UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA EVOLUTIVA E BIOLOGIA MOLECULAR

Ma. Jennifer Jessica Bruscadin

“Uso da análise da expressão alelo-específica e do efeito de origem parental para a predição de variantes reguladoras de fenótipos relacionados ao músculo de bovinos *Bos indicus*”

São Carlos–SP
2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS BIOLÓGICAS E DA SAÚDE
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA EVOLUTIVA E BIOLOGIA MOLECULAR

Ma. Jennifer Jessica Bruscadin

“Uso da análise da expressão alelo-específica e do efeito de origem parental para a predição de variantes reguladoras de fenótipos relacionados ao músculo de bovinos *Bos indicus*”

Tese apresentada ao Programa de Pós-Graduação em Genética Evolutiva e Biologia Molecular do Centro de Ciências Biológicas e da Saúde da Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Doutor em Ciências, área de concentração: Genética e Evolução.

Orientadora: Dr.ª Luciana Correia de Almeida Regitano

Coorientador: Dr. Wellison Jarles da Silva Diniz

São Carlos
2025

I dedicate this work to my parents, Luiz and Luzia, and my husband William for their constant encouragement and unwavering support

ACKNOWLEDGEMENTS

All my gratitude to God for all the opportunities that were granted to me, for the strength in moments of weakness and encouragement in difficult times.

I would like to thank Dr. Luciana Correia de Almeida Regitano for allowing me to participate in the research group, for her assistance, and for her counsel regarding both my personal and academic lives.

Additionally, I would like to express my gratitude to Dr. Wellison Jarles da Silva Diniz for his guidance and assistance in developing this work.

I am grateful to Embrapa Pecuária Sudeste for providing the necessary infrastructure for the execution of this work, as well as to its staff for their always cordial and collaborative support, especially to Dr. Adriana Mércia Guaratini Ibelli for her care with the Animal Biotechnology Laboratory and our group's activities.

I also acknowledge Embrapa Informática Agropecuária for granting access to the server used for the analyses, as well as for their timely assistance in resolving problems and offering advice on how to increase the analyses' effectiveness.

I would like to express my gratitude to the Federal University of São Carlos and the Graduate Program in Evolutionary Genetics and Molecular Biology for their support, encouragement of scientific dissemination, and the courses that provided crucial knowledge for the completion of this project.

This study was financed, in part, by the São Paulo Research Foundation (FAPESP), Brazil. Process Number #2020/01369-1.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001.

I am grateful to the Coordination for the Improvement of Higher Education Personnel (CAPES) and the São Paulo Research Foundation (FAPESP) for awarding the doctoral scholarships. I'm also thankful to FAPESP for granting the Research Internship Abroad Scholarship.

I also thank FAPESP and CNPq for the financial support essential for obtaining all the data used in this work.

I am grateful to my colleagues at the University of California, Davis (UC Davis) for welcoming me to their institution, offering all the support and project opportunities, especially Dr. Huaijun Zhou and Dr. Dailu Guan.

My heartfelt thanks go to Dr. Tainã Figueiredo Cardoso and Dr. Juliana Afonso for their academic support and friendship since the start of my master's studies, as well as to all my research group colleagues for the ideas, collaborations, and companionship.

I am deeply grateful to my parents, Luiz Fernando Bruscadin and Luzia de Fátima Jacometo Bruscadin, and to my husband, William Uliana Albarice Bittencourt, who, despite my frequent absences due to the studies, offered unconditional emotional support, encouragement and understanding in the most challenging moments.

I also thank my daughter, Lavínia Bruscadin Bittencourt, for giving me so much strength and inspiring me even before her birth.

“Sonho que se sonha só é só um sonho;
Mas sonho que se sonha junto é realidade.”

Raul Seixas (Prelúdio)

"Pass on what you have learned. Strength.
Mastery. But weakness, folly, failure also"

Mestre Yoda (Star Wars: Episódio VI - O Retorno de Jedi)

RESUMO

O tecido muscular de animais da raça Nelore (*Bos indicus*) é alvo de estudos que visam o melhoramento de características de qualidade de carne, que afetam a aceitação do consumidor. Assim, é de grande importância a investigação de mecanismos de regulação que impactam a expressão de genes funcionalmente relacionados a tais fenótipos. Análises de expressão alelo-específica (ASE, do inglês *allele-specific expression*) são ferramentas poderosas na identificação de regiões sob efeito de regulação diferencial entre dois alelos pois evidenciam a existência de variantes em *cis* capazes de regular a transcrição em cada haplótipo. Se a expressão alélica depende da origem parental do alelo, ocorre o *imprinting* genômico. Mecanismos epigenéticos, como a metilação do DNA, e *cis*-reguladores, como a ligação de fatores de transcrição (TFs) e de micro RNAs (miRNAs), podem ser responsáveis por esses padrões de expressão quando o SNP regulador estiver em heterozigose. Assim, a fim de analisar a implicação da ASE sobre a regulação gênica e variação fenotípica em músculo de animais Nelore, esse estudo identificou SNPs localizados em regiões de ASE (ASE SNPs) em todo o genoma, apontando quais podem ser associados a fenótipos de qualidade de carne, e predizendo as respectivas variantes potencialmente reguladoras. Além disso, através de análises de metilação alelo-específica (ASM, do inglês *allele-specific methylation*) e de associação de SNPs com o efeito de origem parental sobre fenótipos (PO-QTLs, do inglês *parent-of-origin quantitative trait loci*) sugerimos potenciais candidatos ao *imprinting* genômico em bovinos e os potenciais impactos nos fenótipos de interesse. A partir desse projeto foi identificado o maior conjunto de ASE SNPs em músculo bovino, com 38.177 ASE SNPs, e 24.650 variantes *cis*-reguladoras (entre aseQTLs, associadas à ASE e *cis*-eQTLs, associadas à transcrição) putativas capazes de regular tais padrões de expressão. A variação de 55 fenótipos de qualidade de carne foi associada a um total de 1.479 ASE SNPs (então chamados DASE SNPs, do inglês *differential ASE SNPs*) e a 1.817 PO-QTLs. Foram identificadas 8.074 regiões com ASM, apontando novos genes candidatos ao *imprinting* genômico para futura validação. Foram hipotetizados os mecanismos de regulação e os processos biológicos envolvidos com os nossos achados. Tais resultados contribuíram para o aumento do conhecimento acerca da regulação gênica em músculo bovino, propondo regiões genômicas candidatas à validação e aplicação em programas de seleção genômica que visam o melhoramento da qualidade da carne de animais Nelore.

Palavras-chave: expressão alelo-específica, melhoramento animal, genômica, zebu, epigenética

ABSTRACT

The muscle tissue of Nelore cattle (*Bos indicus*) is the focus of studies aimed at improving meat quality traits that influence consumer acceptance. Therefore, investigating regulatory mechanisms that affect the expression of genes functionally related to these phenotypes is of great importance. Allele-specific expression (ASE) analyses are powerful tools for identifying regions under differential regulatory effects between two alleles, as they highlight the presence of *cis*-regulatory variants impacting transcription on each haplotype. If the allelic expression depends on the parental origin of the allele, genomic imprinting occurs. Epigenetic and *cis*-regulatory mechanisms can be responsible for these expression patterns, such as the binding of transcription factors (TFs), microRNAs (miRNAs), and epigenetic mechanisms like DNA methylation when the regulatory SNP is heterozygous. Thus, to analyze the implications of ASE on gene regulation and phenotypic variation in Nelore cattle muscle, this study identified SNPs located in ASE regions (ASE SNPs) across the genome, pointing out those potentially associated with meat quality traits and identifying putative regulatory variants. Furthermore, through analyses of allele-specific methylation (ASM) and association of SNPs with parental origin effects on phenotypes (PO-QTLs, parent-of-origin quantitative trait *loci*), we suggest potential candidates for genomic imprinting in cattle and their possible impacts on traits of interest. This project identified the largest set of ASE SNPs in bovine muscle to date, comprising 38.177 ASE SNPs, and 24.650 putative *cis*-regulatory variants (aseQTLs, associated with the ASE and *cis*-eQTLs, associated to the transcription) able to regulate such expression patterns. The variation in 55 meat quality phenotypes was associated with a total of 1.479 ASE SNPs (referred to as DASE SNPs, *differential* ASE SNPs) and with 1.817 PO-QTLs. Additionally, 8.074 ASM regions were identified, highlighting new candidate genes for genomic imprinting for future validation. The regulatory mechanisms and biological processes involved in our findings were hypothesized. These results contribute to advancing knowledge about gene regulation in bovine muscle, proposing genomic candidate regions for validation and application in genomic selection programs aimed at improving the meat quality of Nelore cattle.

Keywords: allele-specific expression, animal breeding, genomics, zebu, epigenetics

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

LISTA DE ABREVIATURAS

- a* or AFAT - Redness Color Of Fat
- AEI – Allelic Expression Imbalance
- ASE – Allele-Specific Expression
- ASE GENE – Gene Containing An ASE SNP
- ASE SNP – Single Nucleotide Polymorphism Marking A Region With Allele-Specific Expression
- AseQTL – Allele-Specific Expression Quantitative Trait Loci
- ASM – Allele-Specific Methylation
- b* Or BFAT - Yellowness Color Fat
- BFT – Backfat Thickness
- BP – Biological Process
- *Cis*-eQTL– Cis-Acting Expression Quantitative Trait Loci
- CNV – Copy Number Variations
- DASE – Differential Allele-Specific Expression
- DASE GENE – Gene Containing An DASE SNP
- DASE SNP – Single Nucleotide Polymorphism Marking A Region With Differential Allele-Specific Expression
- DEG – Differentially Expressed Gene
- DM – Differentially Methylated
- eQTL – Expression Quantitative Trait Loci
- FA – Fatty Acids
- FAANG – Functional Annotation Of Animal Genomes
- FC – Fold Change
- FDR – False Discovery Rate
- GEBV – Genomic Estimated Breeding Value
- GO – Gene Ontology
- GSEA – Gene Set Enrichment Analysis
- GWAS – Genome-Wide Association Study
- HDL – High-Density Lipoprotein
- IMF – Intramuscular Fat
- LD – Linkage Disequilibrium
- LDL – Low-Density Lipoprotein
- LFAT – Luminosity Index Of Fat
- MAF – Minor Allele Frequency
- MUFA – Monounsaturated Fatty Acids
- N3 – Omega 3 Fatty Acid
- N6 – Omega 6 Fatty Acid
- ORA – Overrepresentation Analysis
- PO – Parent-Of-Origin
- POE – Parent-Of-Origin Effect

