991	*Slc22a18*	Solute-carrier family 22 member 18	Yes	2	3
ENSBIXG00005004279	*Kcnq1*	Potassium voltage-gated channel subfamily Q member 1	Yes	9	5
ENSBIXG00005007141	*Ppp1r9a*	Protein phosphatase 1 regulatory subunit 9A	Yes	11	5
Dam of origin comparison					
ENSBIXG00005019306	*Zc3h12c*	Zinc finger CCCH-type containing 12C	Yes	5	1
ENSBIXG00005007141	*Ppp1r9a*	Protein phosphatase 1 regulatory subunit 9A	Yes	3	2
Sire of origin comparison					
ENSBIXG00005007073	*Dscam*	DS cell adhesion molecule	No	38	5
ENSBIXG00005021735	*Htr2a*	5-hydroxytryptamine receptor 2A	Yes	18	2

ENSBIXG00005012203	*Nnat*	Neuronatin	No	8	0
ENSBIXG00005009822	*Dgat1*	Diacylglycerol O-acyltransferase 1	Yes	0	1
ENSBIXG00005016997	*Axl*	AXL receptor tyrosine kinase	Yes	9	1
ENSBIXG00005025694	*Ndn*	Necdin MAGE family member	No	0	1
ENSBIXG00005024991	*Slc22a18*	Solute-carrier family 22 member 18	Yes	0	0
ENSBIXG00005013434	*Tfpi2*	Tissue factor pathway inhibitor 2	No	25	0

* Increased expression in Brahman denotes genes that were significantly more highly expressed in Brahman than in Angus. “No” denotes that gene was significantly more highly expressed in Angus.

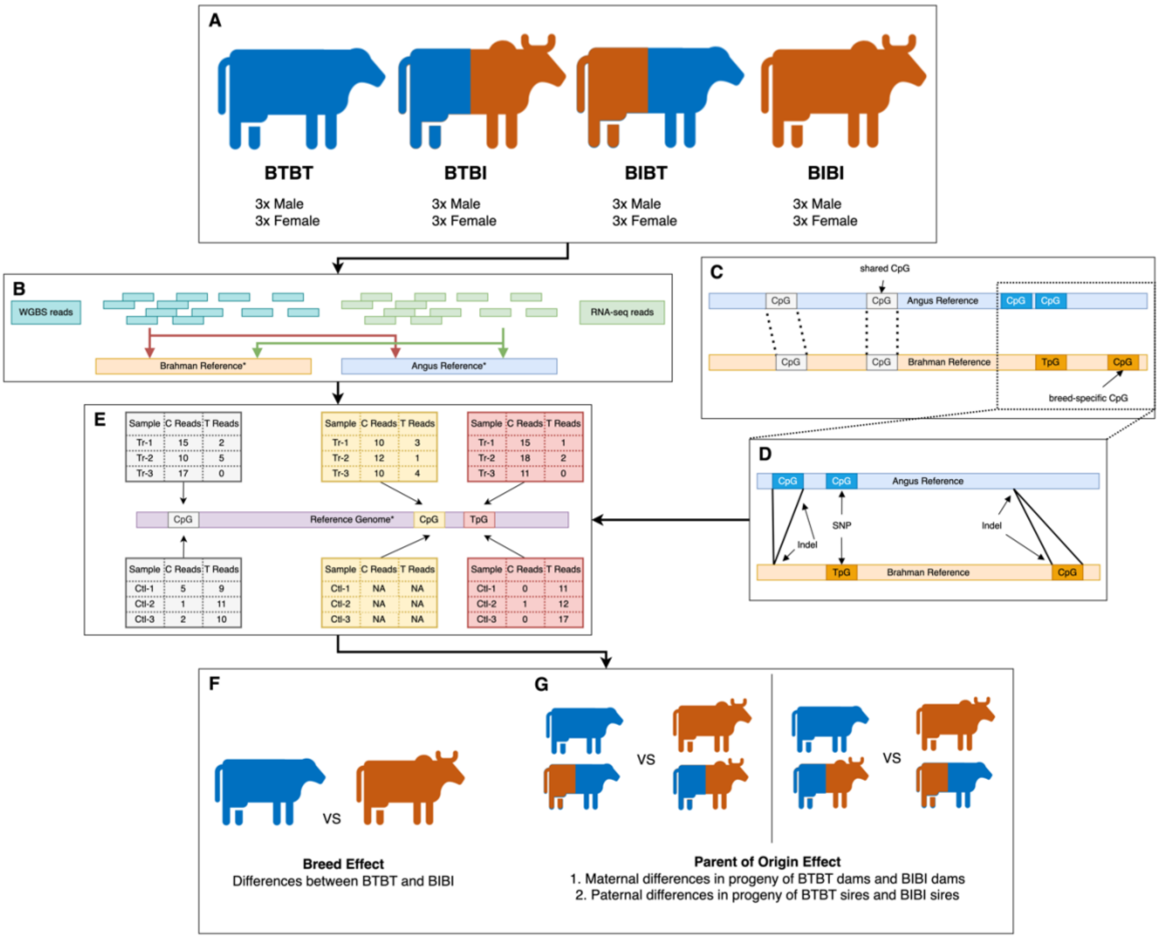
Table 4. Significant imprinted DEGs and their overlap with DMRs when using the Angus reference genome.

Gene ID	Gene name	Protein name	Increased expression in Angus*	Number of hypo-DMRs in Angus	Number of hyper-DMRs in Angus
ENSBIXG00000024138	*Esr2*	Estrogen receptor 2	No	4	8
ENSBIXG00000021864	*Nnat*	Neuronatin	Yes	2	32
ENSBIXG00000015750	*Zfp90*	Zinc finger protein 90	Yes	2	1
ENSBIXG00000012277	*Dlgap2*	DLG-associated protein 2	No	11	23
ENSBIXG00000005197	*Ppp1r9a*	Protein phosphatase 1 regulatory subunit 9A	No	9	15
Dam of origin comparison					
ENSBIXG00000011151	*Zc3h12c*	Zinc finger CCCH-type containing 12C	No	0	4
ENSBIXG00000005197	*Ppp1r9a*	Protein phosphatase 1 regulatory subunit 9A	No	6	3
Sire of origin comparison					
ENSBIXG00000027129	*Dscam*	DS cell adhesion molecule	Yes	5	35
ENSBIXG00000024138	*Esr2*	Estrogen receptor 2	No	3	1
ENSBIXG00000008539	*Htr2a*	5-hydroxytryptamine receptor 2A	No	1	17
ENSBIXG00000021864	*Nnat*	Neuronatin	Yes	1	11
ENSBIXG00000012321	*Dgat1*	Diacylglycerol O-acyltransferase 1	No	0	0
ENSBIXG00000010895	*Sdhd*	Succinate dehydrogenase complex subunit D	Yes	0	5
ENSBIXG00000016809	*Axl*	AXL receptor tyrosine kinase	No	0	8
ENSBIXG00000015087	*Mkrn3*	Makorin ring finger protein 3	Yes	1	1
ENSBIXG00000015080	*Ndn*	Necdin MAGE family member	Yes	1	0
ENSBIXG00000012277	*Dlgap2*	DLG-associated protein 2	No	1	11

ENSBIXG00 000028529	*Slc22a18*	Solute-carrier family 22 member 18	No	0	2

* Increased expression in Angus denotes genes that were significantly more highly expressed in Angus than in Brahman. “No” denotes that gene was significantly more highly expressed in Brahman.

Figures



Sample	C Reads	T Reads
Tr-1	15	2
Tr-2	10	5
Tr-3	17	0

Sample	C Reads	T Reads
Tr-1	10	3
Tr-2	12	1
Tr-3	10	4

Sample	C Reads	T Reads
Tr-1	15	1
Tr-2	18	2
Tr-3	11	0

Sample	C Reads	T Reads
Ctl-1	5	9
Ctl-2	1	11
Ctl-3	2	10

Sample	C Reads	T Reads
Ctl-1	NA	NA
Ctl-2	NA	NA
Ctl-3	NA	NA

Sample	C Reads	T Reads
Ctl-1	0	11
Ctl-2	1	12
Ctl-3	0	17

Figure 1. Overview of methods. A.) Representation of the four genetic groups used in this study. The blue cow represents pure Angus individuals (BTBT). The blue then orange cow represents individuals with an Angus sire and Brahman dam (BTBI). The orange then blue cow represents individuals with a Brahman sire and Angus dam (BIBT). The orange cow represents pure Brahman individuals (BIBI). **B.)** Process of mapping WGBS reads (light green-blue), and RNA-seq reads (green) to both the Brahman and Angus reference genomes. **C.)** Simple representation of shared and breed-specific CpG sites between Brahman and Angus reference genomes. **D.)** Breed-specific CpGs arise from a single nucleotide polymorphism between Brahman and Angus, such as spontaneous deamination of the C to a T. Structural variants, such as

indels between the two genomes, can introduce or remove CpGs in one genome relative to the other. **E.)** Simple representation of how differential methylation can be influenced by breed-specific CpGs. The grey boxes demonstrate how a differentially methylated cytosine is identified when both breeds share that site. Essentially, one compares the number of Cs and Ts in group 1 against the number of Cs and Ts in group 2. If one group reports significantly more Cs than the other, it is considered differentially methylated. The yellow boxes represent a breed-specific CpG where only samples from one group have that CpG, so differential methylation cannot be determined. The red boxes represent a situation where the CpG is present in one subspecies, but spontaneous deamination has mutated the CpG site into a TpG site in the other subspecies. In this case, differential methylation can be calculated. However, it will be erroneous as only one group has a true CpG at that site. **F.)** Graphical representation of how breed differences were determined. We compared methylation and gene expression between BTBT and BIBI samples. **G.)** Graphical representation of how we determined parent-of-origin effects (POEs). Maternal POEs were determined by comparing BTBT and BIBT against BIBI and BTBI. Paternal POEs were determined by comparing BTBT and BTBI against BIBI and BIBT.

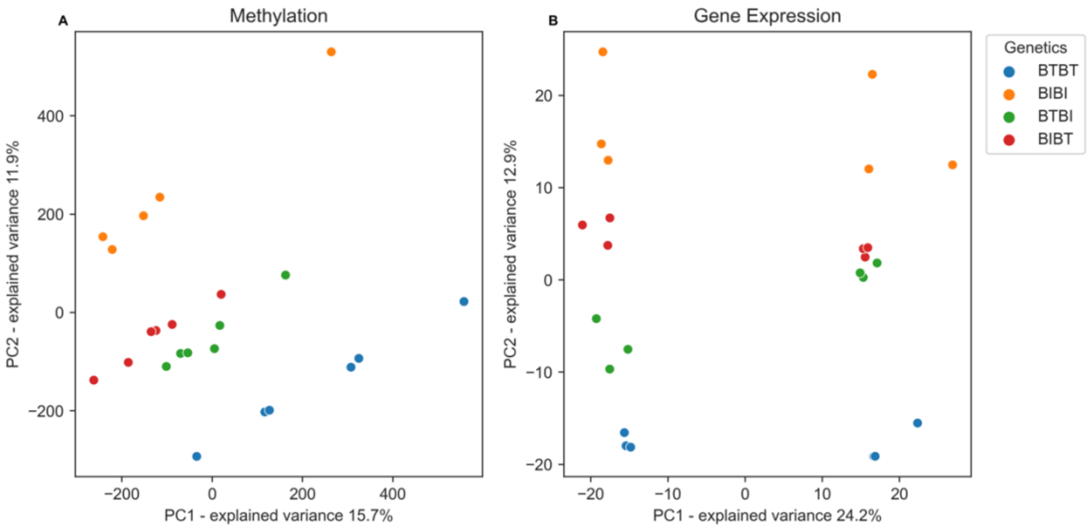


Figure 2. A.) PCA plot showing separation of genetic groups by methylation. Blue represents BTBT, orange represents BIBI, green represents BTBI and red represents BIBT. The X axis is principal component 1, and the Y axis is principal component 2. **B.)** PCA plot showing separation of genetic groups by gene expression data; colours are same as **A**. The X axis is the first dimension of the logFC, and the Y axis is the second dimension of the logFC.

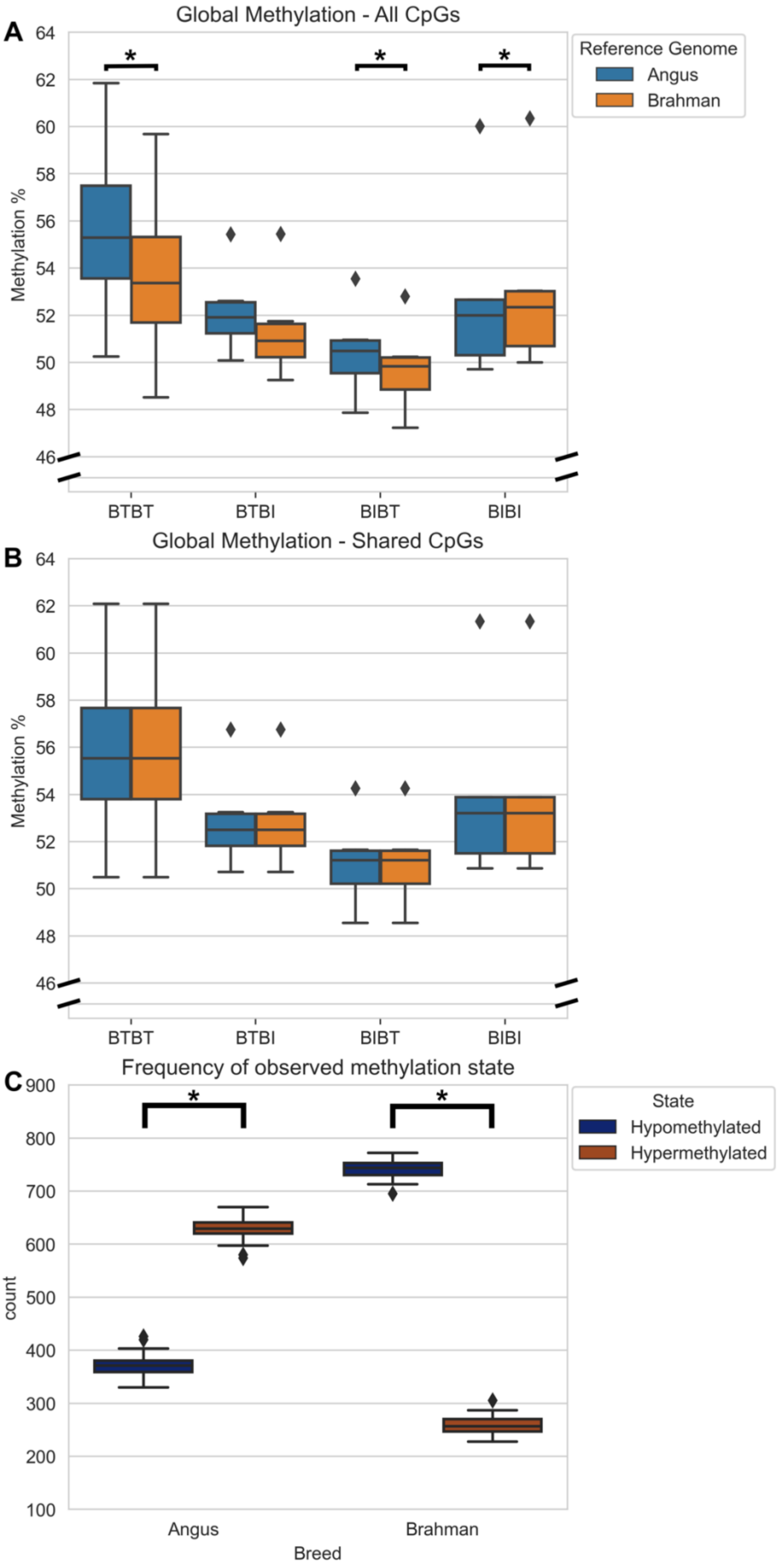

A
Global Methylation - All CpGs
Reference Genome
Angus
Brahman
Methylation %
64
62
60
58
56
54
52
50
48
46
*
*
*
BTBT
BTBI
BIBT
BIBI
B
Global Methylation - Shared CpGs
Methylation %
64
62
60
58
56
54
52
50
48
46
BTBT
BTBI
BIBT
BIBI
C
Frequency of observed methylation state
State
Hypomethylated
Hypermethylated
count
900
800
700
600
500
400
300
200
100
*
*
Angus
Brahman
Breed

Figure 3. A.) Boxplot showing mean global CpG methylation for samples belonging to the four genetic groups. When Brahman and Angus are mapped to their respective genomes, they tend to be more methylated than when mapped to the incorrect reference. The hybrids (BTBI and BIBT) tend toward hypermethylation when mapped to the Angus reference though this is only significantly different in BIBT. * denotes p-value < 0.05, Wilcox test. **B.)** Boxplot showing mean global CpG methylation for samples belonging to each of the four genetic groups mapped to each reference genome; however, this time, only the shared CpGs were considered. Here, there is no significant difference in methylation within genetic groups regardless of which reference genome is used. **C.)** Boxplot showing the mean frequencies methylation states observed in Angus and Brahman after 100 permutations. Dark blue represents hypomethylated CpG sites (methylation ≤ 35%). Dark orange bars represent hypermethylated CpG sites (methylation ≥ 65%). * denotes p-value < 0.05, Mann-Whitney U-test. The X-axis denotes the breed, either Angus or Brahman. The Y-axis represents the count, i.e., the number of sites that fell into the hypo- or hypermethylated categories.

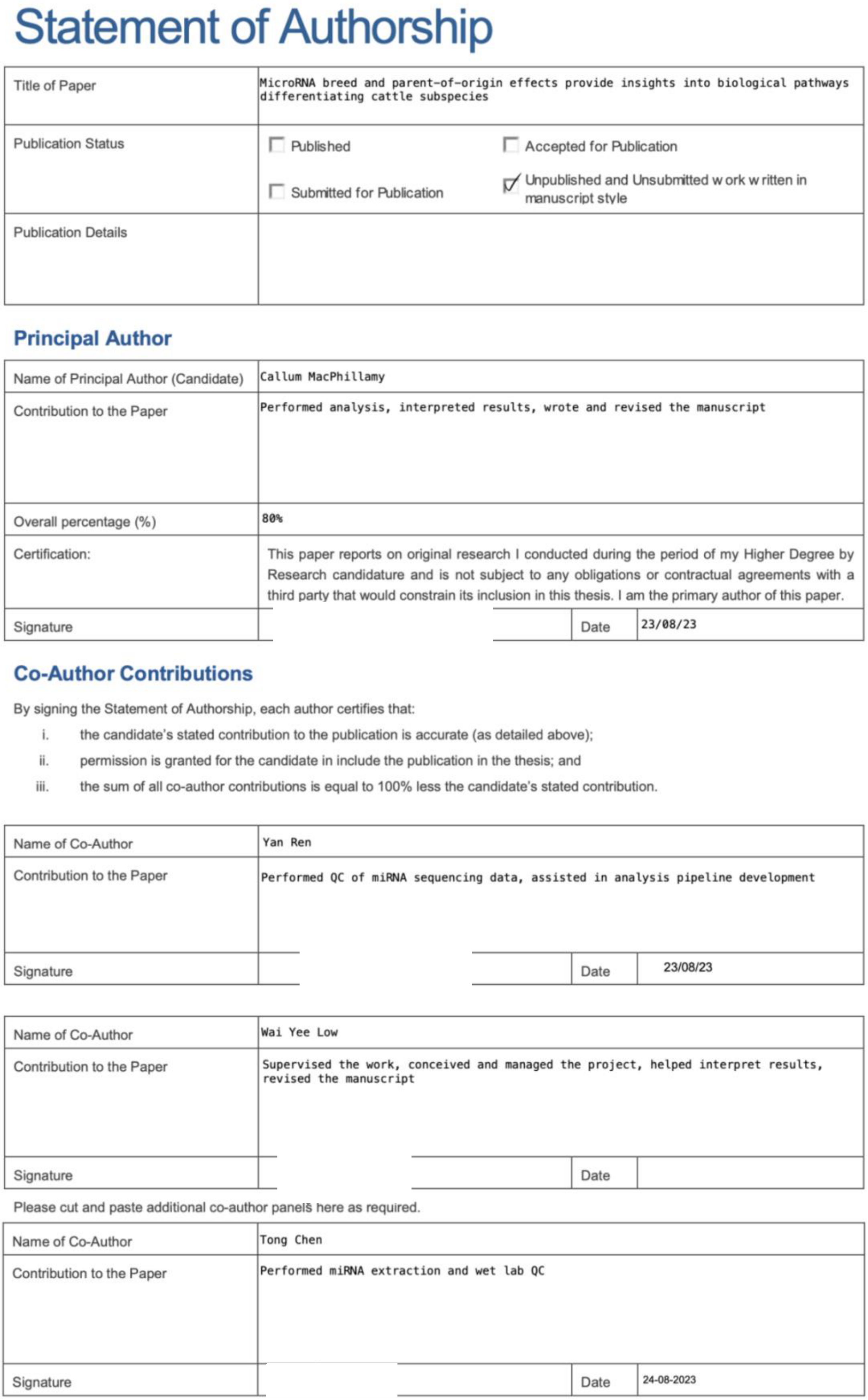
Supplementary materials for Chapter 4 can be in Appendix II.

Chapter 5: MicroRNA breed and parent-of-origin effects provide insights into biological pathways differentiating cattle subspecies

Contextual Statement

Chapter 5 further investigates the possible causes of gene expression differences between Brahman and Angus cattle by comparing microRNA expression differences between the two breeds. MicroRNAs play an essential role in regulating diverse biological processes, and to date, no studies have examined these differences between Brahman and Angus using an essential metabolic organ like the liver. This study built on Chapter 4 by investigating differential microRNA expression between Brahman and Angus in a breed- and parent-of-origin-specific manner. MicroRNA expression was correlated with mRNA expression from the same samples, and standard tools for microRNA and mRNA expression analysis were used alongside KEGG pathway enrichment to identify differentially expressed microRNAs, mRNAs and pathways that may be contributing to phenotypic differences observed between these two breeds.

Statement of authorship



Statement of Authorship

Title of Paper	MicroRNA breed and parent-of-origin effects provide insights into biological pathways differentiating cattle subspecies	
Publication Status	□ Published □ Submitted for Publication	□ Accepted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details		

Principal Author

Name of Principal Author (Candidate)	Callum MacPhillamy		
Contribution to the Paper	Performed analysis, interpreted results, wrote and revised the manuscript		
Overall percentage (%)	80%		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	23/08/23

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

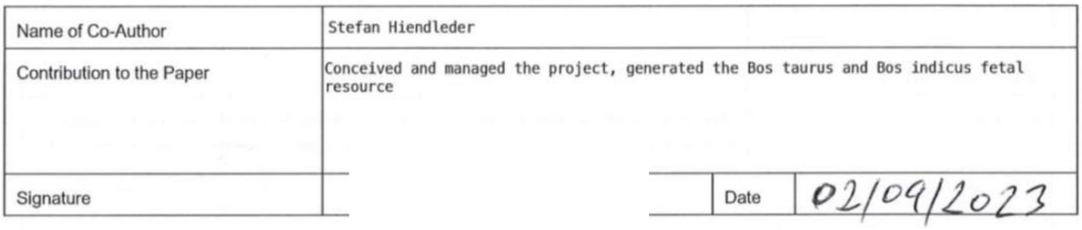
iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Yan Ren		
Contribution to the Paper	Performed QC of miRNA sequencing data, assisted in analysis pipeline development		
Signature		Date	23/08/23

Name of Co-Author	Wai Yee Low		
Contribution to the Paper	Supervised the work, conceived and managed the project, helped interpret results, revised the manuscript		
Signature		Date	

Please cut and paste additional co-author panels here as required.

Name of Co-Author	Tong Chen		
Contribution to the Paper	Performed miRNA extraction and wet lab QC		
Signature		Date	24-08-2023



Name of Co-Author	Stefan Hiendleder		
Contribution to the Paper	Conceived and managed the project, generated the Bos taurus and Bos indicus fetal resource		
Signature		Date	02/09/2023

MicroRNA breed and parent-of-origin effects provide insights into biological pathways differentiating cattle subspecies

Authors

Callum MacPhillamy[1], Yan Ren[1], Tong Chen[1], Stefan Hiendleder[1,2], Wai Yee Low[1]

[1]The Davies Research Centre, School of Animal and Veterinary Sciences, University of Adelaide, Roseworthy, SA 5371, Australia

[2]Robinson Research Institute, The University of Adelaide, North Adelaide, SA 5006, Australia

Abstract

MicroRNAs (miRNAs) are small non-coding RNA species that play a crucial role in regulating gene expression during key developmental processes, such as fetal development. Brahman (*Bos taurus indicus*) and Angus (*Bos taurus taurus*) are two economically important cattle breeds with contrasting phenotypes. We analysed miRNA expression data from fetal liver of pure and reciprocally crossed samples of Angus and Brahman to investigate breed and parent-of-origin differences between these two phenotypically diverse breeds. There were 14 differentially expressed miRNAs (DEMs) between the two pure breeds. Correlation of gene expression modules and miRNAs by breed and parent-of-origin differences revealed an enrichment of genes associated with breed traits like heat tolerance in Brahman and fat gain in Angus. We demonstrate that genes predicted to be targeted by DEMs were more likely to be differentially expressed than non-targets (p-value <0.05). We identified several miRNAs (bta-miR-187, bta-miR-216b, bta-miR-2284c, bta-miR-2285c, bta-miR-2285cp, bta-miR-2419-3p, bta-miR-2419-5p, bta-miR-11984) that showed similar correlation patterns as bta-miR-2355-3p which was a miRNA that may have a role in heat tolerance in Brahman and several Angus-specific miRNAs (bta-miR-2313-5p, bta-miR-490, bta-miR-2316, bta-miR-11990) that that may be involved in fat gain in Angus. Furthermore, we showed that the differentially expressed fetal miRNAs we identified tend to target Rap1, MAPK and Ras signalling pathways. This work sheds light on miRNA expression patterns that contribute to gene expression differences that drive phenotypic changes in indicine and taurine cattle.

Introduction

The two main lineages of modern cattle breeds are generally accepted to have been derived from two separate domestication events of the wild auroch (*Bos primigenius*) (McTavish *et al.* 2013). The first domestication event occurred in the Fertile Crescent around 10,000 years ago and gave rise to *Bos taurus taurus* from the wild auroch, *B. p. primigenius* (Bruford *et al.* 2003; Ajmone-Marsan *et al.* 2010; MacHugh *et al.* 2017). A second domestication event occurred in the Indus Valley, ~1,500 years later, from *B. p. nomadicus*, which separated from the *B. p. primigenius* around 250-330,000 years ago (Loftus *et al.* 1994) and gave rise to *Bos taurus indicus*. The subspecies are referred to here as taurine and indicine cattle, respectively (McTavish *et al.* 2013). In this work, the Angus breed represents taurine cattle and Brahman is representative of indicine cattle. Angus and Brahman have contrasting phenotypes, e.g., Angus have been bred for meat production and growth traits (Elzo *et al.* 2012), and hence have shorter gestation length and lower calving difficulty (Casas *et al.* 2011). In contrast, Brahman cattle have superior heat and disease tolerance traits that enable them to adapt to tropical environments but mature slowly (Dikmen *et al.* 2018; Goszczynski *et al.* 2018).

The genetic regulatory changes that drive the phenotypic differences between taurine and indicine cattle have been of substantial interest from both scientific and economic standpoints due to the value of understanding and combining desirable traits from both subspecies (Li *et al.* 2023). Angus and Brahman differ by ~1% genetically (Low *et al.* 2020) but can be mated to produce fertile hybrids, which has enabled researchers to investigate factors driving their phenotypic differences (Hiendleder *et al.* 2008; Akanno *et al.* 2018; Gobena *et al.* 2018; Andrade *et al.* 2022). One of the main factors behind their phenotypic differences could be changes in their microRNA (miRNA) expression, but this has never been explored before in economically important breeds.

MiRNAs are small non-coding RNAs, around 22bp long, that primarily inhibit gene expression at the post-transcriptional level (Kim *et al.* 2008; Ha & Kim 2014). They are essential regulators of gene expression and have been implicated in a wide range of biological processes, including growth and developmental processes (Awamleh *et al.* 2019; Morales-Roselló *et al.* 2022), differentiation (Galagali & Kim 2020), metabolism (Rottiers & Näär 2012), and responding to environmental stimuli (Vrijens *et al.* 2015; Liu *et al.* 2020). In general, miRNAs are highly conserved across species (Macfarlane & Murphy 2010). Interestingly there are species- and tissue-specific miRNA expression patterns (Jopling 2012; Sun *et al.* 2014). Regulatory changes by miRNA can lead to changes in gene expression, which contributes to the emergence of new traits and phenotypic diversity (Yan *et al.* 2013; Li *et al.* 2020; Hao *et al.* 2023).

In cattle, several miRNAs have been implicated in adipogenesis and myogenesis. Bta-miR-424 has been shown to promote adipogenesis by binding to and upregulating serine/threonine kinase 11 (*stk11*) (Wang *et al.* 2020). Conversely, bta-miR-148-3p inhibition promotes muscle differentiation by downregulating the Krueppel-like factor 6 (*klf6*) gene (Song *et al.* 2019). Similarly, bta-miR-206, bta-miR-1, bta-miR-133 and two novel miRNAs were highly expressed in muscle-related tissues and organs, suggesting a possible role in myogenesis (Sun *et al.* 2013).

Several studies have investigated the differential expression of miRNAs among different cattle breeds in an effort to elucidate possible regulatory mechanisms of certain traits. A comparison of miRNA expression in Sahiwal (*Bos indicus*) and Frieswal (*Bos indicus* x *Bos taurus*) cattle in response to heat stress found bta-miR-150, bta-miR-16a and bta-miR-181b were upregulated in Frieswal cattle (Deb & Sengar 2021). These three miRNAs were negatively correlated with heat shock protein 70.1 (*HSP70.1*) expression (Deb & Sengar 2021), suggesting a possible cause of the difference in heat tolerance between the two breeds. Similarly, a comparison of miRNA expression in the mammary glands of two dairy breeds revealed 17 differentially expressed miRNAs (DEMs) that were predicted to target genes that are likely important for milk synthesis, such as those involved in glucose and lipid metabolism (Billa *et al.* 2019).

Studies in cattle are limited, but recent work using mouse embryonic stem cells has revealed that miRNAs originating from maternally inherited regions of the genome antagonise paternally driven gene programmes by targeting paternally expressed transcripts (Whipple *et al.* 2020). This finding suggests that miRNAs may have a role in the parental genome conflict (Haig 2014), where paternally expressed genes like *Igf2* promote the uptake of nutrients in the fetus (DeChiara *et al.* 1991), while maternally expressed genes, like *Igf2r*, restrict growth (Barlow *et al.* 1991), thus conserving maternal resources. Given that Brahman and Angus have evolved under different selection pressures, their parent-of-origin-specific miRNA expression is likely different, though this has not been explored.

We have previously investigated the gene expression (Liu *et al.* 2021) and methylation differences (Chapter 4) between purebred and reciprocal crosses of Angus and Brahman. Here we report differential miRNA expression by breed and parent-of-origin effects in Angus and Brahman cattle and explore their correlation with mRNA expression and potential biological pathways targeted by these miRNAs. This study aims to shed light on miRNA expression differences that may contribute to phenotypic differences between these two economically valuable cattle breeds.

Materials and methods

Study Animals and Sample Collection

Liver samples of concepti used in the present study were the same as those described in our previous work (Liu *et al.* 2021). All animal experiments and procedures described in this study complied with Australian guidelines, approved by the University of Adelaide's Animal Ethics Committee and followed the ARRIVE Guidelines (https://arriveguidelines.org/) (Approval No. S-094-2005). The samples were created as described in Liu *et al.* (2021). Briefly, the parents were purebred Angus (*Bos taurus taurus*) and purebred Brahman (*Bos taurus indicus*), denoted as BT and BI, respectively. Heifers and fetuses were ethically sacrificed at day 153 of gestation, then snap-frozen in liquid nitrogen and stored at -80°C until further use. The liver samples represented three female and three male individuals from each of the four genetic crosses: BT x BT, BT x BI, BI x BT, and BI x BI. The four genetic groups of the offspring were denoted by the breed of their parents, with the paternal breed being first. For example, BIBT represents a fetus whose sire was Brahman (BI) and whose dam was Angus (BT). The groups were BTBT, BTBI, BIBT and BIBI.

DNA extraction and sequencing

MicroRNA was extracted from frozen fetal liver tissue samples using the Bioo Scientific® NEXTflex™ Small RNA-Seq kit v3 according to the manufacturer's recommendations. All samples were sequenced in a single batch at Australian Cancer Research Foundation, Adelaide, Australia using an Illumina® NextSeq 500. Individual sample names and their corresponding genetic group are outlined in Supplementary Table 1.

Sequence alignment

MicroRNA sequences were aligned using a custom Nextflow pipeline (Di Tommaso *et al.* 2017). Sequencing reads were first checked for quality using FastQC (v. 0.12.1) (Andrews 2010). Adapters from the 5' and 3' ends of the reads were then trimmed using CutAdapt (v. 4.3) (Martin 2011). We then filtered reads to retain those within the 17-28bp range with a mean sequence quality of 25 using Prinseq (v. 0.20.4) (Schmieder & Edwards 2011). Reads passing the filtering were reassessed with FastQC. Reads were then filtered for various bovine small RNA species that included ribosomal RNA (rRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), and small nucleolar RNA (snoRNA). These small RNAs were downloaded from the Rfam database (Kalvari *et al.* 2018; Kalvari *et al.* 2020) via RNA Central and accessed on the 28th of April, 2023. Next, we filtered reads against non-coding RNA (ncRNA) and coding DNA (cDNA) from Ensembl release 109 for ARS-UCD1.2. Reads that did not map to these RNA species were considered potential miRNAs and used in the subsequent analyses. Code relating to all analyses is available at: https://github.com/DaviesCentreInformatics/MicroRNA_BiVsBt. MiRNA sequencing reads are available from BioProject: PRJNA626458.