- PO-GWAS – Parent-Of-Origin Genome-Wide Association Study
- PO-QTL – Parent-Of-Origin Quantitative Trait Loci
- PUFA – Polyunsaturated Fatty Acids
- QTL – Quantitative Trait Loci
- REA – Ribeye Area
- RRBS – Reduced Representation Bisulfite Sequencing
- SFA – Saturated Fatty Acids
- SNP – Single Nucleotide Polymorphisms
- TF – Transcription Factor
- TFBS – Transcription Factor Binding Site
- *Trans*-eQTL – Trans-Acting Expression Quantitative Trait Loci
- WBSF0 – Warner-Bratzler Shear Force Measured 24 Hours After Slaughter
- WBSF14 – Warner-Bratzler Shear Force Measured 14 Days After Slaughter
- WBSF7 – Warner-Bratzler Shear Force Measured 7 Days After Slaughter
- WHC – Water Holding Capacity

1 CONTEXTUAL FRAMEWORK OF THE THESIS

1.1 Introduction

1.1.1 The Nelore cattle in Brazil

Beef is widely known as one of the main food sources on the globe. The *Bos indicus* breed, zebu, is the most common in Brazilian herds, with about 218 million animals in 2019 (UTSUNOMIYA et al., 2019). Zebu cattle are characterized by their distinctive hump, which produces a highly valued cut of meat in Brazil, as well as large ears and excess skin. The most prominent attribute of zebus is their excellent adaptation to tropical climates, tolerating high temperatures and lower-quality feed, and demonstrating superior resistance to parasites compared to taurine cattle (UTSUNOMIYA et al., 2019).

The most predominant zebu breed in Brazil is Nelore, accounting for approximately 80% of all Brazilian cattle, including both purebred animals and crossbreeds (SANTANA et al., 2016). However, Nelore cattle produce meat with lower tenderness (BRESSAN et al., 2011) and marbling compared to *Bos taurus* cattle (COOKE et al., 2020), features that directly influence consumer acceptance.

1.1.2 Relevant traits for beef improvement

For traits related to growth performance, traditional *B. indicus* cattle breeding efforts have produced significant annual genetic improvements (FERNANDES JÚNIOR et al., 2022), but for meat quality traits it is still a work in process. Tenderness is one of the most important focuses in genetic improvement programs for zebu breeds due to being the most important factor for consumer acceptance (KOOHMARAIE; WHEELER; SHACKELFORD, 1995). Shear force is the most used method to measure beef tenderness, equivalent to the force needed to cut a steak, that is negatively correlated with beef tenderness (*i.e.*, the shear force increases, while tenderness is decreased). This measurement is typically obtained using the Warner-Bratzler equipment (WHEELER; SHACKELFORD; KOOHMARAIE, 1997).

Postmortem muscle is tender immediately after slaughter, getting tougher with the absent production of ATP, resulting in the *rigor mortis* process in the following 24 hours (GUO; GREASER, 2022; KOOHMARAIE; WHEELER;

SHACKELFORD, 1995). The tenderization process occurs through proteolytic processes that degrade structural muscle proteins (KOOHMARAIE; WHEELER; SHACKELFORD, 1995). The muscle fiber comprises the sarcolemma, T-tubules, sarcoplasmic reticulum (SR), mitochondria, and myofibrils. The SR stores and regulates calcium ions, essential for muscle contraction, while mitochondria handle oxidative metabolism and ATP production. Myofibrils consist of sarcomeres, the smallest contractile units, with thin actin filaments attached to Z-lines and the protein titin providing structural support. Titin aligns actin and myosin, prevents overstretching, and contributes to passive tension. ATP depletion causes *rigor mortis*, where actomyosin bonds become permanent, influencing sarcomere length, water-holding capacity, and meat tenderness. Postmortem, myofibrillar proteins (*e.g*., desmin, TnT, titin) are degrade, primarily through the calcium-activated calpain system (KOOHMARAIE; WHEELER; SHACKELFORD, 1995), with cathepsins, also potentially involved in tenderization processes (GUO; GREASER, 2022). The distribution of these proteins in the muscle fiber is shown in Figure 1.1.

As mentioned before, calcium plays an important role in muscle tenderization. Mineral concentration in meat can reflect its nutritional impact on human health, being part of the process of muscle homeostasis, and associated with meat quality and livestock performance (DINIZ; BANERJEE; REGITANO, 2019). Sodium is a mineral required for multiple processes such as the active membrane transport of biomolecules, but it's overconsume can be associated with hypertension, harming human health (DOYLE; GLASS, 2010). Iron is responsible for the transport and attachment of oxygen in the blood. Beef is one of the major sources of iron in the human diet, which functions in the transport and attachment of oxygen in the blood. Iron deficiency is associated with anemia, but its higher consumption can increase the risk of type II diabetes and cancer (CZERWONKA; TOKARZ, 2017). In addition, the content of sodium, magnesium, phosphorus, and sulfur macro-minerals in beef was positively correlated with the beef color indices a* (redness) and b* (yellowness), while a negative correlation was found between the essential minerals (sulfur, potassium, and zinc) and cooking losses (PATEL et al., 2019). Another study

also showed that meat flavor significantly correlates with copper, iron, magnesium, phosphorus, potassium, sodium, and zinc (GARMYN et al., 2011).

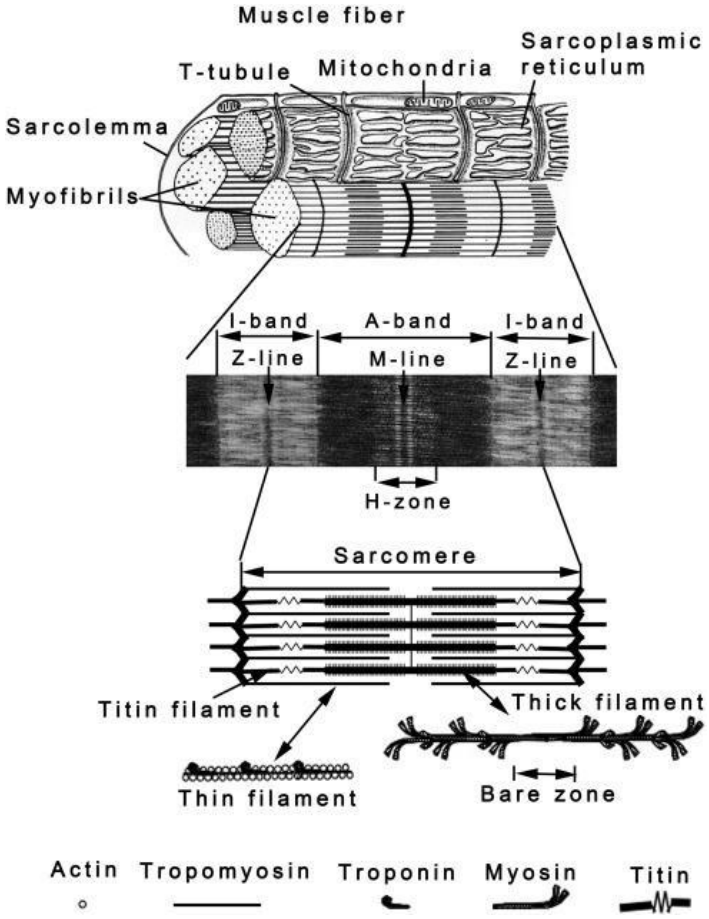


Figure 1.1 - Microstructure of a muscle fiber. Source: Guo and Greaser (2022).

Diniz, Banerjee, and Regitano (2019) reviewed studies that showed evidence of the relationship between minerals and the lipid profile in cattle muscle. As mentioned, *B. indicus* cattle have less marbling than other taurine breeds. Thus, the improvement of marbling, also known as intramuscular fat (IMF), and a reasonable increase in fat deposition can benefit *B. indicus* cattle. Additionally, fatty acids have been extensively studied, evaluating their effects on human health (DUGAN et al., 2011). Saturated fatty acids (SFAs) are known to increase low-density lipoprotein (LDL) cholesterol, popularly known as the “bad” type of cholesterol. However, early studies showed that SFAs are still important for maintaining desirable high-density

lipoprotein (HDL) levels and are not associated with cardiovascular disease (VAHMANI et al., 2015). Beef can be a valuable source of polyunsaturated fatty acids (PUFAs) as long-chain omega-3 (n3) and healthy *cis*-monounsaturated fatty acids (MUFAs) (VAHMANI et al., 2015). A diet with a balanced omega-6 (n6)/n3 ratio and n3-PUFAs is essential for physiological functions and helps reduce the risk of several diseases (ZÁRATE et al., 2017).

Fatty acids are also related to beef palatability. From a correlation study with 1,737 *Longissimus muscle* samples from Angus cattle with multiple collected phenotypes, stearic acid (C18:0), linoleic acid (C18:2), arachidonic acid (C20:4), and PUFA were negatively correlated with tenderness. At the same time, MUFA showed a positive correlation with this phenotype. Juiciness was negatively correlated with C18:2, **α**-linoleic acid (C18:3), C20:4, and PUFA (GARMYN et al., 2011). Given the increasing world's demand for high-quality beef, applying genomic selection techniques can be very beneficial for improving meat quality traits in Nelore cattle (FERNANDES JÚNIOR et al., 2022).

1.1.3 Genomic selection and association studies

Beginning with initiatives to enhance dairy cows, pedigree records and performance data were foundation for developing genetic selection programs in the pre-genomic era. In the early selection procedures, the producers selected animals with desirable milk production, followed by comparisons of this trait between daughters and dams. After that, the researchers observed the need to consider the environmental factors as the contemporary comparisons in genetic models (WEIGEL et al., 2017). Next-generation sequencing technologies have emerged as powerful tools for genome-wide variant screening, helping to increase the coverage of mutations throughout the genome. The availability of genomic data enabled the inclusion of genomic relationship matrices in genetic evaluation models, leading to the genomic selection era (WEIGEL et al., 2017).

Genomic selection for meat quality phenotypes allows greater accuracy in predicting the animal's genetic potential, usually determined indirectly by measuring phenotypes after slaughtering individuals from the same family. Compared to the early methods to estimate breeding values, the genomic selection makes possible

the identification of quantitative trait loci (QTLs) (WEIGEL et al., 2017), which consist of functional variants associated with phenotypic variation obtained from methods as the genome-wide association studies (GWAS) (HAYES; GODDARD, 2010). The QTLs can comprise genomic windows with multiple variants or single nucleotide polymorphisms (SNPs), markers widely used in large-scale genotyping (CHANG et al., 2018) corresponding to a single nucleotide variation that doesn't match the reference sequence.