Messenger RNA from RNA-seq reads of the same samples (BioProject PRJNA626458) were aligned to ARS-UCD1.2 (Rosen *et al.* 2020) using the parameters described in Liu *et al.* (2021). Reads were quantified using featureCounts from the Rsubread package (v. 2.14.2) (Liao *et al.* 2019) and the ARS-UCD1.2 Ensembl gene annotation version 109.

Quantification of known and discovery of novel miRNAs

Potential miRNA reads were used as input for the miRDeep2 pipeline (Friedländer *et al.* 2011). We first used 'mapper.pl' from miRDeep2 with default parameters, except '-l', which we set to 17. This step produced collapsed reads and alignments in the miRDeep2 'arf' format, which were then used in the miRDeep2 quantification and discovery steps. As input to 'miRDeep2.pl', we used the collapsed reads from the 'mapper.pl' step, the ARS-UCD1.2 reference genome, the '.arf' file from the 'mapper.pl' step, mature and hairpin miRNAs belonging to *Bos taurus,* which is denoted as bta from miRbase (Kozomara & Griffiths-Jones 2011). We also used mature miRNAs from *Capra hircus* (chi) and *Ovis ares* (oar) in the miRDeep2 pipeline. We used a mirDeep2 score of ≥ 4, an estimated probability of being a true positive ≥ 70%, a significant Randfold p-value and the precursor location as an ID to identify novel miRNAs (Mukiibi *et al.* 2020).

Identification of differentially expressed miRNAs and mRNAs

We used the output from miRDeep2 to generate counts for differentially expressed miRNA (DEM) analysis. We followed a standard differential expression workflow using DESeq2 (Love *et al.* 2014). All miRNA samples were sequenced in a single batch. The model parameters were breed and sex, with the contrasts being the six pairwise comparisons of each breed and the comparison between males and females.

The mRNA counts were quantified using featureCounts, and the output was used as input to DESeq2 for differential expression analyses (Love *et al.* 2014). Here, the model parameters were breed, sex and batch. The contrasts were the same as those used in the miRNA analyses. Only DEGs and DEMs with an adjusted p-value less than 0.05 were considered significant.

miRNA target prediction

The miRanda (v. 3.3a) (Enright *et al.* 2003) miRNA target prediction software was used to identify possible gene targets of miRNAs. We extracted the three prime untranslated regions (3'UTRs) from the ARS-UCD1.2 Ensembl release 109. We then extracted the sequences of all mature miRNAs of interest, e.g., DEMs and group-specific miRNAs. We then aligned mature miRNA sequences to the 3'UTRs of ARS-UCD1.2 using miRanda with default parameters (Enright *et al.* 2003).

miRNA and mRNA co-expression

The weighted gene co-expression network analysis (WGCNA) package (v. 1.72) was used to identify miRNA and mRNA co-expression networks (Langfelder & Horvath 2008). WGCNA enables users to identify which genes have similar expression profiles (likely co-expressed) and to correlate that gene expression with other data like phenotype data or miRNA expression.

WGCNA removes genes with insufficient counts across samples, outlier genes, and outlier samples from the normalised count matrix. We used a variance stabilising transformation to normalise the matrix, as recommended by Langfelder and Horvath (2008). The input matrix is of the form $m \times n$ where m is the number of samples and n is the number of genes.

We then constructed the gene co-expression network by calculating an adjacency matrix from the filtered and normalised count matrix. The adjacency matrix is an $n \times n$ matrix, where n is the number of genes in the count matrix. The adjacency matrix is populated with values between 0 and 1, such that a_{ij} gives the connection strength between gene i and gene j. We used a 'softPower' of five and seven in calculating the adjacency matrix for the parent of origin and breed comparisons, respectively. We then transformed the adjacency matrix into a topological overlap matrix (TOM) to calculate which genes have high topological overlap, i.e., are connected to roughly the same genes as one another. We then identified the dissimilarity TOM by subtracting the TOM from 1. Following this, we performed hierarchical clustering to identify genes that grouped into modules of co-expressed genes. Each module needed to contain at least 90 genes to be considered separate from another module, and modules needed a correlation of at least 0.8 to be merged. We then correlated the expression of these gene modules with the genetic groups and miRNA expression data. We considered any correlation between mRNA and miRNA and mRNA and genetic group with a p-value below 0.05 as significant.

DEGs in miRNA targets

To determine if the predicted targets of a given miRNA were more likely to be DE within the present study, we tested how likely the targets of each miRNA were to be DE over all DE genes. Here, we defined the number of trials (*n*) as the number of predicted targets for a given miRNA, I.e., there were *n* targets that could be DE or not. The successes (*k*) were the number of predicted targets that were also DE. Finally, the probability of success (*p*) was the number of DEGs in a given comparison divided by the total number of genes in the count matrix. We paired the DEM and DEG comparisons such that if we were comparing DEMs identified between BIBI and BTBT, we only considered DEGs that were also identified between BIBI and BTBT.

KEGG pathway analyses

The clusterProfiler R package (v. 4.8.1) (Yu *et al.* 2012) was used to perform KEGG pathway analysis. The 'enrichKEGG' function was used to identify KEGG pathways significantly enriched in target genes. We used a cut-off of 0.05 for p-values, and the Benjamini-Hochberg method was used to adjust p-values to give q-values, which we also used a cut-off of 0.05 for significance testing. We performed KEGG analyses for the predicted targets of each miRNA that was either DE or specific to a particular group.

Determination of most frequently targeted pathways

To determine which pathways were most frequently targeted by either DE or group-specific miRNAs, we first performed target prediction for each miRNA. Next, we performed KEGG pathway enrichment with clusterProfiler (v. 4.8.1) (Yu *et al.* 2012) for the genes targeted by each miRNA. We then identified all unique pathways and counted the number of times they were targeted by a different miRNA. Word clouds to represent frequency of pathways were made using wordcloud (v. 1.9.2. http://amueller.github.io/word_cloud/) against Angus and Brahman cattle outline as background.

Results

miRNA and mRNA sequence quality

There was a sample average of ~16.9 million reads before filtering (S. table 2). No reads were shorter than our minimum length threshold, but around 588,000 reads per sample were longer than our maximum length threshold and thus were removed. This initial filtering left around 15.5 million reads to be assessed as potential miRNAs.

Just over 50% of reads were found to be other RNA species, i.e., not miRNAs (Figure 1A). Most non-miRNA RNA species were non-coding RNA (ncRNA), with an average of 7 million reads per sample mapped to known bovine ncRNA sequences (Figure 1A; S. table 2). After filtering, an average of ~8.2 million candidate miRNA reads remained for downstream analyses.

Our previous work has reported the RNA-seq alignment quality in detail (Liu *et al.* 2021). Briefly, after read trimming, an average of 49.9 million reads remained for alignment, and between 72 and 77% of them could be assigned to genes.

Known miRNA expression and novel miRNA profiles

There were 414 and 416 known miRNAs with a normalised count of at least one in all BIBI and BTBT samples, respectively (Figure 1B; S. table 3). Most (388) known miRNAs were shared between BIBI and BTBT samples (Figure 1B). BTBT samples had 28 miRNAs that were not expressed in BIBI samples, and BIBI had 26 miRNAs not expressed in BTBT samples (Figure 1B). 21 of the 24 samples had bta-miR-122 as the most highly expressed miRNA, with an average of 39% of the reads for each sample mapped to this miRNA.

When we separated samples based on the breed of their parents, there were between 376 and 398 miRNAs with a normalised count of one in all 12 samples for each parental group, i.e., maternal BI, maternal BT, paternal BI, paternal BT (Figure 1C and 1D). We observed 361 miRNAs shared between the maternal BT and maternal BI groups; this was mirrored when we compared paternal BI and paternal BT groups. The maternal BI group samples had 37 miRNAs not expressed in the maternal BT group. In contrast, we identified 15 miRNAs unique to the maternal BT group compared to the maternal BI group (Figure 1B). The inverse of this pattern was observed between paternal groups, where samples from paternal BT had 33 unique miRNAs compared to 16 in those from the paternal BI group (Figure 1C). There were 163 known miRNAs from miRbase with no expression in any sample.

We identified seven novel miRNAs in BTBT and BIBI samples (S. table 4). We considered these high-confidence novel miRNAs as they were found in all samples within a particular group. Six of these novel miRNAs were shared between BIBI and BTBT samples, with one being exclusive to each group, i.e., one novel miRNA was found in all six BTBT samples but not BIBI samples and vice versa (S. table 4).

Ten miRNAs can differentiate the breeds

We performed principal component analyses to determine if the samples clustered by breed using the miRNA data as with RNA-seq (Liu *et al.* 2021) and whole genome bisulfite sequencing data (Chapter 4). We observed no discernible clustering pattern when using all miRNAs (Figure 1E). We then filtered the miRNA data and used the top ten most variable miRNAs, as measured by variance. We subsequently observed a clear demarcation between BTBT and BIBI samples (Figure 1F). The hybrid groups (BIBT; BTBI) showed considerable overlap with one another and the purebred samples. The BTBT samples tended to cluster more closely to other samples within their group than any other group with BIBT, BTBI and BIBI samples spanning the full range of PC1 (Figure 1F).

Gene expression modules correlated with breed

Using the WGCNA method (Langfelder & Horvath 2008), we identified seven gene modules with significant correlations and one close to our significance cut-off (p = 0.09) with BIBI and BTBT (Figure 2A). These modules comprised 154, 343, 3,603, 1,630, 221, 780, 787 and 189 genes, with 1, 10, 141, 41, 2, 18, 21 and 5 being differentially expressed between BIBI and BTBT samples, respectively. We identified 2,525 DEGs between BIBI and BTBT, and together (Table 1), these eight modules contained ~9% of DEGs identified between the two groups.

A significant correlation is defined as modules with a p-value < 0.05. Modules A, B, D and E displayed significant positive correlations with BTBT samples and significant negative correlations with BIBI samples. Module C had a positive trend with BTBT samples and a negative trend with BIBI samples but was not statistically significant (p = 0.09). Modules F, G and H displayed significant negative correlations with BTBT samples and significant positive correlations with BIBI samples.

We then performed KEGG pathway analysis on the genes within each module to determine if they were enriched in any pathways. Among the eight gene modules correlated with breed, four modules were enriched with pathways. Module br-C, while not significantly correlated with breed (p = 0.09), displayed an enrichment of 32 pathways, including thermogenesis, TGF-beta signalling, mTOR signalling, neurotrophin signalling and Ras signalling (S. table 5). Module br-D was enriched in eight pathways; amino sugar and nucleotide sugar metabolism, biosynthesis of nucleotide sugars, apoptosis, MAPK signalling, phosphatidylinositol signalling, one carbon pool by folate and lipid and atherosclerosis pathways. Module br-F displayed a single enrichment in the glutamatergic synapse pathway. Module br-G had enrichments in relaxin signalling, chemical carcinogenesis – reactive oxygen species, nicotinate and nicotinamide metabolism, thermogenesis, retrograde endocannabinoid signalling, GnRH signalling and purine metabolism pathways. Modules br-A, br-B, br-E and br-H were not enriched in any pathways.

Gene expression modules correlated with dam of origin

Three gene modules (m-A, m-B, m-C) significantly correlated with maternal BI or BT groups (Figure 2B). These modules contained 1,457, 283 and 281 genes, respectively, with 130, 71 and 13 differentially expressed between maternal groups. There were 635 DEGs identified between the maternal groups (Table 1), and ~34% were within the three correlated gene modules.

Modules m-A and m-C were positively correlated with the maternal BI group and negatively correlated with maternal BT. Module m-B was negatively correlated with maternal BI samples and positively correlated with maternal BT samples (Figure 2B).

All three modules that correlated with the maternal breed were enriched in pathways (S. table 5). Module m-A showed enrichment for purine metabolism, bile secretion, glutamatergic synapse and pancreatic secretion pathways. Module m-B was enriched in ECM-receptor interaction, focal adhesion, human papillomavirus infection, amoebiasis, PI3K-Akt signalling, small cell lung cancer and AGE-RAGE signalling pathway in diabetic complications pathways. Module m-C displayed enrichments in neutrophil extracellular trap formation, systemic lupus erythematosus, alcoholism, cell cycle, viral carcinogenesis, DNA replication, amyotrophic lateral sclerosis, nucleocytoplasmic transport, gap junction and small cell lung cancer.

Gene expression modules correlated with sire of origin

We identified four gene modules with significant correlations and one module close to our cut-off (p = 0.07) with the paternal groups (Figure 2C). There were 1,457, 7,071, 3,852, 118 and 1,307 genes in each module, respectively. Of these, 626, 701, 468, 10 and 98 were differentially expressed, meaning ~65% of the 2,097 DEGs identified between paternal groups were contained within these five gene modules.

Modules p-A and p-B were positively correlated with paternal BI samples and negatively correlated with paternal BT samples. The remaining modules (pC-E) were negatively correlated with paternal BI and positively correlated with paternal BT samples (Figure 2C).

All five modules correlated with paternal breeds were enriched in pathways (S. table 5). Purine metabolism, bile secretion, glutamatergic synapse and pancreatic secretion were all enriched in module p-A. Module p-B was enriched in 84 pathways; too many to list; refer to (S. table 5). Module p-C saw an enrichment in valine, leucine and isoleucine degradation, sphingolipid metabolism, MAPK signalling, proteoglycans in cancer, propanoate metabolism, sphingolipid signalling pathway and pantothenate and CoA biosynthesis pathways. Module p-D was enriched for the spliceosome pathway. Module p-E was enriched for 20 pathways and among these were thermogenesis and purine metabolism.

Breed-specific DEMs

There were 14 differentially expressed miRNAs (DEMs) between BIBI and BTBT samples (Figure 2D; Table 1). Three DEMs displayed higher expression in BTBT samples; bta-miR-11998, bta-miR-2313-5p, and bta-miR-6518. These three miRNAs showed significant positive correlations with at least one breed-correlated module (Figure 2D). Bta-miR-11998 was negatively correlated with module br-H and displayed a trend with module br-G (p = 0.05). Modules br-C to E were positively correlated with bta-miR-2313-5p, and module br-B was positively correlated with bta-miR-6518.

The remaining 11 DEMs identified between BIBI and BTBT (bta-miR-11984, bta-miR-187, bta-miR-2284c, bta-miR-2285aj-5p, bta-miR-2285cp, bta-miR-2285cz, bta-miR-2355-3p, bta-miR-2419-3p, bta-miR-2419-5p, bta-miR-2481, bta-miR-6522) were upregulated in BIBI samples (Table 1; S. table 6). All 11 displayed positive correlations with modules positively correlated with the BIBI breed (Figures 2A and 2D). Module br-A displayed significant negative correlations with bta-miR-19984, bta-miR-2285cz, bta-miR-2355-3p and bta-miR-2481. Module br-B exhibited negative correlations with bta-miR-187, bta-miR-2284c, bta-miR-2285-aj-5p, bta-miR-2285cp, bta-miR-2419-3p, bta-miR-2419-5p, bta-miR-2481 and bta-miR-6522. Module br-C displayed no significant correlations with any miRNAs. MiRNAs bta-miR-11984, bta-miR-187, bta-miR-2284c, bta-miR-2285cp, bta-miR-2285cz, bta-miR-2355-3p, bta-miR-2419-3p, bta-miR-2419-5p and bta-miR-2481 all displayed negative correlations with module br-D. Module br-E showed negative correlations with bta-miR-2419-3p, bta-miR-2419-5p and bta-miR-2481. The miRNAs bta-miR-187, bta-miR-2284c, bta-miR-2285aj-5p, bta-miR-2285cp, bta-miR-2285cz, bta-miR-2355-3p, bta-miR-2419-3p, bta-miR-2419-5p, bta-miR-2481 and bta-miR-6522 were all positively correlated with module br-F. These miRNAs were also positively correlated with module br-G except for bta-miR-11984, which was only positively correlated with br-G, not br-F and bta-miR-6522, which was not significantly correlated with br-G. Lastly, module br-H displayed positive correlations with bta-miR-2284c, bta-miR-2285aj-5p, bta-miR-2285cp, bta-miR-2419-5p and bta-miR-2481.

There were 26 BIBI-specific miRNAs, i.e., had a count of at least one in all samples within the BIBI group, with fifteen of these being significantly correlated with breed-correlated modules. Bta-miR-11985 was positively correlated with module br-C, and bta-miR-12003 was negatively correlated with this module. Bta-miR-190a was positively correlated with module br-C. Bta-miR-216b displayed negative correlations with modules br-A and br-D and a positive correlation with br-G. Bta-miR-2284b was negatively correlated with module br-B and positively correlated with br-F and br-H. This pattern was also observed in bta-mir-2284d, with the addition of a positive correlation with module br-G. Bta-mir-2285c was solely negatively correlated with module br-A. Bta-miR-677 was only positively correlated with module br-E.

Thirteen of the 28 BTBT-specific miRNAs displayed significant correlations with gene modules associated with breed (Figure 2D). Bta-miR-11990 was positively correlated with modules br-C and br-D. Bta-miR-1301 was only positively correlated with module br-A. Modules br-D and br-E were positively correlated with bta-miR-184. Module br-G and br-H were positively correlated with bta-miR-2285az. Module br-G displayed a negative correlation with bta-miR-2285cx. Modules br-D and br-F displayed positive and negative correlations with bta-miR-2316, respectively. Module br-A was positively correlated with bta-miR-2330-3p. Bta-miR-2415-3p was positively correlated with modules br-D and br-E and negatively correlated with modules br-F to br-H. Bta-miR-2440 was positively correlated with modules br-B, br-D, and br-E and negatively correlated with modules br-F to br-H. Bta-miR-490 was positively correlated with modules br-C and br-D.

Dam of origin-specific miRNAs

Only one miRNA was differentially expressed between maternal groups (bta-miR-187). Module m-A was positively correlated with this miRNA, and module m-B displayed a negative correlation (Figure 2E). Most of the miRNAs identified as only occurring in the maternal BI or maternal BT groups occurred in the maternal BI group

(Figure 2E). The maternal BI-specific miRNAs with significant correlations were bta-miR-11984, bta-miR-2284c, bta-miR-2285c, bta-miR-2285cp, bta-miR-2285cz and bta-miR-2355-3p; these were positively correlated with module m-A. MiRNAs bta-miR-2285cz and bta-miR-2355-3p were the only miRNAs that had a significant correlation with module m-B, both of which were negative. Of the miRNAs only expressed in the maternal BT group, bta-miR-11998, bta-miR-2331-5p, bta-miR-2415-3p, bta-miR-2440 and bta-miR-4449 were all negatively correlated with module m-A. Module m-C exhibited no significant correlations.

Sire of origin-specific miRNAs

Five DEMs were identified between paternal BI and paternal BT groups. Bta-miR-184, bta-miR-2313-5p and bta-miR-6518 were up-regulated in the paternal BT group, with bta-miR-2285c and bta-miR-2419-5p being upregulated in the paternal BI group (S. table 6). Among the miRNAs that were paternal BI-specific, bta-miR-11984, bta-miR-187, bta-miR-216b, bta-miR-2284c, bta-miR-2285cp, bta-miR-2355-3p and bta-miR-2419-3p were positively correlated with module p-A. Bta-miR-2285cp was negatively correlated with module p-C, whereas bta-miR-2397-3p was positively correlated; bta-miR-383 was negatively correlated with module p-E.

Despite paternal BT having many more unique miRNAs, only seven exhibited significant correlations with any paternal group correlated gene modules. Bta-miR-11998 was positively correlated with module p-A and negatively correlated with module p-C. Bta-miR-1277 was positively correlated with p-B, and bta-miR-2284n was positively correlated with module p-D. Bta-miR-2415-3p displayed a negative correlation with module p-A, as did bta-miR-2440 and bta-miR-365-3p. Bta-miR-365-

3p and bta-miR-4449 were also positively correlated with module p-C. Bta-miR-4449 was also negatively correlated with module p-B (Figure 2F). No significant correlations were identified between bta-miR-184 and the paternal gene modules. Bta-miR-2285c displayed positive correlations with module p-A and a negative correlation with module p-C. Bta-miR-2313-5p displayed a single significant correlation with module p-A. Bta-miR-2419-5p was positively correlated with module p-A. Bta-miR-6518 displayed negative correlations with modules p-A and p-B and a positive correlation with module p-C (Figure 2F).

Predicted targets of DEMs more likely to be DEGs

We compared the predicted targets of each DEM identified between BIBI and BTBT samples and determined whether they were more likely to be DE than the background genes, i.e., genes not predicted to be targeted by a given miRNA. We found that predicted targets of 12 of the 14 DEMs identified between BIBI and BTBT samples were significantly more likely to be DE than non-target genes (binomial test, p <0.05) (S. table 7). Furthermore, the targets of the single DEM identified between maternal BI and maternal BT were significantly more likely to be DE (binomial test, p <0.05). The targets of four of the five DEMs identified between paternal groups were significantly more likely to be DE than non-target genes (binomial test, p <0.05).

Pathways predicted to be targeted by DE and group-specific miRNAs

There were 23, 15, 33, 29, 37 and 17 miRNAs upregulated or only expressed in breed BTBT, maternal BT, paternal BT, breed BIBI, maternal BI and paternal BI, respectively (Figure 3A; S. table 8). We used the miRanda target prediction software (v. 3.3a) (Enright *et al.* 2003) to identify potential targets of each miRNA (Figure 3).

Between 300 and 5,115 targets were predicted among these miRNAs (S. table 9), which were then used to perform KEGG pathway analysis. Genes involved in the MAPK and Rap1 signalling pathways were consistently targeted by miRNAs, with between 37% and 56% of miRNAs in each group targeting these pathways (S. table 8). All pathways targeted by miRNAs expressed in each group can be seen in Figure 3 and Supplementary Table 8.

No DEMs were identified between the sexes

To determine if any differences between the sexes existed, independent of breed differences, we compared all males against all females to search for DEMs by sex. We observed no significant DEMs between the male and female samples. We also found no significant correlations between gene modules and sex. As we observed no significant correlations between sex and gene modules and no significant DEMs by sex, we performed no pathway analysis for this comparison.

Discussion

To our knowledge, this study is the first to report on miRNA expression differences between taurine and indicine cattle breeds and correlating the results with mRNA expression. This study utilised miRNA expression data from fetal liver samples representing male and female individuals of Brahman, Angus and their reciprocal crosses. The inclusion of reciprocal crosses enables us to disentangle not only miRNA expression differences that might be related to breed but also whether the genetics of the dam or sire has an impact on miRNA expression differences, which is a novel aspect of this work. Previous efforts to understand what drives these breed differences have focused on the adult stage of development. For example, Deb and

Sengar (2021) investigated the miRNA expression profiles between Sahiwal (indicine) and Frieswal (indicine x taurine) cattle breeds in response to summer heat stress, identifying DEMs that interact with heat shock protein 70 (*Hsp*70). Similarly, Dong *et al.* (2023) investigated miRNA expression differences in the testes of Mongolian (taurine) and Hainan (indicine) cattle, determining that breed differences in spermatogenesis-related miRNAs existed. While this study did investigate differences between taurine and indicine cattle, the niche nature of the breeds and tissue studied potentially limits the applicability of their findings. Although these studies provide valuable insight into miRNA differences between cattle subspecies, they have not utilised reciprocal crosses and thus have been unable to differentiate the impact of maternal or paternal genetics on miRNA expression.

Our previous study using mRNA expression data from these samples was able to clearly distinguish BTBT, BIBT, BTBI and BIBI samples from each other (Liu *et al.* 2021). In contrast, only BIBI and BTBT samples were clearly distinguishable in this study. Furthermore, despite observing over thousands of differentially expressed genes between Brahman and Angus samples, we identified only 14 DEMs between the two breeds. This pattern is similar to a study that observed only 23 DEMs when comparing miRNA expression in Hereford x Limousine (beef breed) and Holstein-Friesian (dairy breed) muscle cells during myogenic differentiation (Sadkowski *et al.* 2018). As a single miRNA target multiple mRNAs (Peterson *et al.* 2014), the relatively low number of DEMs is not unexpected and these miRNAs profoundly impact differentially expressed genes between breeds.

Using the mRNA expression data, we were able to identify gene modules that were correlated with the two cattle breeds. Moreover, by correlating this same gene expression data with corresponding miRNA expression, we could identify miRNAs that may be contributing to the gene expression that distinguishes the two breeds. For example, module br-C was positively correlated with BTBT samples and was enriched for several pathways, including thermogenesis and Ras signalling. Ras signalling plays an essential role in adipogenesis (Murholm *et al.* 2010), with adipocyte hyperplasia underway in cattle at this developmental stage (Zhao *et al.* 2019). Interestingly, bta-miR-11990 was predicted to target genes involved in the Ras signalling pathway, and this miRNA shares a similar correlation pattern with several BTBT-specific miRNAs (bta-miR-2313-5p, bta-miR-490, bta-miR-2316). As Angus cattle are known to have superior fat gain performance in cold climates (Boyles & Riley 1991), it may be possible that these miRNAs are contributing to post-transcriptional regulation that conveys this trait to Angus cattle.

An advantage of including reciprocal crosses in comparisons of breeds is that it enables the investigation of any dam or sire of origin effects. We observed a single DEM between different maternal groups. In addition to the single DEM (bta-miR-187), we observed several miRNAs that were only expressed in maternal BI samples (bta-miR-2284c, bta-miR-2285c, bta-miR-2285cp, bta-miR-2285cz, bta-miR-2355-3p, bta-miR-11984) that was only found in the maternal BI group. These miRNAs were positively correlated with module m-A, and one of them (bta-miR-2355-3p) was predicted to target genes involved in the glutamatergic synapse pathway. Module m-A was positively correlated with the maternal BI group and was enriched for several pathways, including glutamatergic synapse. We observed a similar pattern in the sire of origin comparison, with bta-miR-2355-3p being positively correlated to module p-A, which was also enriched for the glutamatergic synapse pathway. Several miRNAs (bta-miR-187, bta-miR-216b, bta-miR-2284c, bta-miR-2285c, bta-miR-2285cp, bta-miR-2419-3p, bta-miR-2419-5p, bta-miR-11984) shared a similar correlation pattern to bta-miR-2355-3p. The glutamatergic synapse has a known role in heat tolerance of an individual as glutamatergic neurons transmit peripheral and central heat signals to the hypothalamic preoptic area of the brain (Sun *et al.* 2022), which then begins a coordinated response to lower the temperature. The liver is known to play an integral role in coordinating this heat stress response via increased production of heat shock proteins, increasing metabolic rate and increased vasodilation (Thorne *et al.* 2020).

Further evidence to support the possible role of glutamatergic synapses in conveying heat tolerance in cattle can be found in another recent study of dairy cattle (Cheruiyot *et al.* 2021). Given the developmental timeline of the liver and the observed pathway enrichments (Tiniakos *et al.* 1996; Giancotti *et al.* 2019), it is possible that bta-miR-2355-3p and those miRNAs with similar correlation patterns play a role in modulating neuron development in the liver, priming it for hotter temperatures later in life and that this can be conveyed by either a Brahman sire or dam.

While we observed several DEMs that were correlated to gene modules, many more miRNAs were observed that were only expressed in one of the two groups in each comparison, e.g., bta-miR-490 in the breed comparison, bta-miR-11998 in the maternal breed comparison and bta-miR-187 in the paternal breed comparison. There was some overlap between DEMs and group-specific miRNAs, e.g. bta-miR-184 in the paternal-breed comparison. However, this was due to the cut-off used to identify miRNAs being calculated differently from how DESeq2 determines if a gene or miRNA should be retained. In any case, there were considerably more miRNAs expressed in a single group than there were differentially expressed between groups. This pattern leads one to posit that group-specific miRNAs may be more important in driving gene regulatory differences than DEMs between Brahman and Angus cattle.

We did not observe any DEMs between males and females, which is consistent with the limited number of DEGs observed between male and female samples. The lack of sex-specific miRNA expression could be a result of the limited sex-specific expression that is observed in the liver prior to puberty (Conforto & Waxman 2012) and other mechanisms are driving DEGs by sex.

We identified a range of predicted targets for each of the differentially expressed and group-specific miRNAs and performed biological pathway analyses on these targets to gain insights into the pathways that may be affected. Notably, a substantial proportion of the targeted pathways identified in each of the three comparisons were signalling pathways, such as the Rap1, MAPK and Ras signalling pathways. Each of these pathways plays important roles in fetal development. Rap1 signalling is important for vascular morphogenesis (Chrzanowska-Wodnicka 2013), and has also been implicated as a possible cause of the differences between high and low-performing meat goat breeds (Shen *et al.* 2022). Additionally, Rap1 signalling ablation in the brain of mice has been shown to protect mice from high-fat diet-induced obesity (Kaneko *et al.* 2016), suggesting a critical role of Rap1 signalling in fat storage. While this study investigated Rap1 signalling in the brain, given the close interactions between the brain and liver to monitor glucose and lipid homeostasis, it could be that Rap1 signalling in the liver also has a role in fat storage. Ras and MAPK signalling pathways are critical to cell proliferation and differentiation (Zhang & Liu 2002). Moreover, the Ras signalling pathway is purported to have a role in adipogenesis as ectopic expression of the pathway can induce preadipocyte formation in the absence of insulin and insulin-like growth factor 1 (*Igf-1*) (MacDougald & Lane 1995), suggesting a possible role in modulating adipogenesis in the fetus. In addition, MAPK signalling can act as a negative regulator of muscle development (Xie *et al.* 2018), suggesting that differences in how this pathway is regulated may influence how the fetus develops muscle. The differences in miRNA expression between taurine and indicine cattle may result in differential modulation of signalling cascades involved in fetal liver development and growth regulation, which in turn contributes to the observed phenotypic differences.

Conclusions

This study sheds light on the miRNA expression differences between taurine and indicine cattle using two global economically important breeds. We identified several miRNAs that may play a role in controlling economically important traits in these breeds, such as fat gain and heat tolerance. This study has identified miRNAs that may play important roles in how traits may be conferred to the developing fetus and provides valuable biological insight into the possible mechanisms of how these traits are controlled.

References

Ajmone-Marsan P., Garcia J.F. & Lenstra J.A. (2010) On the origin of cattle: How aurochs became cattle and colonized the world. Evolutionary Anthropology: Issues, News, and Reviews 19, 148-57. https://doi.org/10.1002/evan.20267

Akanno E.C., Chen L., Abo-Ismail M.K., Crowley J.J., Wang Z., Li C., Basarab J.A., MacNeil M.D. & Plastow G.S. (2018) Genome-wide association scan for heterotic quantitative trait loci in multi-breed and crossbred beef cattle. Genetics Selection Evolution 50, 48. 10.1186/s12711-018-0405-y

Andrade A.N.N., Hernandez A., Rodriguez E.E., Davila K.M.S., Mateescu R. & Rodríguez E. (2022) 38 Effect of Breed Composition and Genome-Wide Association Study on Epidermis Thickness in a Multibreed Angus-Brahman Population. Journal of Animal Science 100, 9-. 10.1093/jas/skac247.015

Andrews S. (2010) FastQC.

Awamleh Z., Gloor G.B. & Han V.K.M. (2019) Placental microRNAs in pregnancies with early onset intrauterine growth restriction and preeclampsia: potential impact on gene expression and pathophysiology. BMC Medical Genomics 12, 91. 10.1186/s12920-019-0548-x

Barlow D.P., Stöger R., Herrmann B., Saito K. & Schweifer N. (1991) The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. Nature 349, 84-7.

Billa P.A., Faulconnier Y., Ye T., Chervet M., Le Provost F., Pires J.A.A. & Leroux C. (2019) Deep RNA-Seq reveals miRNome differences in mammary tissue of lactating Holstein and Montbéliarde cows. BMC Genomics 20, 621. 10.1186/s12864-019-5987-4

Boyles S.L. & Riley J.G. (1991) Feedlot performance of Brahman x Angus versus Angus steers during cold weather. Journal of Animal Science 69, 2677-84. 10.2527/1991.6972677x

Bruford M.W., Bradley D.G. & Luikart G. (2003) DNA markers reveal the complexity of livestock domestication. Nature Reviews Genetics 4, 900-10. 10.1038/nrg1203

Casas E., Thallman R. & Cundiff L. (2011) Birth and weaning traits in crossbred cattle from Hereford, Angus, Brahman, Boran, Tuli, and Belgian Blue sires. Journal of Animal Science 89, 979-87.