Among the limitations of GWAS methods is the requirement for a large initial sample set and considerable genome coverage (CHANG et al., 2018; SCHAID; CHEN; LARSON, 2018). Additionally, QTLs typically have a low effect on complex phenotypes (HAYES; GODDARD, 2010), that are often epistatic or pleiotropic, reflecting multifaceted dynamics between markers and genes. Furthermore, identifying causal variants is difficult due to the linkage disequilibrium (WEIGEL et al., 2017), which refers to the non-random association of alleles at different loci. The genetic complexity of the phenotype must be known prior to analysis, controlling relevant variables by estimating the Genomic Estimated Breeding Values (GEBVs, *i.e.*, the genetic merit of animals based on phenotypes corrected by known environmental factors) or to include these effects directly in the association model (CHANG et al., 2018; SCHAID; CHEN; LARSON, 2018).

Another important factor to consider in GWAS is the relatively high costs to obtain high-density SNP genotypes in large sample sizes. An alternative to reduce costs while still obtaining a suitable dataset is SNP imputation methods (BALDING, 2006; SCHAID; CHEN; LARSON, 2018). Imputation techniques fill gaps in missing genotypes by leveraging the linkage disequilibrium of known variants from a subset of samples with higher coverage, applying maximum likelihood or probabilistic approaches. Thus, imputation techniques significantly enhance the number of SNPs available for association analysis (SCHAID; CHEN; LARSON, 2018).

Expression information can help us understand the functional relevance of the related genes with metabolic pathways (NICOLAE et al., 2010), helping us understand the mechanisms involved in the relationship between genotype, transcription, and phenotype. Thus, in addition to phenotypes, other dependent

variables can be used to associate variants with other factors, such as gene expression, and identify eQTLs.

The integration of GWAS results with gene expression data obtained through RNA sequencing (RNA-seq), often demonstrates that QTLs associated with traits of interest can also be identified as eQTLs, meaning SNPs associated with gene expression (NICOLAE et al., 2010; SCHAID; CHEN; LARSON, 2018). Although it is arbitrary, eQTLs can have their regulatory potential classified based on their distance from the regulated gene: if such variants are located within 1 Mb of the genes, they are classified as *cis*-eQTLs, and if the distance between the eQTLs and the genes exceeds 1 Mb, they are classified as *trans*-eQTLs (YAO et al., 2017). Cesar et al. (2018) identified *cis* and *trans*-eQTLs associated with intramuscular fat phenotypes in muscle samples from a Nelore population.

1.1.4 Allele-specific expression as a tool to identify cis-regulation

Since alleles in the same DNA strand can be affected by *cis*-eQTLs (MASON et al., 2018), these elements may be linked to genes that exhibit allele-specific expression (ASE) (KHANSEFID et al., 2018; MASON et al., 2018). ASE analysis, which is used to integrate genotypic and RNA sequencing data (genome and transcriptome) (CASTEL et al., 2015; CHAMBERLAIN et al., 2015), is a very helpful tool for identifying *cis*-regulatory causal variants by observing which allele is being expressed or repressed (ZOU et al., 2019). ASE is defined by the imbalanced expression of two alleles of the same locus of a diploid organism (PASTINEN, 2010), observed only in heterozygous *loci* where RNA-seq counts for one allele are more abundant than the ones from the other allele (Figure 1.2).

The presence of genes with ASE in bovine is common and tissue-specific (CHAMBERLAIN et al., 2015). The ASE pattern was found in muscle from *Bos taurus* cattle, where the authors identified 5,658 SNPs marking regions with ASE (ASE SNPs) from whole-genome sequencing data from 19 Limousin animals, with 2,107 overlapping meat and carcass quality QTLs (GUILLOCHEAU et al., 2019). From genotypes obtained with the Illumina BovineHD BeadChip (770 K SNPs) and 190 Nelore samples, 820 ASE SNPs were identified in *Bos indicus* muscle cattle, affecting genes functionally related to muscle homeostasis (DE SOUZA et al., 2020).

These manuscripts suggest that ASE analyses are highly dependent on dense genotype coverage and large sample sizes, showing that ASE SNPs are extensively spread in bovine muscle and may influence meat quality characteristics.

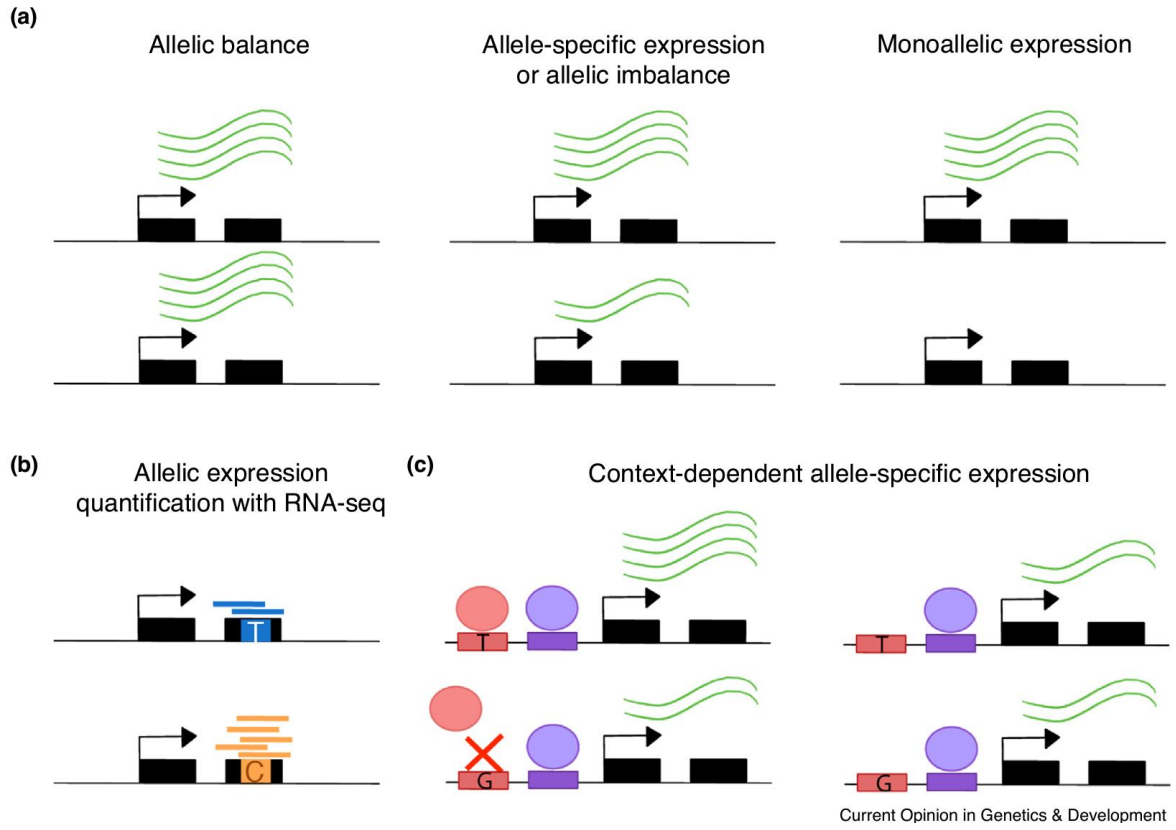


Figure 1.2 - Expression patterns. a) Representation of expression patterns; b) Allele-specific expression detection from RNA-seq; c) Potential mechanisms regulating allele-specific expression in a context-dependent manner. Source: Robles-Espinoza et al., (2020)

There are different categories for the ASE (Figure 1.2). The allelic imbalance can be categorized as monoallelic (Figure 1.2A, third example), *i.e*., if only one allele is expressed or there is a substantial differential expression between the alleles. It can be context-specific (Figure 1.2C), reflecting the ASE level difference between samples submitted to different conditions, treatments, environments (KNOWLES et al., 2017), or phenotypes; there is no convention for naming this class of ASE, but we will call it differential ASE (DASE) in this entire manuscript.

As ASE is calculated for each sample, the impact of technical issues present in total expression investigations is minimized (CASTEL et al., 2015). A heterozygous genotype for the tested SNP is required for ASE analysis, which counts transcript reads corresponding to each allele independently to identify allelic expression (CASTEL et al., 2015).

An issue of ASE analysis is that RNA-seq reads exhibit alignment bias, favoring the reference alleles' expression (DEGNER et al., 2009). Some methods can be used to prevent this effect, including using a masked reference genome, personalized genomes, or the WASP software (VAN DE GEIJN et al., 2015). The WASP program has been widely used, since it was shown to have the best performance when comparing the efficiency of these methods (CASTEL et al., 2015; VAN DE GEIJN et al., 2015). Finding SNPs in the reference genome, flipping their nucleotides to the alternative variants, and re-mapping the reads with varying mapping efficiencies is the foundation of WASP mapping bias removal (VAN DE GEIJN et al., 2015).

Following alignments, a quality check is carried out based on the minimal mapping quality, base quality, and number of reads per SNP after eliminating duplicates (ROBLES-ESPINOZA et al., 2021). It is advised to acquire allele-specific counts at each site per animal, using the GATK ASE Read counter tool and the GATK pipeline for RNA-seq processing. The relative expression of reference and alternative alleles can be computed using these counts (CASTEL et al., 2015). The binomial test examines the difference between the alleles' expression and the expected proportion value in a biallelic expression (0.5), consisting of one of the most traditional methods for determining the statistical difference between allele-specific reads (CASTEL et al., 2015).

1.1.5 Parent-of-origin effects and genomic imprinting

If a gene shows monoallelic ASE depending on the parental origin of the allele, it is classified as genomic *imprinting*. This phenomenon, which emerges through epigenetic changes such as DNA methylation, is crucial for the development of mammals (FERGUSON-SMITH; SURANI, 2001). Imprinted genes in the bovine model have been the subject of few published studies (DINDOT et al., 2004;

KARAMI et al., 2019), and until this date, no research on *B. indicus* animals has been found.

Since selection can affect an animal's development, its impact on the expression pattern of *imprinted* genes needs to be carefully considered (SMITH et al., 2015). Furthermore, imprinting may impact how the phenotypes being selected manifest in the progeny. This is because only the allele acquired from a particular parent will be expressed in the offspring for genes that have an imprinting effect.