Cheruiyot E.K., Haile-Mariam M., Cocks B.G., MacLeod I.M., Xiang R. & Pryce J.E. (2021) New loci and neuronal pathways for resilience to heat stress in cattle. Scientific Reports 11, 16619. 10.1038/s41598-021-95816-8

Chrzanowska-Wodnicka M. (2013) Distinct functions for Rap1 signaling in vascular morphogenesis and dysfunction. Experimental Cell Research 319, 2350-9. 10.1016/j.yexcr.2013.07.022

Conforto T.L. & Waxman D.J. (2012) Sex-specific mouse liver gene expression: genome-wide analysis of developmental changes from pre-pubertal period to young adulthood. Biology of Sex Differences 3, 9. 10.1186/2042-6410-3-9

Deb R. & Sengar G.S. (2021) Comparative miRNA signatures among Sahiwal and Frieswal cattle breeds during summer stress. 3 Biotech 11, 79. 10.1007/s13205-020-02608-4

DeChiara T.M., Robertson E.J. & Efstratiadis A. (1991) Parental imprinting of the mouse insulin-like growth factor II gene. Cell 64, 849-59. 10.1016/0092-8674(91)90513-x

Di Tommaso P., Chatzou M., Floden E.W., Barja P.P., Palumbo E. & Notredame C. (2017) Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316-9. 10.1038/nbt.3820

Dikmen S., Mateescu R.G., Elzo M.A. & Hansen P.J. (2018) Determination of the optimum contribution of Brahman genetics in an Angus-Brahman multibreed herd for regulation of body temperature during hot weather. Journal of Animal Science 96, 2175-83. 10.1093/jas/sky133

Dong Z., Ning Q., Liu Y., Wang S., Wang F., Luo X., Chen N. & Lei C. (2023) Comparative transcriptomics analysis of testicular miRNA from indicine and taurine cattle. Animal Biotechnology 34, 1436-46. 10.1080/10495398.2022.2029466

Elzo M.A., Johnson D.D., Wasdin J.G. & Driver J.D. (2012) Carcass and meat palatability breed differences and heterosis effects in an Angus–Brahman multibreed population. Meat Science 90, 87-92. https://doi.org/10.1016/j.meatsci.2011.06.010

Enright A.J., John B., Gaul U., Tuschl T., Sander C. & Marks D.S. (2003) MicroRNA targets in Drosophila. Genome Biology 5, R1. 10.1186/gb-2003-5-1-r1

Friedländer M.R., Mackowiak S.D., Li N., Chen W. & Rajewsky N. (2011) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Research 40, 37-52. 10.1093/nar/gkr688

Galagali H. & Kim J.K. (2020) The multifaceted roles of microRNAs in differentiation. Current Opinion in Cell Biology 67, 118-40. 10.1016/j.ceb.2020.08.015

Giancotti A., Monti M., Nevi L., Safarikia S., D’Ambrosio V., Brunelli R., Pajno C., Corno S., Di Donato V., Musella A., Chiappetta M.F., Bosco D., Panici P.B., Alvaro D. & Cardinale V. (2019) Functions and the Emerging Role of the Foetal Liver into Regenerative Medicine. Cells 8. 10.3390/cells8080914

Gobena M., Elzo M.A. & Mateescu R.G. (2018) Population Structure and Genomic Breed Composition in an Angus–Brahman Crossbred Cattle Population. Frontiers in Genetics 9. 10.3389/fgene.2018.00090

Goszczynski D.E., Corbi-Botto C.M., Durand H.M., Rogberg-Muñoz A., Munilla S., Peral-Garcia P., Cantet R.J.C. & Giovambattista G. (2018) Evidence of positive selection towards Zebuine haplotypes in the BoLA region of Brangus cattle. Animal 12, 215-23. https://doi.org/10.1017/S1751731117001380

Ha M. & Kim V.N. (2014) Regulation of microRNA biogenesis. Nature Reviews Molecular Cell Biology 15, 509-24. 10.1038/nrm3838

Haig D. (2014) Coadaptation and conflict, misconception and muddle, in the evolution of genomic imprinting. Heredity (Edinb) 113, 96-103. 10.1038/hdy.2013.97

Hao D., Wang X., Yang Y., Chen H., Thomsen B. & Holm L.-E. (2023) MicroRNA sequence variation can impact interactions with target mRNA in cattle. Gene 868, 147373. https://doi.org/10.1016/j.gene.2023.147373

Hiendleder S., Lewalski H. & Janke A. (2008) Complete mitochondrial genomes of Bos taurus and Bos indicus provide new insights into intra-species variation, taxonomy and domestication. Cytogenetic and Genome Research 120, 150-6. 10.1159/000118756

Jopling C. (2012) Liver-specific microRNA-122: Biogenesis and function. RNA Biol 9, 137-42. 10.4161/rna.18827

Kalvari I., Nawrocki E.P., Argasinska J., Quinones-Olvera N., Finn R.D., Bateman A. & Petrov A.I. (2018) Non-Coding RNA Analysis Using the Rfam Database. Curr Protoc Bioinformatics 62, e51. 10.1002/cpbi.51

Kalvari I., Nawrocki E.P., Ontiveros-Palacios N., Argasinska J., Lamkiewicz K., Marz M., Griffiths-Jones S., Toffano-Nioche C., Gautheret D., Weinberg Z., Rivas E., Eddy S.R., Finn Robert D., Bateman A. & Petrov A.I. (2020) Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Research 49, D192-D200. 10.1093/nar/gkaa1047

Kaneko K., Xu P., Cordonier E.L., Chen S.S., Ng A., Xu Y., Morozov A. & Fukuda M. (2016) Neuronal Rap1 Regulates Energy Balance, Glucose Homeostasis, and Leptin Actions. Cell Rep 16, 3003-15. 10.1016/j.celrep.2016.08.039

Kim D.H., Sætrom P., Snøve O. & Rossi J.J. (2008) MicroRNA-directed transcriptional gene silencing in mammalian cells. Proceedings of the National Academy of Sciences 105, 16230-5. doi:10.1073/pnas.0808830105

Kozomara A. & Griffiths-Jones S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic acids research. 39, D152-D7. 10.1093/nar/gkq1027

Langfelder P. & Horvath S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559. 10.1186/1471-2105-9-559

Li B., Dong J., Yu J., Fan Y., Shang L., Zhou X. & Bai Y. (2020) Pinpointing miRNA and genes enrichment over trait-relevant tissue network in Genome-Wide Association Studies. BMC Medical Genomics 13, 191. 10.1186/s12920-020-00830-w

Li Z., He J., Yang F., Yin S., Gao Z., Chen W., Sun C., Tait R.G., Bauck S., Guo W. & Wu X.-L. (2023) A look under the hood of genomic-estimated breed compositions for brangus cattle: What have we learned? Frontiers in Genetics 14. 10.3389/fgene.2023.1080279

Liao Y., Smyth G.K. & Shi W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Research 47, e47-e. 10.1093/nar/gkz114

Liu G., Liao Y., Sun B., Guo Y., Deng M., Li Y. & Liu D. (2020) Effects of chronic heat stress on mRNA and miRNA expressions in dairy cows. Gene 742, 144550. https://doi.org/10.1016/j.gene.2020.144550

Liu R., Tearle R., Low W.Y., Chen T., Thomsen D., Smith T.P.L., Hiendleder S. & Williams J.L. (2021) Distinctive gene expression patterns and imprinting signatures revealed in reciprocal crosses between cattle sub-species. BMC Genomics 22. 10.1186/s12864-021-07667-2

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 91, 2757-61. doi:10.1073/pnas.91.7.2757

Love M.I., Huber W. & Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550. 10.1186/s13059-014-0550-8

Low W.Y., Tearle R., Liu R., Koren S., Rhie A., Bickhart D.M., Rosen B.D., Kronenberg Z.N., Kingan S.B. & Tseng E. (2020) Haplotype-resolved genomes provide insights into structural variation and gene content in Angus and Brahman cattle. Nature Communications 11, 1-14.

MacDougald O.A. & Lane M.D. (1995) Transcriptional regulation of gene expression during adipocyte differentiation. Annual Review of Biochemistry 64, 345-73. 10.1146/annurev.bi.64.070195.002021

Macfarlane L.A. & Murphy P.R. (2010) MicroRNA: Biogenesis, Function and Role in Cancer. Current Genomics 11, 537-61. 10.2174/138920210793175895

MacHugh D.E., Larson G. & Orlando L. (2017) Taming the Past: Ancient DNA and the Study of Animal Domestication. Annual Review of Animal Biosciences 5, 329-51. 10.1146/annurev-animal-022516-022747

Martin M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, 10. 10.14806/ej.17.1.200

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 110, E1398-E406. doi:10.1073/pnas.1303367110

Morales-Roselló J., Loscalzo G., García-Lopez E.M., García-Gimenez J.L. & Perales-Marín A. (2022) MicroRNA-132 is overexpressed in fetuses with late-onset fetal growth restriction. Health Science Reports 5, e558. https://doi.org/10.1002/hsr2.558

Mukiibi R., Johnston D., Vinsky M., Fitzsimmons C., Stothard P., Waters S.M. & Li C. (2020) Bovine hepatic miRNAome profiling and differential miRNA expression analyses between beef steers with divergent feed efficiency phenotypes. Scientific Reports 10, 19309. 10.1038/s41598-020-73885-5

Murholm M., Dixen K. & Hansen J.B. (2010) Ras signalling regulates differentiation and UCP1 expression in models of brown adipogenesis. Biochim Biophys Acta 1800, 619-27. 10.1016/j.bbagen.2010.03.008

Peterson S., Thompson J., Ufkin M., Sathyanarayana P., Liaw L. & Congdon C.B. (2014) Common features of microRNA target prediction tools. Frontiers in Genetics 5. 10.3389/fgene.2014.00023

Rosen B.D., Bickhart D.M., Schnabel R.D., Koren S., Elsik C.G., Tseng E., Rowan T.N., Low W.Y., Zimin A., Couldrey C., Hall R., Li W., Rhie A., Ghurye J., McKay S.D., Thibaud-Nissen F., Hoffman J., Murdoch B.M., Snelling W.M., McDaneld T.G., Hammond J.A., Schwartz J.C., Nandolo W., Hagen D.E., Dreischer C., Schultheiss S.J., Schroeder S.G., Phillippy A.M., Cole J.B., Van Tassell C.P., Liu G., Smith T.P.L. & Medrano J.F. (2020) De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9, giaa021-giaa. 10.1093/gigascience/giaa021

Rottiers V. & Näär A.M. (2012) MicroRNAs in metabolism and metabolic disorders. Nature Reviews: Molecular Cell Biology 13, 239-50. 10.1038/nrm3313

Sadkowski T., Ciecierska A., Oprządek J. & Balcerek E. (2018) Breed-dependent microRNA expression in the primary culture of skeletal muscle cells subjected to myogenic differentiation. BMC Genomics 19, 109. 10.1186/s12864-018-4492-5

Schmieder R. & Edwards R. (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863-4. 10.1093/bioinformatics/btr026

Shen J., Hao Z., Luo Y., Zhen H., Liu Y., Wang J., Hu J., Liu X., Li S., Zhao Z., Liu Y., Yang S. & Wang L. (2022) Deep Small RNA Sequencing Reveals Important miRNAs Related to Muscle Development and Intramuscular Fat Deposition in Longissimus dorsi Muscle From Different Goat Breeds. Frontiers in Veterinary Science 9. 10.3389/fvets.2022.911166

Song C., Yang J., Jiang R., Yang Z., Li H., Huang Y., Lan X., Lei C., Ma Y., Qi X. & Chen H. (2019) miR-148a-3p regulates proliferation and apoptosis of bovine muscle cells by targeting KLF6. J Cell Physiol 234, 15742-50. 10.1002/jcp.28232

Sun J., Li M., Li Z., Xue J., Lan X., Zhang C., Lei C. & Chen H. (2013) Identification and profiling of conserved and novel microRNAs from Chinese Qinchuan bovine longissimus thoracis. BMC Genomics 14, 42. 10.1186/1471-2164-14-42

Sun J., Zhou Y., Cai H., Lan X., Lei C., Zhao X., Zhang C. & Chen H. (2014) Discovery of Novel and Differentially Expressed MicroRNAs between Fetal and Adult Backfat in Cattle. PLOS ONE 9, e90244. 10.1371/journal.pone.0090244

Sun Y., Wang H., Wang W., Lu J., Zhang J., Luo X., Luan L., Wang K., Jia J., Yan J. & Qin L. (2022) Glutamatergic and GABAergic neurons in the preoptic area of the hypothalamus play key roles in menopausal hot flashes. Frontiers in Aging Neuroscience 14. 10.3389/fnagi.2022.993955

Thorne A.M., Ubbink R., Brüggenwirth I.M.A., Nijsten M.W., Porte R.J. & Meijer V.E.d. (2020) Hyperthermia-induced changes in liver physiology and metabolism: a rationale for hyperthermic machine perfusion. American Journal of Physiology-Gastrointestinal and Liver Physiology 319, G43-G50. 10.1152/ajpgi.00101.2020

Tiniakos D.G., Lee J.A. & Burt A.D. (1996) Innervation of the liver: morphology and function. Liver 16, 151-60. https://doi.org/10.1111/j.1600-0676.1996.tb00721.x

Vrijens K., Bollati V. & Nawrot T.S. (2015) MicroRNAs as potential signatures of environmental exposure or effect: a systematic review. Environmental Health Perspectives 123, 399-411. 10.1289/ehp.1408459

Wang L., Zhang S., Zhang W., Cheng G., Khan R., Junjvlieke Z., Li S. & Zan L. (2020) miR-424 Promotes Bovine Adipogenesis Through an Unconventional Post-Transcriptional Regulation of STK11. Frontiers in Genetics 11. 10.3389/fgene.2020.00145

Whipple A.J., Breton-Provencher V., Jacobs H.N., Chitta U.K., Sur M. & Sharp P.A. (2020) Imprinted Maternally Expressed microRNAs Antagonize Paternally Driven Gene Programs in Neurons. Molecular Cell 78, 85-95.e8. 10.1016/j.molcel.2020.01.020

Xie S.-J., Li J.-H., Chen H.-F., Tan Y.-Y., Liu S.-R., Zhang Y., Xu H., Yang J.-H., Liu S., Zheng L.-L., Huang M.-B., Guo Y.-H., Zhang Q., Zhou H. & Qu L.-H. (2018) Inhibition of the JNK/MAPK signaling pathway by myogenesis-associated miRNAs is required for skeletal muscle development. Cell Death & Differentiation 25, 1581-97. 10.1038/s41418-018-0063-1

Yan X., Huang Y., Zhao J.X., Rogers C.J., Zhu M.J., Ford S.P., Nathanielsz P.W. & Du M. (2013) Maternal obesity downregulates microRNA let-7g expression, a possible mechanism for enhanced adipogenesis during ovine fetal skeletal muscle development. Int J Obes (Lond) 37, 568-75. 10.1038/ijo.2012.69

Yu G., Wang L.G., Han Y. & He Q.Y. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. Omics 16, 284-7. 10.1089/omi.2011.0118

Zhang W. & Liu H.T. (2002) MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Research 12, 9-18. 10.1038/sj.cr.7290105

Zhao L., Huang Y. & Du M. (2019) Farm animals for studying muscle development and metabolism: dual purposes for animal production and human health. Animal Frontiers 9, 21-7. 10.1093/af/vfz015

Tables

Table 1. DEMs and DEGs of breed and parent-of-origin effects

	BIBI - BTBT	Maternal BI – Maternal BT	Paternal BI – Paternal BT
miRNA			
Up	11	1	2
Not significant	910	923	919
Down	3	0	3
Total DE	**14**	**1**	**5**
mRNA			
Up	1,290	279	1,030
Not significant	16,128	18,018	16,556
Down	1,235	356	1,067
Total DE	**2,525**	**635**	**2,097**

The reference groups for the three comparisons are BTBT, Maternal BT and Paternal BT.

Figures

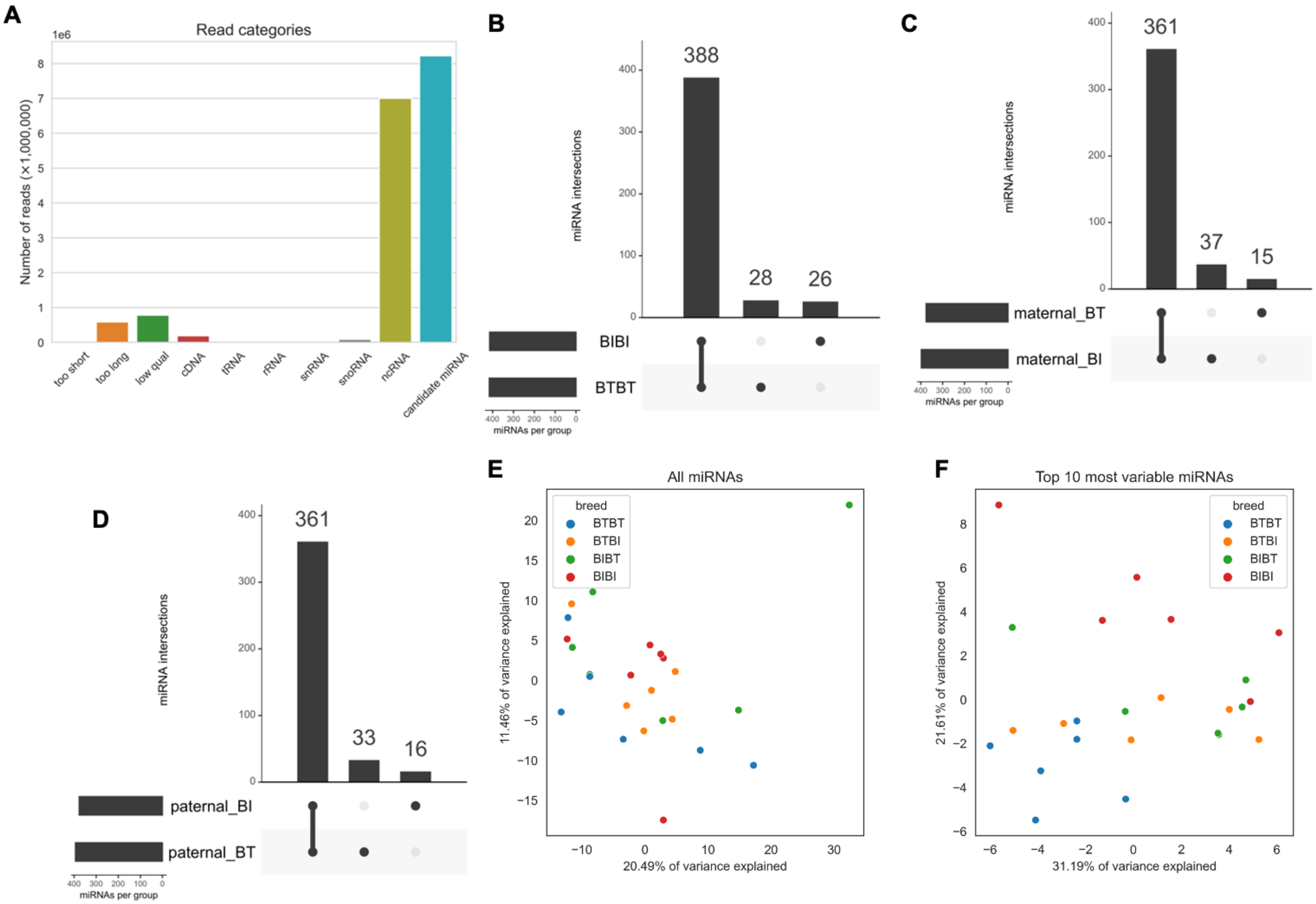

A
Read categories
1e6
Number of reads (×1,000,000)
0
1
2
3
4
5
6
7
8
too short
too long
low qual
cDNA
tRNA
rRNA
snRNA
snoRNA
ncRNA
candidate miRNA
B
miRNA intersections
0
100
200
300
400
388
28
26
BIBI
BTBT
400 300 200 100 0
miRNAs per group
C
miRNA intersections
0
100
200
300
400
361
37
15
maternal_BT
maternal_BI
400 300 200 100 0
miRNAs per group
D
miRNA intersections
0
100
200
300
400
361
33
16
paternal_BI
paternal_BT
400 300 200 100 0
miRNAs per group
E
All miRNAs
breed
BTBT
BTBI
BIBT
BIBI
11.46% of variance explained
−15
−10
−5
0
5
10
15
20
−10
0
10
20
30
20.49% of variance explained
F
Top 10 most variable miRNAs
breed
BTBT
BTBI
BIBT
BIBI
21.61% of variance explained
−6
−4
−2
0
2
4
6
8
−6
−4
−2
0
2
4
6
31.19% of variance explained

Figure 1. A.) Bar chart showing mean proportions of reads belonging to the filtered, cDNA, tRNA, ncRNA, rRNA, snoRNA, snRNA, too short, too long and “low qual” categories. Filtered refers to reads that are not contaminants and passed length and quality thresholds. cDNA, tRNA, ncRNA, snoRNA and snRNA refer to any reads that mapped to them i.e., contaminants. **B.)** Upset plot outlining the intersections of miRNAs with a count of at least one in all samples within a group. The x-axis of the bar plot represents each combination of groups e.g., the first point refers to the group of miRNAs found in all groups. The second point refers to the miRNAs found in BIBI, BTBT and BTBI but not BIBT. The y-axis is the number of miRNAs found in each point on the x-axis. **C.)** PCA plot showing how all 24 samples cluster when using all expressed miRNAs. Samples are coloured by breed, where blue denotes BTBT, orange is BTBI, green is BIBT and red is BIBI. **D.)** Same as **C** but only using the top ten most variable miRNAs determined by variance. The colours are the same as in **C**. **E)** PCA plot showing clustering of samples using all expressed miRNAs. Dots represent an individual, and colours represent genetics. Blue dots represent BTBT, orange dots represent BTBI, green dots represent BIBT, and red dots represent BIBI. The x-axis represents PC1, and the y-axis represents PC2. **F)** PCA plot showing clustering of samples using only the top ten most variable miRNAs based on variance. Colours and axes are the same as in **E**.

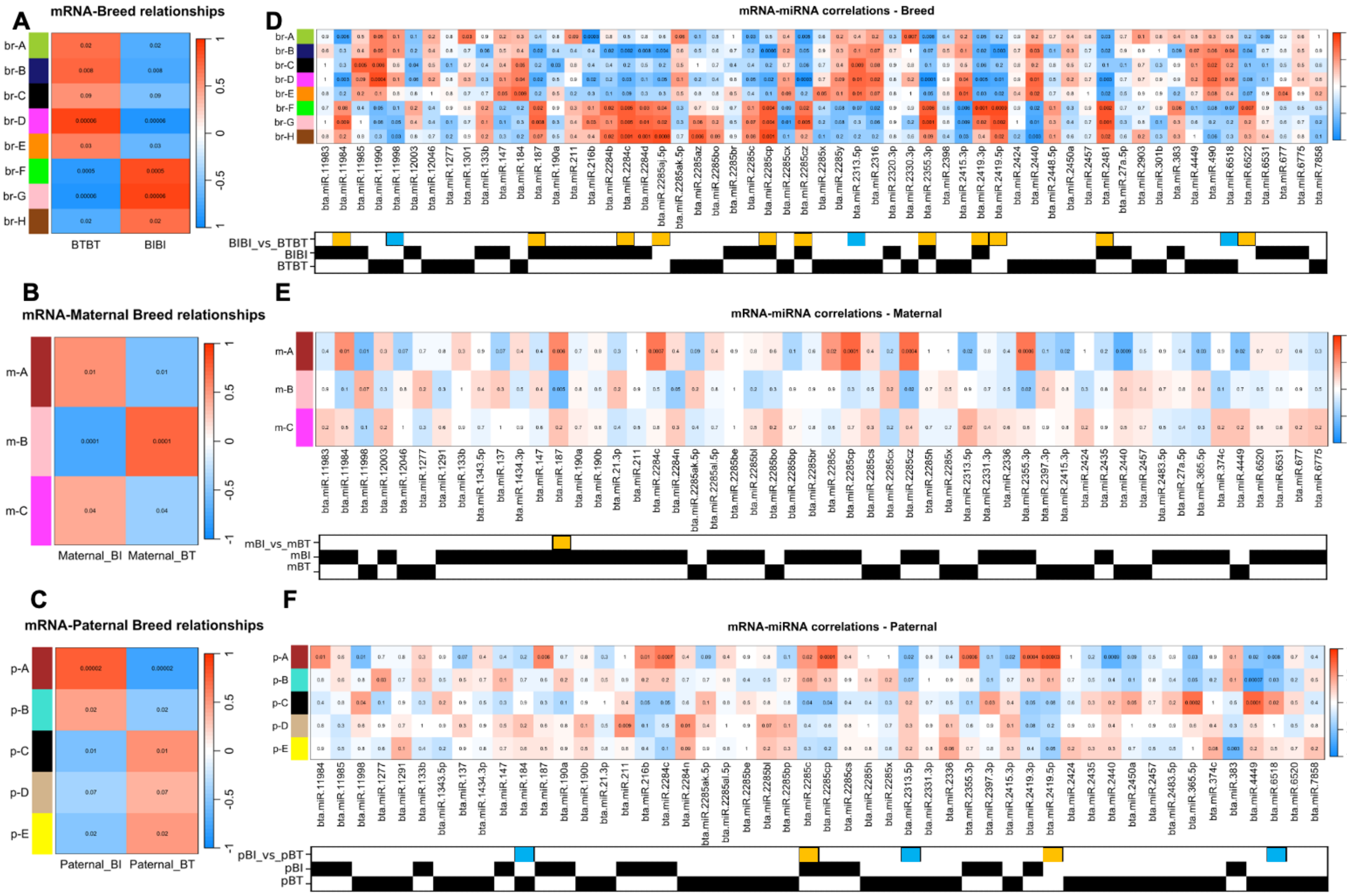
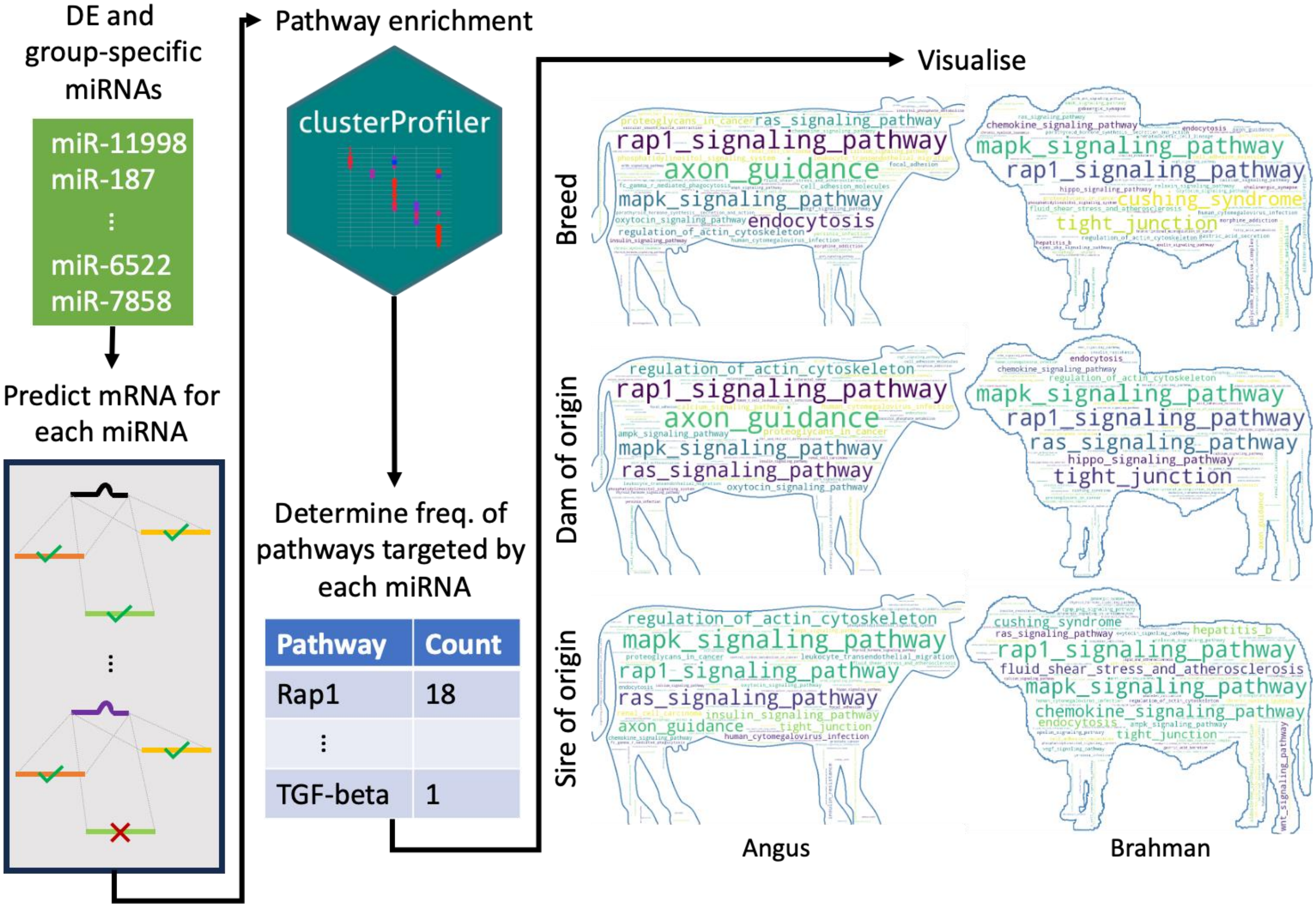

A
mRNA-Breed relationships
br-A
br-B
br-C
br-D
br-E
br-F
br-G
br-H
0.02
0.008
0.09
0.00006
0.03
0.0005
0.00006
0.02
BTBT
BIBI
1
0.5
0
-0.5
-1
B
mRNA-Maternal Breed relationships
m-A
m-B
m-C
0.01
0.0001
0.04
Maternal_BI
Maternal_BT
C
mRNA-Paternal Breed relationships
p-A
p-B
p-C
p-D
p-E
0.00002
0.02
0.01
0.07
0.02
Paternal_BI
Paternal_BT
D
mRNA-miRNA correlations - Breed
bta.miR.11983
bta.miR.11984
bta.miR.11985
bta.miR.11990
bta.miR.11998
bta.miR.12003
bta.miR.12046
bta.miR.1277
bta.miR.1301
bta.miR.133b
bta.miR.147
bta.miR.184
bta.miR.187
bta.miR.190a
bta.miR.211
bta.miR.216b
bta.miR.2284b
bta.miR.2284c
bta.miR.2284d
bta.miR.2285aj.5p
bta.miR.2285ak.5p
bta.miR.2285az
bta.miR.2285bo
bta.miR.2285br
bta.miR.2285c
bta.miR.2285cp
bta.miR.2285cx
bta.miR.2285cz
bta.miR.2285x
bta.miR.2285y
bta.miR.2313.5p
bta.miR.2316
bta.miR.2320.3p
bta.miR.2330.3p
bta.miR.2355.3p
bta.miR.2398
bta.miR.2415.3p
bta.miR.2419.3p
bta.miR.2419.5p
bta.miR.2424
bta.miR.2440
bta.miR.2448.5p
bta.miR.2450a
bta.miR.2457
bta.miR.2481
bta.miR.27a.5p
bta.miR.2903
bta.miR.301b
bta.miR.383
bta.miR.4449
bta.miR.490
bta.miR.6518
bta.miR.6522
bta.miR.6531
bta.miR.677
bta.miR.6775
bta.miR.7858
BIBI_vs_BTBT
BIBI
BTBT
E
mRNA-miRNA correlations - Maternal
bta.miR.11983
bta.miR.11984
bta.miR.11998
bta.miR.12003
bta.miR.12046
bta.miR.1277
bta.miR.1291
bta.miR.133b
bta.miR.1343.5p
bta.miR.137
bta.miR.1434.3p
bta.miR.147
bta.miR.187
bta.miR.190a
bta.miR.190b
bta.miR.21.3p
bta.miR.211
bta.miR.2284c
bta.miR.2284n
bta.miR.2285ak.5p
bta.miR.2285al.5p
bta.miR.2285be
bta.miR.2285bl
bta.miR.2285bo
bta.miR.2285bp
bta.miR.2285br
bta.miR.2285c
bta.miR.2285cp
bta.miR.2285cs
bta.miR.2285cx
bta.miR.2285cz
bta.miR.2285h
bta.miR.2285x
bta.miR.2313.5p
bta.miR.2331.3p
bta.miR.2336
bta.miR.2355.3p
bta.miR.2397.3p
bta.miR.2415.3p
bta.miR.2424
bta.miR.2435
bta.miR.2440
bta.miR.2457
bta.miR.2483.5p
bta.miR.27a.5p
bta.miR.365.5p
bta.miR.374c
bta.miR.4449
bta.miR.6520
bta.miR.6531
bta.miR.677
bta.miR.6775
mBI_vs_mBT
mBI
mBT
F
mRNA-miRNA correlations - Paternal
bta.miR.11984
bta.miR.11985
bta.miR.11998
bta.miR.1277
bta.miR.1291
bta.miR.133b
bta.miR.1343.5p
bta.miR.137
bta.miR.1434.3p
bta.miR.147
bta.miR.184
bta.miR.187
bta.miR.190a
bta.miR.190b
bta.miR.21.3p
bta.miR.211
bta.miR.216b
bta.miR.2284c
bta.miR.2284n
bta.miR.2285ak.5p
bta.miR.2285al.5p
bta.miR.2285be
bta.miR.2285bl
bta.miR.2285bp
bta.miR.2285c
bta.miR.2285cp
bta.miR.2285cs
bta.miR.2285h
bta.miR.2285x
bta.miR.2313.5p
bta.miR.2331.3p
bta.miR.2336
bta.miR.2355.3p
bta.miR.2397.3p
bta.miR.2415.3p
bta.miR.2419.3p
bta.miR.2419.5p
bta.miR.2424
bta.miR.2435
bta.miR.2440
bta.miR.2450a
bta.miR.2457
bta.miR.2483.5p
bta.miR.365.5p
bta.miR.374c
bta.miR.383
bta.miR.4449
bta.miR.6518
bta.miR.6520
bta.miR.7858
pBI_vs_pBT
pBI
pBT

Figure 2. A) Correlation heat map displaying the correlations between traits and the mRNA expression data. The x-axis refers to each phenotypic trait, i.e., breed. The y-axis refers to each of the gene modules identified by WGCNA. Each cell in the heatmap is coloured by the strength of the correlation, where a darker shade of red denotes an increasingly positive correlation, and darker shades of blue denote an increasingly negative correlation. The values in the heatmap denote the p-value associated with the correlation. **B)** Same as **A** but comparing maternal genetics. **C)** Same as **A** and **B** but comparing paternal genetics. **D)** Correlation heat map displaying the correlations between mRNA and miRNA. The x-axis denotes each of the miRNAs deemed of interest, i.e., was differentially expressed or only found in a particular group. The y-axis and correlation colour scheme are the same as **A**. The matrix underneath the heatmap denotes which comparison or group within which a given miRNA was identified. A yellow box denotes a miRNA that was upregulated in BIBI, and a blue box denotes a miRNA that was upregulated in BTBT. A black square denotes the presence of that miRNA in that group. **E)** Same as **D** but comparing miRNAs expressed in samples with different maternal genetics. A yellow box denotes miRNA that was upregulated in maternal BI. **F)** Same as **D** and **E** but comparing samples with different paternal genetics. Yellow boxes denote miRNAs upregulated in paternal BI, and blue boxes denote miRNAs upregulated in paternal BT.