Despite imprinting being the most studied parent-of-origin effect (POE), POEs can also happen without direct association with gene expression. POEs did not follow the Mendelian inheritance, as the contribution of alleles to the phenotypic variation of the progeny is dependent on which progenitor this allele came from (GUILMATRE; SHARP, 2012). This phenomenon can result from maternal/parental genetic effects that are exhibited in the progeny without transmitting any causal mutation, from random monoallelic expression and epigenetic effects in response to the environment.

There has been described evidence of POEs affecting complex traits (LAWSON; CHEVERUD; WOLF, 2013; MOTT et al., 2014). Multiple QTLs for various traits were found showing POEs in the mice model (MOTT et al., 2014). The parental origin contribution to 33 performance traits in pigs was tested using an imprinting model, showing that genomic imprinting accounted for between 5% and 19% of the overall additive genetic variance (NEUGEBAUER; LUTHER; REINSCH, 2010).

Traditional association models usually consider the allelic contribution of the two parents as equivalent, thus failing to identify the influence of heritable components and POEs on the phenotypic variation (GUILMATRE; SHARP, 2012). POEs can emerge from genetic interactions and epigenetic modulations mediated by environmental conditions (GUILMATRE; SHARP, 2012). Most POE/imprinting studies use methods that lie on the trio information, *i.e.*, high-density genotypes from each parent and their offspring. Unfortunately, trio data are costly for large sample datasets.

The evidence that heterozygous samples have more phenotypic variance than homozygous samples at *loci* showing POEs was described. These findings were combined with the expected pleiotropic effect of genes in complex traits, supporting the premise to the development of POIROT software, that is able to study POEs without trio information. POIROT applies a Robust Omnibus Test to compare the phenotypic covariance matrix of heterozygotes and homozygotes (HEAD et al., 2023).

By identifying regions with ASE or *imprinting*, it is possible to identify different *cis*-regulatory or epigenetic polymorphisms associated with these expression patterns, as well as relate them to phenotypes of interest, being an ally to conventional methodologies (CHENG et al., 2015). In addition, the integration between allelic expression information and potentially functional variants in the context of livestock traits should contribute to the mapping of regulatory elements, whose prioritization has great potential to increase the accuracy of genomic selection (XIANG et al., 2019).

1.1.6 Regulatory mechanisms affecting gene expression

Herein, this study has already highlighted how the ASE study is a powerful tool for searching for this expression pattern's causal variants. Since allelic expression can result from *cis*-regulatory elements, aseQTL (Allele-Specific Expression Quantitative Trait *Loci*) identification analyses were developed to identify elements directly responsible for ASE and not generalized to total expression such as eQTLs (BATTLE et al., 2014). However, aseQTL analysis is more conservative than an eQTL association due to the requirement of animal groups that are heterozygotes for SNPs in expressed regions. Thus, both analyses can complement screening for regulatory variants affecting ASE regions.

Once the candidate regulatory variants have been identified, fine mapping is necessary to get close to the causal SNPs (ZOU et al., 2019). It is possible to prioritize SNPs based on the predicted functional consequence by considering their genomic locations. The potential functional effect of the SNP can be classified according to whether it is present in protein-coding DNA regions or non-coding regions (SCHAID; CHEN; LARSON, 2018). The impact of the variant in coding

regions is greater if the nucleotide change of the SNP alters an amino acid and, consequently, the structure of the protein or its isoforms (YUAN et al., 2006). SNPs located in non-protein coding regions are more common in the genome and tend to have a regulatory role (SCHAID; CHEN; LARSON, 2018). When the SNP changes the amino acid, but the peptide is not lost early, the variation is classified as non-synonymous, and if this variation changes the protein structure, it is missense (YUAN et al., 2006). Intronic variants can modify transcription factor binding sites (TFBSs) but with a lower potential impact than expected in promoter regions (YUAN et al., 2006).

Cis-regulatory or epigenetic regulatory mechanisms may act on potentially regulatory variants in an allele-specific manner. Cytosine methylation in CpG dinucleotide-rich regions (CpG islands) is considered a marker for gene silencing and imprinting (PFEIFER, 2000). This happens when proteins bind to the methyl group, preventing transcription factors from accessing the appropriate locations and promoting chromatin condensation, which renders the DNA strand transcriptionally inactive (MOORE; LE; FAN, 2013). Methylation and gene silencing have an intricate association because, as a recent research has demonstrated, significant demethylation at CpG sites follows transcriptional activation in response to an infection, which is a downstream effect of transcription factors (TFs) binding (PACIS et al., 2019).

TFs are proteins responsible for controlling RNA-Polymerase recruitment to the DNA in the transcription, which can activate or block gene expression. These proteins act by binding in TF motifs, generally located in the promoter region (WANG et al., 2015). Gaur et al., (2013) reviewed multiple studies on the regulatory mechanisms affecting allelic imbalance and illustrated that allele-specific differences in TF binding and open chromatin can influence gene expression, partly driven by heritable genetic variation. In a similar dynamic, miRNAs have an affinity to their binding sites, located mostly on the 3’ UTR end of target genes in mammals, degrading the mRNA and blocking translation (CANNELL; KONG; BUSHELL, 2008). As represented in Figure 1.2C, TFBSs or miRNAs motif changes are also known *cis*-regulatory mechanisms that can result in ASE (GAUR et al., 2013).

These exposures demonstrate the importance of conducting ASE studies to better understand the regulatory landscape of gene expression. They help increase knowledge about POEs, such as genomic imprinting, and draft hypotheses about their functional impact on beef cattle phenotypes.

1.2 Justification

Given the increasing number of studies exploring candidate genes that may impact the phenotypic variation of traits relevant for meat quality improvement, the identification of markers responsible for regulating the expression patterns of these genes is highly needed. Since the identification of these regulatory variants is improved through ASE analyses, this study aimed to identify genome-wide ASE SNPs in *B. indicus* muscle, also identifying those that directly impact muscle-related traits (DASE SNPs) and predicting the related *cis*-acting variants and their respective regulatory mechanisms. Furthermore, the identification of PO-QTLs and ASM regions are complementary factors to better understand the ASE pattern, identify genomic locations under parental effect and point out potential candidates for imprinting in cattle. This work complemented results from previous studies and provided multiple novel knowledge layers for the genetic regulation in *B. indicus* muscle. A representation of these variants is represented in Figure 1.3.

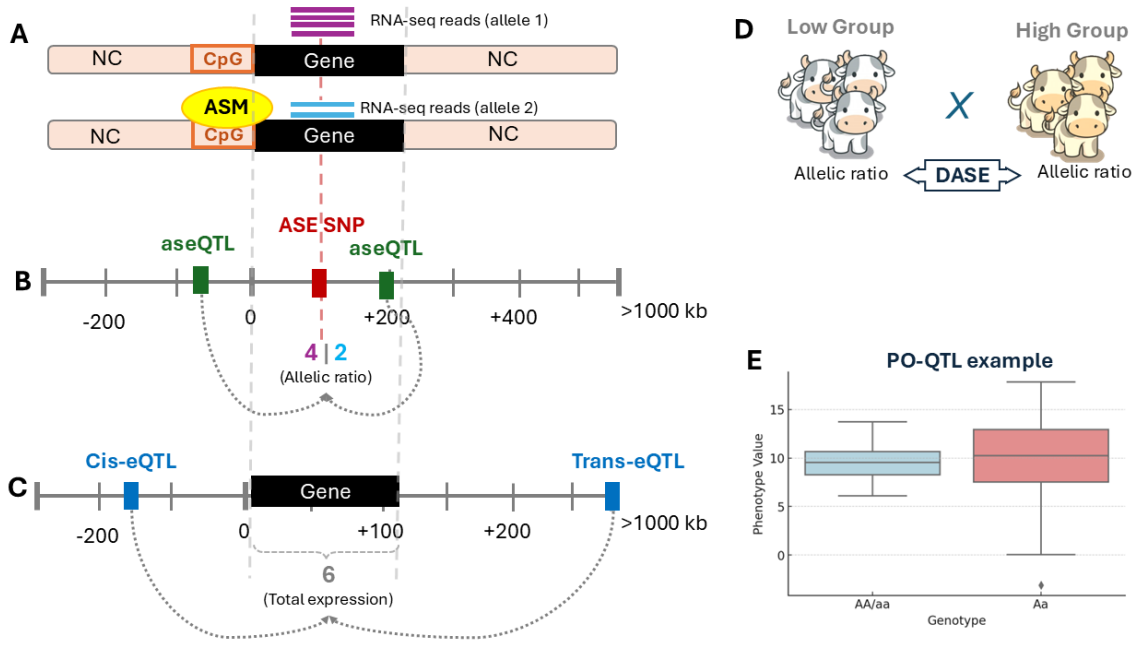


Figure 1.3 – Representation of variants described in this study. A) Schematic representation of a diploid genome, showing RNA-seq reads with different allelic representations (*i.e.,* allele 1 and allele 2) mapping to an example gene on each chromosome. Allele-specific methylation (ASM) refers to methylation occurring in a CpG region specific to only one haplotype. B) Illustration of a SNP marking the allelic expression ratio observed for the example gene (ASE SNP) and aseQTLs, which are *cis*-regulatory variants associated with the level of allelic imbalance in the ASE SNP. C) Representation of eQTLs, which are variants associated with the total expression of the example gene. By definition, a *cis*-eQTL is located within 1 Mb of the coding region, whereas a *trans*-eQTL is located beyond 1 Mb from the coding region. D) Interpretation of Differential Allele-Specific Expression (DASE), which is based on differences in allelic imbalance levels between groups of animals with contrasting phenotypes. E) Explanation of the test used to identify Parent-of-Origin (PO)-QTLs: In a region affected by PO effects, higher phenotypic variance is expected in the heterozygous group.

1.3 Objectives

1.3.1 Main Objective

To identify markers associated with the regulation of muscle phenotypes using approaches that consider ASE and parent-of-origin effects, with potential future applicability in Nelore beef improvement programs.

1.3.2 Specific Objectives

- To identify ASE SNPs in bovine muscle from a dense set of genotypes obtained through imputation and identify *cis*-eQTLs and aseQTLs regulating these genomic regions;
- To associate ASE with meat quality phenotypes by identifying DASE SNPs;
- To identify QTLs associated with meat quality phenotypes that exhibit parent-of-origin effects;
- To identify ASM regions using reduced representation bisulfite sequencing (RRBS) data, highlighting evidence of new candidate imprinted genes;
- To evaluate the most promising variants for further exploration as potential markers for meat quality in cattle.