DE and group-specific miRNAs
miR-11998
miR-187
⋮
miR-6522
miR-7858
Predict mRNA for each miRNA
Pathway enrichment
clusterProfiler
Determine freq. of pathways targeted by each miRNA
Visualise
Breed
Dam of origin
Sire of origin
Angus
Brahman
rap1_signaling_pathway
axon_guidance
mapk_signaling_pathway
endocytosis
ras_signaling_pathway
proteoglycans_in_cancer
regulation_of_actin_cytoskeleton
oxytocin_signaling_pathway
chemokine_signaling_pathway
cushing_syndrome
tight_junction
hippo_signaling_pathway
fluid_shear_stress_and_atherosclerosis
insulin_signaling_pathway
human_cytomegalovirus_infection
hepatitis_b
leukocyte_transendothelial_migration

Pathway	Count
Rap1	18
⋮	
TGF-beta	1

Figure 3. Schematic overview of how the most frequently targeted pathways were identified. For each DE and group-specific miRNA, we predicted their mRNA targets. We then performed pathway enrichment, where the input was a list of genes predicted to be targeted by a given miRNA. We then identified all the unique pathways predicted to be targeted by all the upregulated and expressed miRNAs for a particular group (e.g., all upregulated and expressed miRNAs in the BTBT group) and counted how many miRNAs targeted each pathway. The number of miRNAs upregulated and expressed in each group was BIBI = 29, BTBT = 28, maternal BI = 37, maternal BT = 15, paternal BI = 17 and paternal BT = 33.

Supplementary materials for Chapter 5 can be found in Appendix III.

Chapter 6: Thesis summary and future directions

Thesis summary

Understanding genetic and epigenetic regulation in Angus and Brahman cattle is vital to understanding the gene expression differences that drive their contrasting phenotypes. Critical to this is knowledge of the location of enhancers, regions of differential methylation and differences in microRNA expression, as these can contribute substantially to altered gene expression between Angus and Brahman individuals. Additionally, to understand how these phenotypic differences are first established, it is crucial to investigate them when they first appear. Research presented in this thesis has shed light on genetic and epigenetic differences between Angus and Brahman cattle. The findings reveal multiple genetic and epigenetic differences between Angus and Brahman cattle that may contribute to the contrasting phenotypes observed between these economically important breeds.

Currently, very little is known about the genomic positions of enhancers in cattle. However, a plethora of data exists describing these important genetic features in human and mouse. Since around 2012, machine learning has made substantial strides in various classification tasks, such as sentiment analysis, image classification and object detection. Owing to their ability to detect patterns not readily observed by humans, substantial interest has been placed in machine learning models and their ability to predict enhancers from DNA sequences. Additionally, new approaches have been developed to try and represent non-numerical data like the DNA sequence in a way that is interpretable by these machine learning models. Parallel to these developments, there have been efforts to use these methods to identify enhancers in lesser-studied species. In Chapter 3, evaluations of the best combination of machine learning models and DNA representations were applied to determine which offered

the best cross-species enhancer prediction performance. While all combinations of model and DNA representation performed less effectively at cross-species enhancer prediction than within-species enhancer prediction, the convolutional neural network (CNN) with a one-hot encoded DNA representation and the support vector machine with a *k*mer-proportion representation of the DNA performed the best at cross-species enhancer prediction for deep and shallow learning methods, respectively. Furthermore, their high concordance with the species-specific ChIP-seq data suggests they may have indirectly learnt DNA features associated with H3K27ac and H3K4me1 binding. Moreover, when these models were applied to the genomes of cattle, pig and dog, they predicted a similar proportion of enhancers within these genomes as what ENCODE predicts the enhancer proportion of the human genome to be, suggesting these models can identify enhancer-like elements from DNA sequence alone. Of course, these predicted enhancers would need to be experimentally validated by wet lab experiments that can identify enhancers such as cap analysis of gene expression sequencing (CAGE-seq) (Chang *et al.* 2019; Khor *et al.* 2021), clustered regularly interspaced short palindromic repeats (CRISPR) interference (CRISPRi) (Fulco *et al.* 2016), or micro-C (Hsieh *et al.* 2015; Hsieh *et al.* 2016). As more breed-specific data is generated, such as the recent additions to the cattle reference genome (Salavati *et al.* 2023), and new enhancer identification techniques are developed, these models can be further refined to improve their accuracy.

In Chapter 4, whole-genome bisulfite sequencing was used to investigate DNA methylation differences between Angus and Brahman fetal liver samples at various genomic loci, including enhancers predicted in Chapter 3. The WGBS was used to determine what, if any, DNA methylation differences existed between these two breeds at a critical developmental time point. DNA methylation was compared between the breeds (Brahman versus Angus), the maternal groups (samples with a Brahman dam versus samples with an Angus dam) and paternal groups (samples with a Brahman sire versus samples with an Angus sire). The breed comparison showed the greatest number of differentially methylated regions (DMRs) and differentially expressed genes (DEGs), followed by the paternal comparison and the maternal comparison, with the least. Minimal DMRs were observed in the promoter regions of genes, with the bulk of DMRs occurring in intergenic regions. Furthermore, most of the DMRs near genes fell within the putative enhancer regions (identified in Chapter 3) of those genes. These results suggest that differential enhancer methylation may be an important epigenomic feature driving the differences in the observed gene expression. Additionally, many of the imprinted genes investigated had a DMR within their putative enhancer region, further highlighting a possible role for enhancer methylation in gene regulatory differences between Brahman and Angus. These findings were made using the shared CpGs identified between the Brahman and Angus genomes. However, Chapter 4 also demonstrated that structural variants (SVs) and single nucleotide variants (SNVs) exist between Brahman and Angus and that these disproportionately impact CpGs, highlighting the likelihood that a combination of differential methylation of shared genomic regions between the two breeds as well as breed-specific genomic and epigenomic variants play a role in gene expression differences between these two breeds.

Chapter 5 built on the findings of Chapter 4 by investigating miRNA expression differences between the two breeds and whether this might shed light on possible causes of the contrasting phenotypes observed between Brahman and Angus.

Interestingly, it appeared that at this developmental time point, Brahman and Angus could be separated via the top ten most variable miRNAs, with the breeds being indistinguishable when all miRNAs were used. This poor clustering was unexpected, given the clustering observed in the methylation and gene expression data examined in Chapter 4. Most of the differentially expressed miRNAs (DEM) were identified in the breed comparison, with only one DEM identified between maternal groups and five DEMs between paternal groups. Interestingly, this followed the pattern observed in Chapter 4, where the breed comparison had the greatest number of DMRs and DEGs, followed by the paternal and maternal comparisons. There tended to be a greater number of group-specific miRNAs than differentially expressed miRNAs, highlighting a parallel with Chapter 4. In Chapter 4, it appeared likely that a combination of differential methylation between shared regions and breed-specific methylation impacted gene regulation. Similarly, in Chapter 5, the greater number of breed-specific miRNAs than DEMs also suggests that a combination of DE and breed-specific miRNAs regulates gene expression differences. Weighted correlation network analysis was used to identify gene modules that were positively and negatively correlated with each group and to identify miRNAs that may be regulating those modules. Unsurprisingly, gene modules that were positively correlated with Angus samples were negatively correlated with Brahman samples and vice versa. KEGG pathway analysis revealed that genes related to the glutamatergic synapse were significantly correlated with Brahman cattle. Similarly, KEGG analysis revealed that genes associated with adipogenesis were significantly correlated with Angus cattle. Given the role of the glutamatergic synapse in coordinating the body’s response to hyperthermia and the superior meat quality of Angus cattle, the miRNAs that were significantly correlated with genes associated with glutamatergic synapse and adipogenesis are promising candidates to investigate the origin of these traits. Finally, many DE and breed-specific miRNAs targeted the same pathways in Brahman and Angus cattle, suggesting that differential modulation of these pathways between the breeds may be causing the observed phenotypic differences between Brahman and Angus. This differential regulation via miRNAs may arise from SNVs in the three prime UTRs of the mRNA or within the miRNA.

The results presented in this thesis were obtained via a range of machine-learning, genomic and epigenomic analyses, all aimed at improving our understanding of genetic and epigenetic regulation in Brahman and Angus cattle and how differences between the two breeds may inform their phenotypic differences. This analysis has uncovered various candidate genetic and epigenetic features that may influence breed-specific traits in Brahman and Angus cattle. In summary, we have shown that: (1) machine-learning models can be used to predict enhancers across species and that this can be used to inform subsequence epigenomic analysis. (2) differential methylation of potential enhancers is likely more important than differential methylation of promoters in regulating gene expression in the liver of Angus and Brahman fetuses at gestational day 153. (3) the genetics of the sire (i.e., whether it is Angus or Brahman) likely influences how imprinted genes like *Dgat1* are methylated. (4) breed-specific gene expression patterns are associated with breed-specific traits and are likely under miRNA control. (5) although the impacts of SVs and SNVs on methylation and miRNA binding were unclear, they will form the basis of future studies as we transition to the pangenome era, where variants between genomes can be better accounted for during analyses. In conclusion, the work presented in this thesis adds to the body of knowledge of genetic and epigenetic regulation in Brahman and Angus cattle by providing possible mechanisms for the establishment of the contrasting phenotypes observed between these breeds.

Future directions

Results from this thesis highlight candidate regions and pathways that may contribute to understanding how the contrasting phenotypes between Brahman and Angus are established. Despite this, the exact contribution of each of these components remains unclear. The work presented here has demonstrated the potential for enhancer prediction tools with genomic and epigenomic assays to improve our understanding of how certain traits in cattle come about. As a product of this study, key questions remain to be answered, which could provide the basis for future work.

Can further improvements be made in identifying cattle enhancers?

For all the advantages of working with livestock, such as being able to control mating, there are substantial downsides as well. The most severe of which is the degree of funding available for basic science research on livestock. Given that livestock research will likely never see the same level of investment that human medical research receives, the question remains as to how we can further improve our understanding of cis-regulatory elements like enhancers in livestock species like cattle. This thesis examined the role of machine learning in predicting enhancers in under-studied species; however, with the speed with which the machine learning field moves, in addition to new sequencing data being generated, further improvements can likely be made in this area.

Much attention has been given to deep learning models, specifically image classifiers, and their applicability to genomics. For example, DeepVariant has been shown to outperform GATK in calling genomic variants (Poplin *et al.* 2018). DeepVariant is a convolutional neural network (CNN) based model that converts variants to image-like input to determine variant type. The image-like input consists of the bases, quality scores and additional read features, where each of these inputs represents a colour channel in an image, i.e., red, green or blue. As more epigenomic sequencing data becomes available for a variety of cattle breeds and tissues, such as assay of transposase-accessible chromatin sequencing (ATAC-seq), DNase I hypersensitivity sequencing (DNase-seq) and chromatin immunoprecipitation sequencing (ChIP-seq), there will be an opportunity to develop models that do not solely rely on the DNA sequence. Using the Hilbert Curve representation from Chapter 3, researchers will be able to generate "images" of DNA regions from cattle where one colour channel is the sequence, another colour channel is the ATAC-seq signal or DNase-seq signal, and the third channel is the H3K27ac ChIP-seq signal, for example. The integration of DNA sequence and DNase-seq signal has been used successfully with human data (Yin *et al.* 2019) but has not previously been feasible in cattle owing to the lack of matched ATAC-seq and ChIP-seq datasets. As more are generated as part of the various livestock consortia, this will become a more viable option.

Alternatively, different assay approaches may provide new avenues to pursue to improve our understanding of the cattle genome. Enhancer RNAs (eRNAs) are a recently discovered class of non-coding RNAs that are transcribed from the DNA sequence of enhancers in a tissue-specific manner (Sartorelli & Lauberth 2020). Due to the nascent nature of eRNAs, they can only be captured by more specialised sequencing techniques like global run-on sequencing (GRO-seq) (Step *et al.* 2014; Liang *et al.* 2016; Kim *et al.* 2018; Pan *et al.* 2021), precision run-on sequencing (PRO-seq) (Levandowski *et al.* 2021; Xu *et al.* 2022; Bae *et al.* 2023; Caligiuri *et al.* 2023) and CAGE-seq (Chang *et al.* 2019; Khor *et al.* 2021). CAGE-seq data has recently been generated in cattle to improve the annotation of the cattle genome as part of the BovReg consortium (Salavati *et al.* 2023). As more CAGE-seq and similar data are generated for cattle, it would be interesting to evaluate the performance of enhancer prediction models trained on CAGE-seq data.

Would chromatin conformation capture technologies be a better use of money as they enable the inference of the enhancer target and location?

Parallel to the deep learning and cattle epigenomic dataset developments, chromatin conformation capture (3C) based techniques have become increasingly popular, in particular, micro-C (Hsieh *et al.* 2015; Hsieh *et al.* 2016; Krietenstein *et al.* 2020), micro-ChIP (Mumbach *et al.* 2016; Mumbach *et al.* 2017) and pore-C (Deshpande *et al.* 2022). The appeal of these techniques is that they provide insight into the 3D structure of the genome and, with sufficient resolution, can reveal enhancer-promoter contacts, meaning one can gain insight into the location of enhancers and their target genes. The cost of these technologies is still high and may be prohibitive in many contexts for cattle. However, Hi-ChIP offers a possible middle ground. This assay combines ChIP-seq with 3C technologies, and so by using an enhancer-associated histone modification like H3K27ac, one can enrich the 3C library with DNA fragments bound to this histone mark (Mumbach *et al.* 2016; Mumbach *et al.* 2017), giving insight into DNA regions interacting with potential enhancers at a fraction of the cost of sequencing the entire genome. Of course, the challenge of choosing what tissue and time point to generate the sequencing data remains. As sequencing costs decrease, these 3C-based techniques will arguably provide the best value, even of CAGE-seq, as one can get the location and target of enhancers in one assay instead of just the location.

Can long-read sequencing technologies be used to decipher which parent is contributing DNA methylation and gene expression?

Parent-of-origin effects (POEs) are known to play a role in cattle, with maternally inherited genes disproportionately influencing muscle development in reciprocal crosses (Xiang *et al.* 2013). It is less clear, however, how this differs depending on what breed the dam is, what breed the sire is and whether there is an association between traits and which parent contributes to the DNA methylation and gene expression pattern in the reciprocal cross. To address this, recording various phenotypic traits in the reciprocal crosses and performing Iso-seq and nanopore sequencing of the relevant tissues would be interesting to determine the association between different traits and the parents' genetics. By using these long-read sequencing techniques, it is possible to determine whether any genes show preferential expression from either the maternal or paternal chromosomes. Similarly, by capturing DNA methylation data from the nanopore sequencing, one can investigate whether the preferential expression of the maternal allele is due to higher methylation of the paternal allele, for example. Knowledge of any differential transcript usage would be of particular interest in breeding hybrid progeny, where the aim is to maximise the desirable traits of both breeds in one animal. For example, if there are significant differences in post-birth growth rate between hybrid progeny from an Angus dam compared to a Brahman dam, with Angus dams birthing smaller calves that then have accelerated post-birth weight gain, Iso-seq and nanopore-sequencing enables one to tease apart the possible mechanisms causing this; causative mechanisms could be differential methylation, structural variants, preferential expression of parental allele, differential transcript usage or a combination of all of them. This information can then inform breeding programmes so that only Angus dams with those mechanisms are used for breeding.