REFERENCES

BATTLE, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. **Genome Research**, v. 24, n. 1, p. 14–24, jan. 2014.

BRESSAN, M. C. et al. Physicochemical properties of meat from Bos taurus and Bos indicus. **Revista Brasileira de Zootecnia**, v. 40, n. 6, p. 1250–1259, jun. 2011.

CANNELL, I. G.; KONG, Y. W.; BUSHELL, M. How do microRNAs regulate gene expression? **Biochemical Society Transactions**, v. 36, n. 6, p. 1224–1231, 1 dez. 2008.

CASTEL, S. E. et al. Tools and best practices for data processing in allelic expression analysis. **Genome Biology**, v. 16, n. 1, p. 195, dez. 2015.

CHAMBERLAIN, A. J. et al. Extensive variation between tissues in allele specific expression in an outbred mammal. **BMC Genomics**, v. 16, n. 1, p. 993, dez. 2015.

CHANG, L.-Y. et al. High density marker panels, SNPs prioritizing and accuracy of genomic selection. **BMC Genetics**, v. 19, n. 1, p. 4, dez. 2018.

CHENG, H. H. et al. Fine mapping of QTL and genomic prediction using allele-specific expression SNPs demonstrates that the complex trait of genetic resistance to Marek's disease is predominantly determined by transcriptional regulation. **BMC Genomics**, v. 16, n. 1, p. 816, dez. 2015.

COOKE, R. F. et al. Cattle adapted to tropical and subtropical environments: social, nutritional, and carcass quality considerations. **Journal of Animal Science**, v. 98, n. 2, p. skaa014, 1 fev. 2020.

CZERWONKA, M.; TOKARZ, A. Iron in red meat–friend or foe. **Meat Science**, v. 123, p. 157–165, jan. 2017.

DE SOUZA, M. M. et al. Allele-specific expression is widespread in Bos indicus muscle and affects meat quality candidate genes. **Scientific Reports**, v. 10, n. 1, p. 10204, 23 jun. 2020.

DEGNER, J. F. et al. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. **Bioinformatics**, v. 25, n. 24, p. 3207–3212, 15 dez. 2009.

DINDOT, S. V. et al. Conservation of genomic imprinting at the XIST, IGF2, and GTL2 loci in the bovine. **Mammalian Genome**, v. 15, n. 12, p. 966–974, dez. 2004.

DINIZ, W. J. D. S.; BANERJEE, P.; REGITANO, L. C. A. Cross talk between mineral metabolism and meat quality: a systems biology overview. **Physiological Genomics**, v. 51, n. 11, p. 529–538, 1 nov. 2019.

DOYLE, M. E.; GLASS, K. A. Sodium Reduction and Its Effect on Food Safety, Food Quality, and Human Health. **Comprehensive Reviews in Food Science and Food Safety**, v. 9, n. 1, p. 44–56, jan. 2010.

DUGAN, M. et al. Review: *Trans*- forming beef to provide healthier fatty acid profiles. **Canadian Journal of Animal Science**, v. 91, n. 4, p. 545–556, dez. 2011.

FERGUSON-SMITH, A. C.; SURANI, M. A. Imprinting and the Epigenetic Asymmetry Between Parental Genomes. **Science**, v. 293, n. 5532, p. 1086–1089, 10 ago. 2001.

FERNANDES JÚNIOR, G. A. et al. Current applications and perspectives of genomic selection in Bos indicus (Nellore) cattle. **Livestock Science**, v. 263, p. 105001, set. 2022.

GARMYN, A. J. et al. Estimation of relationships between mineral concentration and fatty acid composition of longissimus muscle and beef palatability traits1. **Journal of Animal Science**, v. 89, n. 9, p. 2849–2858, 1 set. 2011.

GAUR, U. et al. Research progress in allele-specific expression and its regulatory mechanisms. **Journal of Applied Genetics**, v. 54, n. 3, p. 271–283, ago. 2013.

GUILLOCHEAU, G. M. et al. Survey of allele specific expression in bovine muscle. **Scientific Reports**, v. 9, n. 1, p. 4297, 12 mar. 2019.

GUILMATRE, A.; SHARP, A. Parent of origin effects. **Clinical Genetics**, v. 81, n. 3, p. 201–209, mar. 2012.

GUO, W.; GREASER, M. L. Muscle structure, proteins, and meat quality. Em: **New Aspects of Meat Quality**. [s.l.] Elsevier, 2022. p. 15–37.

HAYES, B.; GODDARD, M. Genome-wide association and genomic selection in animal breedingThis article is one of a selection of papers from the conference “Exploiting Genome-wide Association in Oilseed Brassicas: a model for genetic improvement of major OECD crops for sustainable farming”. **Genome**, v. 53, n. 11, p. 876–883, nov. 2010.

HEAD, S. T. et al. POIROT: a powerful test for parent-of-origin effects in unrelated samples leveraging multiple phenotypes. **Bioinformatics**, v. 39, n. 4, p. btad199, 3 abr. 2023.

KARAMI, K. et al. Characterization of bovine (Bos taurus) imprinted genes from genomic to amino acid attributes by data mining approaches. **PLOS ONE**, v. 14, n. 6, p. e0217813, 6 jun. 2019.

KHANSEFID, M. et al. Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. **BMC Genomics**, v. 19, n. 1, p. 793, dez. 2018.

KNOWLES, D. A. et al. Allele-specific expression reveals interactions between genetic variation and environment. **Nature Methods**, v. 14, n. 7, p. 699–702, jul. 2017.

KOOHMARAIE, M.; WHEELER, T. L.; SHACKELFORD, S. D. Beef tenderness: regulation and prediction. **Clay Center, NE: USDA-ARS U.S. Meat Animal Research Center**, p. 24, 1995.

LAWSON, H. A.; CHEVERUD, J. M.; WOLF, J. B. Genomic imprinting and parent-of-origin effects on complex traits. **Nature Reviews Genetics**, v. 14, n. 9, p. 609–617, set. 2013.

MASON, V. C. et al. eQTL discovery and their association with severe equine asthma in European Warmblood horses. **BMC Genomics**, v. 19, n. 1, p. 581, dez. 2018.

MOORE, L. D.; LE, T.; FAN, G. DNA Methylation and Its Basic Function. **Neuropsychopharmacology**, v. 38, n. 1, p. 23–38, jan. 2013.

MOTT, R. et al. The Architecture of Parent-of-Origin Effects in Mice. **Cell**, v. 156, n. 1–2, p. 332–342, jan. 2014.

NEUGEBAUER, N.; LUTHER, H.; REINSCH, N. Parent-of-origin effects cause genetic variation in pig performance traits. **Animal**, v. 4, n. 5, p. 672–681, 2010.

NICOLAE, D. L. et al. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. **PLoS Genetics**, v. 6, n. 4, p. e1000888, 1 abr. 2010.

PACIS, A. et al. Gene activation precedes DNA demethylation in response to infection in human dendritic cells. **Proceedings of the National Academy of Sciences**, v. 116, n. 14, p. 6938–6943, 2 abr. 2019.

PASTINEN, T. Genome-wide allele-specific analysis: insights into regulatory variation. **Nature Reviews Genetics**, v. 11, n. 8, p. 533–538, ago. 2010.

PATEL, N. et al. Relationships of a Detailed Mineral Profile of Meat with Animal Performance and Beef Quality. **Animals**, v. 9, n. 12, p. 1073, 3 dez. 2019.

PFEIFER, K. Mechanisms of Genomic Imprinting. **The American Journal of Human Genetics**, v. 67, n. 4, p. 777–787, out. 2000.

ROBLES-ESPINOZA, C. D. et al. Allele-specific expression: applications in cancer and technical considerations. **Current Opinion in Genetics & Development**, v. 66, p. 10–19, fev. 2021.

SANTANA, M. L. et al. Structure and genetic diversity of Brazilian Zebu cattle breeds assessed by pedigree analysis. **Livestock Science**, v. 187, p. 6–15, maio 2016.

SCHAID, D. J.; CHEN, W.; LARSON, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. **Nature Reviews Genetics**, v. 19, n. 8, p. 491–504, ago. 2018.

SMITH, L. C. et al. Epigenetic consequences of artificial reproductive technologies to the bovine imprinted genes SNRPN, H19/IGF2, and IGF2R. **Frontiers in Genetics**, v. 6, 25 fev. 2015.

UTSUNOMIYA, Y. T. et al. Genomic clues of the evolutionary history of *Bos indicus* cattle. **Animal Genetics**, v. 50, n. 6, p. 557–568, dez. 2019.

VAHMANI, P. et al. The scope for manipulating the polyunsaturated fatty acid content of beef: a review. **Journal of Animal Science and Biotechnology**, v. 6, n. 1, p. 29, dez. 2015.

VAN DE GEIJN, B. et al. WASP: allele-specific software for robust molecular quantitative trait locus discovery. **Nature Methods**, v. 12, n. 11, p. 1061–1063, nov. 2015.

WANG, G. et al. Understanding Transcription Factor Regulation by Integrating Gene Expression and DNase I Hypersensitive Sites. **BioMed Research International**, v. 2015, p. 1–7, 2015.

WEIGEL, K. A. et al. A 100-Year Review: Methods and impact of genetic selection in dairy cattle—From daughter–dam comparisons to deep learning algorithms. **Journal of Dairy Science**, v. 100, n. 12, p. 10234–10250, dez. 2017.

WHEELER, T. L.; SHACKELFORD, S. D.; KOOHMARAIE, M. Standardizing Collection and Interpretation of Warner-Bratzler Shear Force and Sensory Tenderness Data. **50th Annual Reciprocal Meat Conference Proceedings**, v. 50, p. 9, 1997.

XIANG, R. et al. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. **Proceedings of the National Academy of Sciences**, v. 116, n. 39, p. 19398–19408, 24 set. 2019.

YAO, C. et al. Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. **The American Journal of Human Genetics**, v. 100, n. 4, p. 571–580, abr. 2017.

YUAN, H.-Y. et al. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. **Nucleic Acids Research**, v. 34, n. Web Server, p. W635–W641, 1 jul. 2006.

ZÁRATE, R. et al. Significance of long chain polyunsaturated fatty acids in human health. **Clinical and Translational Medicine**, v. 6, n. 1, p. e25, dez. 2017.

ZOU, J. et al. Leveraging allelic imbalance to refine fine-mapping for eQTL studies. **PLOS Genetics**, v. 15, n. 12, p. e1008481, 13 dez. 2019.