Determining whether it is a maternal or paternal copy of a gene being expressed and its association with DNA methylation will also enable a better understanding of imprinting in cattle. Much of what we know about imprinted genes comes from human and mouse studies. However, there is evidence to suggest that imprinting is not well-conserved among mammals (Monk *et al.* 2006; Khatib *et al.* 2007). Given the uncertainty around imprinting conservation, it is likely that cattle researchers cannot solely rely on human and mouse studies to inform on imprinting in cattle. As a result, long-read sequencing will provide a valuable tool to help us investigate imprinting in a wide array of species.

What impacts do SVs and SNVs have on genomic and epigenomic differences between the breeds?

The advent of long-read sequencing technologies has enabled us, for the first time, to sequence highly repetitive regions of the genome, resulting in the first telomere-to-telomere (T2T) genome assemblies (Nurk *et al.* 2022; Chen *et al.* 2023; Wang *et al.* 2023). These highly accurate reference genomes can then form the backbone of pangenome graphs, which capture a high degree of within-species diversity. These pangenome graphs will be crucial to enabling accurate breed comparisons, like Brahman and Angus. As Chapter 4 demonstrated, SNVs and SVs disproportionately affected CpGs between Brahman and Angus. By aligning to a pangenome graph that accounts for these variants, we can better understand how genetic and epigenetic differences between the breeds might contribute to gene regulation. For example, analysis of DNA methylation data with a pangenome graph may reveal a Brahman-specific insertion that introduces a CpG island near a gene implicated in conveying heat tolerance.

Analysing miRNA data with pangenome graphs may reveal SNVs within the seed region of miRNAs that alter the target specificity of that miRNA. Similarly, SNVs in the 3'UTR of genes may change the binding strength of miRNAs. Regardless of whether the SNV occurs in the miRNA seed region or the 3'UTR, both have the potential to alter miRNA-mRNA interactions, which can cause changes in how genes are regulated between Brahman and Angus cattle. Moreover, pangenome graphs have the potential to simplify analyses as all genomic features will be on the same coordinate system, enabling accurate comparisons across breeds to be made.

Does more need to be done to apply this knowledge to breeding programmes?

The main driver of much of the research in cattle is how this understanding of the underlying biology can be used to improve breeding programmes. Examining first the enhancer predictions, these have the potential to help prioritise SNVs within a breeding programme. By prioritising SNVs, one can reduce the number of SNVs that need to be examined for genomic prediction. The use of regulatory elements to prioritise SNVs has been previously investigated in cattle (Xiang *et al.* 2019). However, there is still a need for breed-specific identification of cis-regulatory elements, like enhancers, as there are likely important breed-specific differences in these elements. As more breed-specific data are generated, we can develop breed-specific prioritisation of SNVs that will further improve the accuracy of these genomic predictions.

Investigating DNA methylation differences between Brahman and Angus during fetal development is an important step in understanding how differential methylation may contribute to breed differences during the development of muscle and fat, essential tissues for meat production. While this is important for understanding how fetal programming may improve the productivity of the progeny, it may be too invasive and costly for producers. With this in mind, collecting methylation data from muscle and fat tissues in cattle around the age of slaughter would be worthwhile. The reason is that several studies have identified DNA methylation quantitative trait loci (meQTLs) associated with phenotypes, such as tender and tough meat (de Souza *et al.* 2022), milk fat content (Wang *et al.* 2021) and muscle growth (Van Laere *et al.* 2003). By capturing DNA methylation profiles when carcass traits are recorded, one can gain an insight into which SNVs are affected by DNA methylation. This information can then be used to prioritise SNVs, thus reducing the number of loci in the genomic prediction model and improving prediction.

Finally, the use of miRNAs to improve genomic prediction in cattle could be utilised in a similar way as enhancers and DNA methylation have been proposed above. SNVs that occur in miRNA seed regions or 3'UTRs of important genes could help reduce the number of loci needed in genomic prediction. This miRNA-SNV prioritisation has been shown to be effective in Holstein, Jersey and Nordic red cattle in prioritising SNVs associated with milk production and mastitis traits (Fang *et al*. 2018), but further work is needed to evaluate the usefulness of this in other genomic prediction contexts.

References

Bae S., Kim K., Kang K.S., Kim H., Lee M.J., Oh B., Kaneko K., Ma S.K., Choi J.H., Kwak H., Lee E.Y., Park S.H. & Park-Min K.H. (2023) RANKL-responsive epigenetic mechanism reprograms macrophages into bone-resorbing osteoclasts. Cellular & Molecular Immunology 20, 94-109. 10.1038/s41423-022-00959-x

Caligiuri M., Williams G.L., Castro J., Battalagine L., Wilker E., Yao L.L., Schiller S., Toms A., Li P., Pardo E., Graves B., Azofeifa J., Chicas A., Herbertz T., Lai M.R., Basken J., Wood K.W., Xu Q.L. & Guichard S.M. (2023) FT-6876, a Potent and Selective Inhibitor of CBP/p300, is Active in Preclinical Models of Androgen Receptor-Positive Breast Cancer. Targeted Oncology 18, 269-85. 10.1007/s11523-023-00949-7

Chang H.C., Huang H.C., Juan H.F. & Hsu C.L. (2019) Investigating the role of super-enhancer RNAs underlying embryonic stem cell differentiation. BMC Genomics 20. 10.1186/s12864-019-6293-x

Chen J., Wang Z., Tan K., Huang W., Shi J., Li T., Hu J., Wang K., Wang C., Xin B., Zhao H., Song W., Hufford M.B., Schnable J.C., Jin W. & Lai J. (2023) A complete telomere-to-telomere assembly of the maize genome. Nature Genetics 55, 1221-31. 10.1038/s41588-023-01419-6

de Souza M.M., Niciura S.C.M., Rocha M.I.P., Pan Z., Zhou H., Bruscadin J.J., da Silva Diniz W.J., Afonso J., de Oliveira P.S.N., Mourao G.B., Zerlotini A., Coutinho L.L., Koltes J.E. & de Almeida Regitano L.C. (2022) DNA methylation may affect beef tenderness through signal transduction in Bos indicus. Epigenetics Chromatin 15, 15. 10.1186/s13072-022-00449-4

Deshpande A.S., Ulahannan N., Pendleton M., Dai X., Ly L., Behr J.M., Schwenk S., Liao W., Augello M.A., Tyer C., Rughani P., Kudman S., Tian H., Otis H.G., Adney E., Wilkes D., Mosquera J.M., Barbieri C.E., Melnick A., Stoddart D., Turner D.J., Juul S., Harrington E. & Imieliński M. (2022) Identifying synergistic high-order 3D chromatin conformations from genome-scale nanopore concatemer sequencing. Nature Biotechnology 40, 1488-99. 10.1038/s41587-022-01289-z

Fang L., Sørensen P., Sahana G., Panitz F., Su G., Zhang S., Yu Y., Li B., Ma L., Liu G., Lund M.S. & Thomsen B. (2018) MicroRNA-guided prioritization of genome-wide association signals reveals the importance of microRNA-target gene networks for complex traits in cattle. Sci Rep 8, 9345. 10.1038/s41598-018-27729-y

Fulco C.P., Munschauer M., Anyoha R., Munson G., Grossman S.R., Perez E.M., Kane M., Cleary B., Lander E.S. & Engreitz J.M. (2016) Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science 354, 769-73. doi:10.1126/science.aag2445

Hsieh T.-H.S., Fudenberg G., Goloborodko A. & Rando O.J. (2016) Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. Nature Methods 13, 1009-11. 10.1038/nmeth.4025

Hsieh T.H., Weiner A., Lajoie B., Dekker J., Friedman N. & Rando O.J. (2015) Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell 162, 108-19. 10.1016/j.cell.2015.05.048

Khatib H., Zaitoun I. & Kim E.-S. (2007) Comparative analysis of sequence characteristics of imprinted genes in human, mouse, and cattle. Mammalian Genome 18, 538-47.

Khor J.M., Guerrero-Santoro J., Douglas W. & Ettensohn C.A. (2021) Global patterns of enhancer activity during sea urchin embryogenesis assessed by eRNA profiling. Genome Research 31, 1680-92. 10.1101/gr.275684.121

Kim Y.J., Xie P., Cao L., Zhang M.Q. & Kim T.H. (2018) Global transcriptional activity dynamics reveal functional enhancer RNAs. Genome Research 28, 1799-811. 10.1101/gr.233486.117

Krietenstein N., Abraham S., Venev S.V., Abdennur N., Gibcus J., Hsieh T.S., Parsi K.M., Yang L., Maehr R., Mirny L.A., Dekker J. & Rando O.J. (2020) Ultrastructural Details of Mammalian Chromosome Architecture. Molecular Cell 78, 554-65 e7. 10.1016/j.molcel.2020.03.003

Levandowski C.B., Jones T., Gruca M., Ramamoorthy S., Dowell R.D. & Taatjes D.J. (2021) The Delta 40p53 isoform inhibits p53-dependent eRNA transcription and enables regulation by signal-specific transcription factors during p53 activation. PLoS Biology 19. 10.1371/journal.pbio.3001364

Liang J., Zhou H., Gerdt C., Tan M., Colson T., Kaye K.M., Kieff E. & Zhao B. (2016) Epstein-Barr virus super-enhancer eRNAs are essential for MYC oncogene expression and lymphoblast proliferation. Proceedings of the National Academy of Sciences of the United States of America 113, 14121-6. 10.1073/pnas.1616697113

Monk D., Arnaud P., Apostolidou S., Hills F., Kelsey G., Stanier P., Feil R. & Moore G. (2006) Limited evolutionary conservation of imprinting in the human placenta. Proceedings of the National Academy of Sciences 103, 6623-8.

Mumbach M.R., Rubin A.J., Flynn R.A., Dai C., Khavari P.A., Greenleaf W.J. & Chang H.Y. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nature Methods 13, 919-22. 10.1038/nmeth.3999

Mumbach M.R., Satpathy A.T., Boyle E.A., Dai C., Gowen B.G., Cho S.W., Nguyen M.L., Rubin A.J., Granja J.M., Kazane K.R., Wei Y., Nguyen T., Greenside P.G., Corces M.R., Tycko J., Simeonov D.R., Suliman N., Li R., Xu J., Flynn R.A., Kundaje A., Khavari P.A., Marson A., Corn J.E., Quertermous T., Greenleaf W.J. & Chang H.Y. (2017) Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nature Genetics 49, 1602-12. 10.1038/ng.3963

Nurk S., Koren S., Rhie A., Rautiainen M., Bzikadze A.V., Mikheenko A., Vollger M.R., Altemose N., Uralsky L., Gershman A., Aganezov S., Hoyt S.J., Diekhans M., Logsdon G.A., Alonge M., Antonarakis S.E., Borchers M., Bouffard G.G., Brooks S.Y., Caldas G.V., Chen N.-C., Cheng H., Chin C.-S., Chow W., de Lima L.G., Dishuck P.C., Durbin R., Dvorkina T., Fiddes I.T., Formenti G., Fulton R.S., Fungtammasan A., Garrison E., Grady P.G.S., Graves-Lindsay T.A., Hall I.M., Hansen N.F., Hartley G.A., Haukness M., Howe K., Hunkapiller M.W., Jain C., Jain M., Jarvis E.D., Kerpedjiev P., Kirsche M., Kolmogorov M., Korlach J., Kremitzki M., Li H., Maduro V.V., Marschall T., McCartney A.M., McDaniel J., Miller D.E., Mullikin J.C., Myers E.W., Olson N.D., Paten B., Peluso P., Pevzner P.A., Porubsky D., Potapova T., Rogaev E.I., Rosenfeld J.A., Salzberg S.L., Schneider V.A., Sedlazeck F.J., Shafin K., Shew C.J., Shumate A., Sims Y., Smit A.F.A., Soto D.C., Sović I., Storer J.M., Streets A., Sullivan B.A., Thibaud-Nissen F., Torrance J., Wagner J., Walenz B.P., Wenger A., Wood J.M.D., Xiao C., Yan S.M., Young A.C., Zarate S., Surti U., McCoy R.C., Dennis M.Y., Alexandrov I.A., Gerton J.L., O'Neill R.J., Timp W., Zook J.M., Schatz M.C., Eichler E.E., Miga K.H. & Phillippy A.M. (2022) The complete sequence of a human genome. Science 376, 44-53. doi:10.1126/science.abj6987

Pan C.W., Wen S.M., Chen L., Wei Y.L., Niu Y.J. & Zhao Y. (2021) Functional roles of antisense enhancer RNA for promoting prostate cancer progression. Theranostics 11, 1780-94. 10.7150/thno.51931

Poplin R., Chang P.-C., Alexander D., Schwartz S., Colthurst T., Ku A., Newburger D., Dijamco J., Nguyen N., Afshar P.T., Gross S.S., Dorfman L., McLean C.Y. & DePristo M.A. (2018) A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology 36, 983-7. 10.1038/nbt.4235

Salavati M., Clark R., Becker D., Kuehn C., Plastow G., Dupont S., Moreira G.C.M., Charlier C., Clark E.L. & BogReg C. (2023) Improving the annotation of the cattle genome by annotating transcription start sites in a diverse set of tissues and populations using Cap Analysis Gene Expression sequencing. G3-Genes Genomes Genetics 13. 10.1093/g3journal/jkad108

Sartorelli V. & Lauberth S.M. (2020) Enhancer RNAs are an important regulatory layer of the epigenome. Nature Structural & Molecular Biology 27, 521-8. 10.1038/s41594-020-0446-0

Step S.E., Lim H.W., Marinis J.M., Prokesch A., Steger D.J., You S.H., Won K.J. & Lazar M.A. (2014) Anti-diabetic rosiglitazone remodels the adipocyte transcriptome by redistributing transcription to PPAR gamma-driven enhancers. Genes & Development 28, 1018-28. 10.1101/gad.237628.114

Van Laere A.-S., Nguyen M., Braunschweig M., Nezer C., Collette C., Moreau L., Archibald A.L., Haley C.S., Buys N., Tally M., Andersson G., Georges M. & Andersson L. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. Nature 425, 832-6. 10.1038/nature02064

Wang M., Bissonnette N., Dudemaine P.L., Zhao X. & Ibeagha-Awemu E.M. (2021) Whole Genome DNA Methylation Variations in Mammary Gland Tissues from Holstein Cattle Producing Milk with Various Fat and Protein Contents. Genes (Basel) 12. 10.3390/genes12111727

Wang Y.-H., Liu P.-Z., Liu H., Zhang R.-R., Liang Y., Xu Z.-S., Li X.-J., Luo Q., Tan G.-F., Wang G.-L. & Xiong A.-S. (2023) Telomere-to-telomere carrot (Daucus carota) genome assembly reveals carotenoid characteristics. Horticulture Research 10. 10.1093/hr/uhad103

Xiang R., Berg I.V.D., MacLeod I.M., Hayes B.J., Prowse-Wilkins C.P., Wang M., Bolormaa S., Liu Z., Rochfort S.J., Reich C.M., Mason B.A., Vander Jagt C.J., Daetwyler H.D., Lund M.S., Chamberlain A.J. & Goddard M.E. (2019) Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proc Natl Acad Sci U S A 116, 19398-408. 10.1073/pnas.1904159116

Xiang R., Ghanipoor-Samami M., Johns W.H., Eindorf T., Rutley D.L., Kruk Z.A., Fitzsimmons C.J., Thomsen D.A., Roberts C.T., Burns B.M., Anderson G.I., Greenwood P.L. & Hiendleder S. (2013) Maternal and Paternal Genomes Differentially Affect Myofibre Characteristics and Muscle Weights of Bovine Fetuses at Midgestation. PLOS ONE 8, e53402. 10.1371/journal.pone.0053402

Xu W.Q., He C.X., Kaye E.G., Li J.H., Mu M.D., Nelson G.M., Dong L., Wang J.H., Wu F.Z., Shi Y.G., Adelman K., Lan F., Shi Y. & Shen H.J. (2022) Dynamic control of chromatin-associated m(6)A methylation regulates nascent RNA synthesis. Molecular Cell 82, 1156-+. 10.1016/j.molcel.2022.02.006

Yin Q., Wu M., Liu Q., Lv H. & Jiang R. (2019) DeepHistone: a deep learning approach to predicting histone modifications. BMC Genomics 20, 193. 10.1186/s12864-019-5489-4

Appendix I

Supplementary materials for Chapter 3 are available online at https://doi.org/10.1016/j.ygeno.2022.110454

Appendix II

Due to the size of the Supplementary information for Chapter 4, it has been made available online at: https://github.com/cmacphillamy/Thesis_supplementary_information/tree/main/Chapter_4_supplementary

Appendix III

Supplementary information for Chapter 5 is available online at: https://github.com/cmacphillamy/Thesis_supplementary_information/tree/main/Chapter_5_supplementary