2 IDENTIFICATION OF ASE SNPS, ASEQTLS, AND *CIS*-EQTLS FROM A DENSE GENOTYPE PANEL

In this chapter, we will describe the obtention of the genotype and allelic counts files, with the related RNA-seq mapping procedure used for multiple purposes of this Doctorate Project. These results were published in the Journal BBA Gene Regulatory Mechanisms on October 8th, 2022. The manuscript is transcribed into the following subsections and its header is shown in Figure 2.1.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2022.194886.


BBA - Gene Regulatory Mechanisms 1865 (2022) 194886

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Allele-specific expression reveals functional SNPs affecting muscle-related genes in bovine

Jennifer Jessica Bruscadin [a,b], Tainã Figueiredo Cardoso [b], Wellison Jarles da Silva Diniz [c], Juliana Afonso [b], Marcela Maria de Souza [d], Juliana Petrini [e], Bruno Gabriel Nascimento Andrade [f], Vinicius Henrique da Silva [e], José Bento Sterman Ferraz [g], Adhemar Zerlotini [h], Gerson Barreto Mourão [e], Luiz Lehmann Coutinho [e], Luciana Correia de Almeida Regitano [b,*]

[a] Center of Biological Sciences and Health, Federal University of São Carlos, São Carlos, SP, Brazil
[b] Embrapa Pecuária Sudeste, São Carlos, SP, Brazil
[c] Department of Animal Sciences, Auburn University, AL, USA
[d] Department of Animal Science, Iowa State University, Ames, IA, USA
[e] Center for Functional Genomics, Department of Animal Science, University of São Paulo (ESALQ-USP), Piracicaba, SP, Brazil
[f] Computer Science Department, Munster Technological University, MTU/ADAPT, Cork, Ireland
[g] Department of Veterinary Medicine, University of São Paulo (FMVZ-USP), Pirassununga, SP, Brazil
[h] Embrapa Informática Agropecuária, Campinas, SP, Brazil

Figure 2.1 - Header of the published manuscript: “Allele-specific expression reveals functional SNPs affecting muscle-related genes in bovine.

2.1 Abstract

Single nucleotide polymorphisms showing allele-specific expression (ASE SNPs) are useful for *cis*-regulatory variants discovery. Despite this potential, there are expensive costs involved in genome-level ASE analysis for large sample sizes. If different data resolutions are available, genotype imputation can be used to mitigate this limitation. Aiming to increase the power to detect regulatory variants, we used a large dataset (>4 million) of imputed SNP genotypes and RNA-Seq data from 190 Nelore steers. Differences between major and minor allele expressions in muscle were tested with a Binomial Test. We identified 38,177 ASE SNPs (FDR ≤ 0.05) within 7304 linkage disequilibrium blocks. After that, we searched for aseQTLs (i.e., neighboring SNPs potentially regulating the ASE SNPs' allelic expression) by comparing the ASE of heterozygous to homozygous sample groups under a Wilcoxon Rank Sum test. We identified 21,543 aseQTLs potentially regulating 430 ASE SNPs (FDR ≤ 0.05). A total of 3333 *cis*-eQTLs (being 2098 ASE SNPs and 1075 aseQTLs) were associated with the expression of 758 transcripts (FDR ≤ 0.05), demonstrating the *cis*-regulatory effect of these ASE SNPs and aseQTLs. Data integration showed reproducibility with previous studies in bovine ASE and genomic imprinting. Furthermore, we identified 36,756 novel ASE regions due to the imputation approach. Comparisons with epigenetics data from Functional Annotation of Animal Genomes (FAANG) suggest a regulatory potential of the ASE-related SNPs. The affected genes were enriched in metabolic pathways essential for muscle homeostasis. These findings reinforce the potential of using ASE for discovering *cis*-regulatory SNPs that may affect muscle-related traits.

Keywords*:* Cattle *Cis*-regulation Epigenetics Genomics Data integration

2.2 Introduction

Single nucleotide polymorphisms (SNPs) are markers used in large-scale genotyping to assess the associations between phenotypes and genomic variants [1]. The new generation sequencing enables fast and accurate high-density genotyping, but it is expensive for large sample populations. The SNPs' imputation can reduce costs by inferring unobserved genotypes in lower density panels, mediated by linked SNPs with known genotypes from sequencing data obtained from fewer samples [2,3].

The use of genotypic information and RNA sequencing (RNA-Seq) data from the same population enables the identification of SNPs showing allele-specific expression (ASE SNPs), *i.e.*, regions presenting imbalanced expression between the alleles inherited from each parental chromosome [4,5]. When the differential expression between two alleles is parentally-dependent, the observed ASE pattern responds to genomic imprinting and commonly presents monoallelic expression [6]. Finding expressed or repressed alleles assist in the search for causal *cis*-regulatory variants [7,8].

The ASE SNP identification is insufficient to know which SNPs regulate its expression pattern, as regulatory variants are usually located in non-coding regions [9]. Thereby, ASE is caused by multiple cumulative *cis*-regulatory mechanisms affecting the allele [10], *e.g.*, transcription factor (TF) binding within accessible chromatin regions, which are modulated by DNA methylation and histones modifications [11]. These mechanisms can act as aseQTLs, *i.e.*, SNPs directly associated with the allelic ratio disequilibrium of the ASE SNPs [12], and *cis*-eQTLs, *i.e.*, SNPs that regulate the expression of nearby transcripts in the genome [13].

Recently, the presence of the ASE has been investigated in cattle [5,7,9,14,15], exploring related *cis*-regulatory mechanisms [7,16]. Chamberlain et al. [5] reported tissue-specific allelic imbalance in Holstein cattle. Studies on Nelore and Limousin cattle muscle have shown that ASE is well distributed across the chromosomes [7,14] and varies among samples [14]. The studies mentioned above suggest that genes presenting allelic expression imbalance (ASE genes) are relevant for muscle maintenance and have been associated with beef quality traits

[7,14]. In addition, Silva et al. identified copy number variation regions (CNVs) associated with tenderness in Nelore muscle [17]. CNVs result from losses and gains of sequence due to structural alterations that can affect the results of ASE analysis [18].

As ASE is detected by sample, the number of variants available to test is more informative for evaluating the ASE profile than the population size. In a previous study performed in the muscle of 190 Nelore steers, based on the 770 K genotypes from Illumina Bovine HDBead Chip, De Souza et al. identified a total of 820 ASE SNPs within 530 genes, whereas with whole-genome sequencing data from an experimental population of 19 Limousin muscle samples, Guillocheau et al. [7] identified 6908 ASE SNPs within 2451 genes, although this later study used the previous UMD3.1 bovine assembly.

In commercial genotyping chips, the SNPs are selected to represent linkage disequilibrium (LD) regions, so multiple SNPs cannot be tested for ASE if the selected SNP is not heterozygous in a given sample. Our assumption is that using a larger genotype dataset could increase the probability of finding heterozygous SNPs and considerably improve the coverage of ASE regions identified in this 190-steers Nelore population. The increased coverage would thus provide more SNPs to be tested and encompass more muscle-transcribed genes. Therefore, the higher density of the ASE profile in Nelore muscle would benefit the discovery of causal SNPs, *i.e.*, the variants affected by *cis*-regulatory SNPs, such as *cis*-eQTLs and aseQTLs, or those showing a regulatory potential by themselves.

In this study, we used a large dataset of imputed SNP genotypes to increase the power to detect regulatory variants in bovine muscle. We identified ASE SNPs and aseQTLs, pointing out those that are also *cis*-eQTLs in the muscle of a Nelore population. The results were compared to previous data in the same population from the Illumina BovineHD BeadChip genotypes (770 K SNPs) [14] and ASE data detected in the muscle of Limousin cattle [7]. Altogether, we increased the available variants to about 4 million SNPs, enhanced the coverage of discovered ASE in bovine muscle, and showed consistency with previous reports. Furthermore, we

presented potential regulatory variants and regulatory mechanisms related to the ASE pattern and its role in muscle homeostasis.

2.3. Methods

2.3.1 Experimental animals and sample collection

All the experimental procedures were carried out according to the corresponding guidelines provided by the Institutional Animal Care and Use Committee Guidelines of Embrapa Pecuária Sudeste (São Carlos, São Paulo, Brazil. Protocol CEUA 01/2013), that approved all experimental protocols (CEUA Process 01/2013). The animals used in this manuscript are a subset of a Nelore steer experimental population described elsewhere [19]. *Longissimus thoracis* muscle samples, between the 11th and 13th ribs, were collected from all animals at slaughter and stored at - 80 ◦C until RNA extraction, as reported by Tizioto et al. [19]. We selected 190 animals with RNA-Seq and high-density genotypic data for this work, being the same samples used in previous studies of our group [14,16], to ensure comparability.

2.3.2 RNA extraction and sequencing

RNA extraction was carried out with 100 mg of muscle tissue per sample using TRIzol reagent (Life Technologies, Carlsbad, CA, USA) following the manufacturer's protocol. Details on RNA extraction, quality control, and sequencing were previously published [20]. The RNA sequencing was performed in an Illumina HiSeq 2500 equipment (Illumina, San Diego, CA, USA) at the Genomics Center (ESALQ, Piracicaba, São Paulo, Brazil). Sequencing adapters and low complexity reads were removed with SeqyClean software (https://github.com/ibes t/seqyclean), and quality control was carried out using the FASTQC software version 0.10.1 (https://www.bioinformatics.babraham.ac. uk/projects/fastqc/). The RNA-Seq data is available on the European Nucleotide Archive (ENA) under the following accession codes: PRJEB13188, PRJEB10898, and PRJEB19421. The RNA-Seq data were mapped to the reference genome ARS-UCD1.2 [21] using the STAR software (v2.7.8) [22] with the default parameters. The bam files were used as inputs for the

WASP pipeline [23] to find and remove duplicated reads and the ones showing mapping bias.

2.3.3 DNA extraction and genotyping

DNA extraction and quality control were executed as previously described [24]. Genotypes were called in the Illumina GenomeStudio software and reported in previous studies [19,24]. BovineHD genotypes were phased using Eagle [25] and imputed using the Minimac3 program [26] as described in detail previously [27].

Briefly, the major panel used in SNP imputation was constructed with whole-genome sequence data from 26 progenitors of our subset population, processed according to the 1000 Bulls Genomes Project (http://www.1000bullgenomes.com/) guidelines. PLINK software (v1.9) [28] was used to filter the SNPs selecting only informative ones, *i.e.*, autosomal and with minor allele frequency (MAF) > 0.01. To compute the imputation accuracy, a leave-one-out cross-validation method was applied. To do this, one progenitor was sorted out and included as a target sample for imputation, considering the same SNP density as the progeny. This process was repeated for the 26 progenitors. The ratio between the number of incorrect imputed alleles and the total of alleles imputed in the tested sample is the allelic imputation error rate. The correlation between the true and imputed genotype is the imputation accuracy. For quality control, we used the PLINK software. SNPs showing imputation accuracy <0.98 or allelic imputation error rate >0.02 were filtered out. Additionally, only autosomal SNPs with a minor allele frequency higher than 0.01 were kept for further analysis [27]. All analyses were carried out using the ARS-UCD1.2 assembly of the bovine reference genome [21].

In this work, the filtered imputed genotypes were phased using Beagle (v5.2) [29], following the default parameters. The SNP files were indexed with BCFTools (v0.1.19) [30], concatenated, and sample filtered with VCFTools (v0.1.15) [31]. The resulting file contained 4,812,912 SNPs and 190 animals. We obtained the linkage disequilibrium (LD) blocks from this genotype file using the PLINK software (v1.9) [28], with the default parameters. The method for accounting for the LD was the D prime (D′), *i.e.*, the normalized measure of allelic association, which estimates the recombination history between pairs of SNPs [32].

We used the following parameters: testing the LD blocks with variants up to 200 kb distant from each other, with MAF ≥ 0.05. As the D′ method did not consider the variant frequency, the MAF threshold was applied to avoid inflation of LD with rare variants, as better described previously [32]. Considering a 90% D′ confidence interval, strong LD was observed between a D′ lower bound above 0.7 and an upper bound with at least 0.98. Upper bounds <0.9 indicate strong evidence of historical recombination.

2.3.4 ASE analysis

The major and minor alleles counts were obtained from heterozygous SNPs using GATK ASE Read Counter from GATK v4.2 [4]. To avoid inflation of allelic ratios in low expressed SNPs, we required ten reads that mapped an SNP to test this variant in ASE analysis. Throughout the manuscript we consider reference alleles to be the most frequent in our Nelore population (*i.e.*, major alleles), and alternative alleles are those least frequent (*i.e.*, minor alleles). Using an R script (R version: 4.1.1), major and minor alleles' counts for each SNP were compared per animal against the expected value of 0.5 (a scenario with a balanced allelic expression ratio) using the Binomial Test, which is a traditional test for ASE analysis [33]. We used the Benjamini-Hochberg method to correct the False Discovery Rate (FDR) of the *p*-values. The tests resulting in FDR ≤ 0.05 were considered significant, and the respective variant was classified as an ASE SNP.

2.3.5 aseQTL association

For the aseQTL analysis, we searched for SNPs associated with the allelic imbalance pattern of ASE SNPs in a window of 200 kb upstream and downstream to the ASE SNPs position, since aseQTLs were mostly located within this range in our previous study [16]. Ten heterozygous samples for each ASE SNP and ten heterozygous samples for each SNP located in the genomic range to be tested (candidate SNP) were required. The tests were executed per ASE SNP window, comparing the genotypes for the candidate SNPs to the allelic expression imbalance (AEI) of the ASE SNP, measured per sample, as follows: AEI $=\left|\frac{ref}{ref+alt} - 0.5\right|$ where:

ref are the counts that mapped the major allele and *alt* correspond to the number of counts that mapped the minor allele. When AEI tends to zero, it indicates minimal ASE. When AEI approximates 0.5, it corresponds to a high ASE level. The AEI of ASE SNPs from homozygous and heterozygous samples for the candidate SNP were compared using a Wilcoxon Rank Sum Test. The resulting *p*-values were corrected for FDR using the Benjamini- Hochberg method. The significant aseQTLs were the candidate SNPs in which the heterozygous group showed a higher allelic imbalance than the homozygous group [12,16], with FDR ≤ 0.05.

2.3.6 cis-eQTL screening

We performed a *cis*-eQTL analysis to evaluate the ASE SNPs and aseQTLs effect on the related transcripts' expression. For this analysis, we selected transcripts of all ASE genes from the bovine annotation file downloaded from Ensembl release 103. Salmon software (v1.8.0) [34] adopting full decoy was used to quantify the abundance of all transcripts. The abundances were associated with the genotypes of the related ASE SNPs and aseQTLs. For the association analysis, we used a linear regression model implemented by MatrixEQTL R package (v2.3) [35], including sequencing lane and contemporary group (farm of origin, birth year, and slaughter date) as fixed effects and age at slaughter as a covariate, according to a previous study [20].

The tests were performed pairwise between all transcripts from the same ASE gene and the respective ASE SNPs or aseQTLs. The resulting β values correspond to the slope coefficient, which can be interpreted as the effect size in expression caused by the increase of minor alleles of the tested SNP, considering the remaining variables as constant. If the beta value is positive, the minor allele increases the transcription, and if the beta value is negative, the minor allele decreases the given transcript's expression. The additive effect of genotype on transcripts' abundance was considered significant when it resulted in FDR ≤ 0.05. The eQTLs were classified as *cis*-eQTLs for SNPs located until 1 Mb of genomic distance from the regulated transcript.

2.3.7 SNP functional annotation and gene enrichment analysis

The annotation of ASE SNPs, aseQTLs, and *cis*-eQTLs was conducted using the SnpEff software (v4.5) [36], based on the ARS-UCD1.2 bovine reference genome. ASE genes were functionally annotated in the ClueGo plugin for Cytoscape (v3.9.1) [37]. We searched for enriched biological processes and KEGG metabolic pathways based on the bovine annotation as background and tested the enrichment significance with a right-sided hypergeometric test. Biological Processes and metabolic pathways with FDR ≤ 0.05 were considered significant.

2.*3.8. Data integration*

Our results were integrated with various datasets to explore the relevance of our ASE-related variants (ASE SNPs, aseQTLs, and *cis*-eQTLs) in different aspects: i) the consistency of our results and previous ASE and imprinting data; ii) the effect of ASE genes in relevant livestock phenotypes, and iii) the regulatory potential of our findings. Firstly, we compared all the ASE results found here with ASE SNPs and ASE genes [7,14], aseQTLs [16], *cis*-eQTLs [20], and imprinted genes previously described. The ASE data used in this integration were based on muscle from a *Limousin* cattle population [7] and the same Nelore samples used here, but analyzed previously using a lower density SNP data produced with the Illumina BovineHD BeadChip (770 K) [14].

The aseQTL and *cis*-eQTL results from the same Nelore population were retrieved from previously published papers [16,20]. The imprinted genes were downloaded from the Genomic Imprinting database [38]. The ASE SNPs were also overlapped with CNV regions reported previously in the same population [17] to investigate if the allelic imbalance in these regions resulted from CNVs. Data obtained in the UMD3_1 genome assembly had its original coordinates converted to the ARS-UCD1.2 reference genome using UCSC LiftOver [39]. The second integration approach was performed to relate ASE genes with relevant traits. The identified ASE genes were compared with differentially expressed genes (DEGs) from studies in the same Nelore population. The overlapped DEGs were related to mineral concentration [40,41], carcass quality [42], beef tenderness [43], fatty acid,

and intramuscular fat content [44,45] traits. In the last integration analysis, we searched for regulatory evidence acting on ASE-related SNPs.

We integrated the genomic positions of ASE SNPs, aseQTLs, and *cis*-eQTLs with bovine epigenetic data downloaded from the pilot project of the Functional Annotation of Animal Genomes (FAANG) consortium [46]. We retrieved the ATAC-seq and ChIP-seq data of the two available bovine muscle samples from the NCBI database (Series accession code: GSE158430). The targets of ChIP-seq data obtained by Kern et al. used for comparison were H3K4me3, H3K27me3, H3K27ac, and H3K4me1 histone modifications and the binding protein CTCF. We also downloaded the *cis*-eQTL data of bovine muscle from the cattle GTEx database (https://cgtex.roslin.ed.ac.uk/downloads/) and overlapped it with the genomic position of our ASE SNPs, *cis*-eQTLs, and aseQTLs. The ASE genes were overlapped with those putatively differentially affected by the H3K27me3 histone modification in contrasting mineral sample groups [47] and genes with differentially methylated regions and cytosines [48] obtained in the same Nelore population. We evaluated the significance of the integration using permutation tests, comparing the number of overlaps of ASE-related SNPs in each functional region to the respective number of overlaps obtained from a random resampling of 1000 SNPs from the original dataset, with the regioneR package [49].

2.4 Results

2.4.1. ASE and monoallelic expression are widespread in Nelore muscle

From the 190 samples studied here, there was an average of 95.72 % (ranging from 87.85 % to 97.10 %) of uniquely mapped RNA-seq reads in the bovine reference genome (ARS-UCD1.2). The WASP software filtering removed 6.9 million duplicated reads or those presenting mapping bias on average. As ASE analyses required heterozygous SNPs, this filtering reduced the original genotype file from 4,812,912 SNPs to 1,353,848 SNPs. To select SNPs with considerable expression in our population, we required at least ten total counts of mapped reads for ASE analysis. Thus, 40,969 SNPs were used to test the allelic expression imbalance, covering 0.85 % of our original genotype data.

Significant ASE was frequent in our tested dataset. From the 701,032 tests performed, 427,333 (corresponding to 60.90 % of the tests) were significant (FDR < 0.05), comprising 38,177 unique ASE SNPs (Supplementary Table 1; Fig. 2.2A). Considering all ASE SNPs, 37,128 were distributed in 7304 LD blocks. The highest number of ASE SNPs were located in the BTA3 (Fig. 2.2B). The ASE SNPs identified per chromosome were not significantly correlated with chromosome length.

Although the identification of 38,177 ASE SNPs is an expressive proportion of all SNPs tested (93.00% of the 40,969 tested SNPs), ASE is analyzed per sample, so multiple ASE SNPs were significant in a few animals. For example, 7671 ASE SNPs were identified in only one sample (Fig. 2.3A). The average number of significant animals for each ASE SNP was 11, ranging from one to 96 (Fig. 2.3A), and all animals had at least one ASE SNP identified. Monoallelic expression was widespread among the samples (Fig. 2.4B). window of 1 Mb. B) SNPs tested for ASE, significant and non-significant distributed per chromosome.

On average, 10.55 samples were monoallelic for the ASE SNPs, ranging from one to 96 animals. The major alleles were more expressed than the minor alleles in approximately 90.31 % of the significant tests. The major allele frequency was predominant in the population, with an average MAF of 0.79 (ranging from 0.42 to 0.99, Fig. 2.3C). A total of 37,943 (99.38 %) of the ASE SNPs showed monoallelic expression in at least one sample, from which 37,227 SNPs expressed only the major allele. On the other hand, 8513 SNPs were monoallelic and exclusively expressed the minor allele.

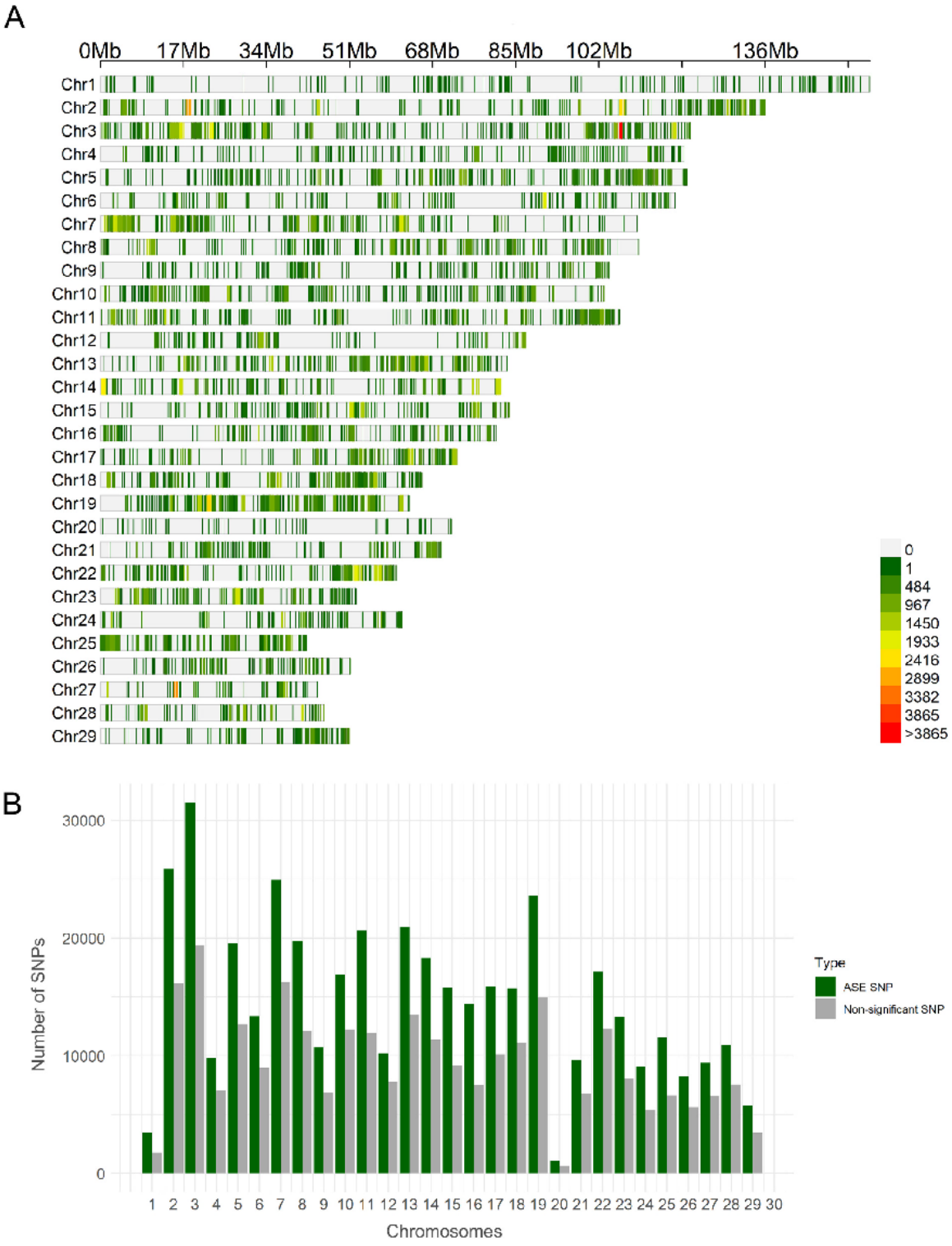


Figure 2.2 - Chromosomal distribution of ASE SNPs. A) Distribution of ASE SNPs. The segment colors correspond to the number of ASE SNPs located in awindow of 1 Mb. B) SNPs tested for ASE, significant and non-significant distributed per chromosome. Source: Author's

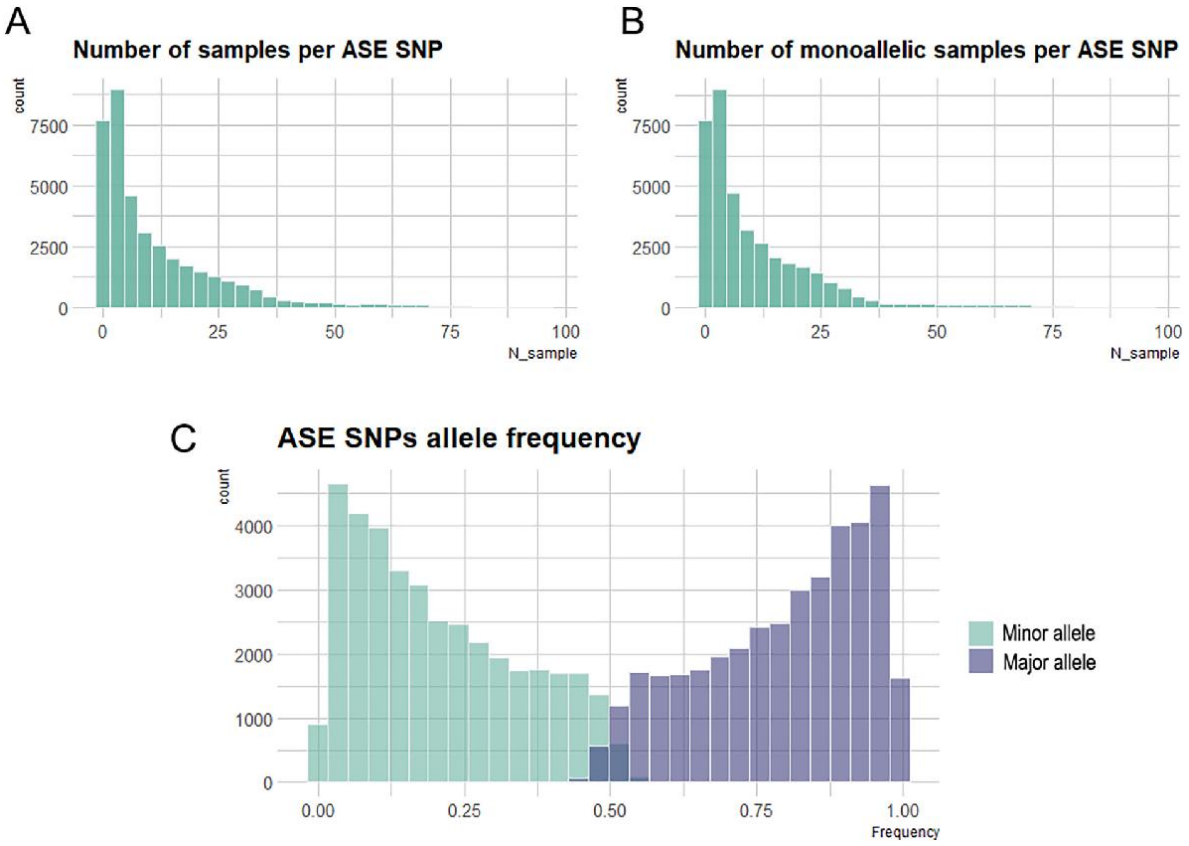


Figure 2.3 - ASE SNPs alleles' and samples' frequency. A) Histogram of the number of significant samples for the ASE SNPs. B) Histogram of the number of monoallelic samples for the ASE SNPs. C) Major and minor allele frequency of ASE SNPs. Source: Author's

2.4.2 Annotation of ASE SNPs and genes predicted regulation of livestock-relevant metabolic pathways

The ASE SNPs annotation by SnpEff software [36] revealed that the ASE SNPs were classified predominantly as downstream gene variants (26.70 %), followed by intron variants (22.70 %), synonymous variants (20.30 %), and 3′UTR variants (10.40 %), corresponding to 80.10 % of all predicted SNP effects (Supplementary Table 1; Fig. 2.4). Some ASE SNPs were predicted to result in high-impact consequences, such as the SnpEff specification (http://pcingola.github.io/SnpEff/se_inputoutput /#effect-prediction-details). The affected genes and the respective ASE SNPs effects were: *MBNL1* (stop lost); *EMCN*, *SQSTM01*, and *EIF3CL* (splice donor); *AGO4* and *RRP7A* (splice acceptor);

and *CSF1*, *ZNF266*, *SHPRH*, *PPM1B*, *ENSBTAG00000049014*, *OR51D1*, *ENSBTAG00000027426*, *CD46*, *ENSBTAG00000000560*, *XAF1*, *NRADD*, *PSMB8*, and *ANKRD42* (stop gained).

The ASE SNPs are within 7448 genes (Supplementary Table 2), corresponding to 52.38 % of all muscle-expressed genes of the bovine genome (*N* = 14,219). The average ASE SNPs per ASE gene was 5.12, ranging from one to 166 variants. The top five ASE genes with more ASE SNPs were *MYOM2* (166 ASE SNPs), *MBP* (133 ASE SNPs), *SVIL* (117 ASE SNPs), *PDE4DIP* (104 ASE SNPs), and *SNTB1* (102 ASE SNPs).

We considered putative monoallelically expressed genes the ones showing all the significant samples expressing only one allele for all the identified ASE SNPs. Thus, a total of 4354 ASE genes were classified here as potential monoallelically-expressed genes in Nelore muscle (Supplementary Table 2). *TIMP4* was the gene with more monoallelic SNPs in the population, expressing only the major allele in 22 ASE SNPs and 69 samples. A similar scenario is observed for the other two genes with more monoallelic expression, *i.e.*, *HNRNPU* (31 monoallelic ASE SNPs in 75 samples) and *DHX30* (21 monoallelic ASE SNPs in 78 samples). Analysis concerning the gene ontology of the ASE genes identified 1742 enriched biological processes (FDR ≤ 0.05) (Supplementary Table 3). The KEGG pathway enrichment retrieved 151 significant metabolic pathways (FDR ≤ 0.05), including protein processing in the endoplasmic reticulum, autophagy, endocytosis, focal adhesion, MAPK signaling pathway, and proteasome (Supplementary Table 4).

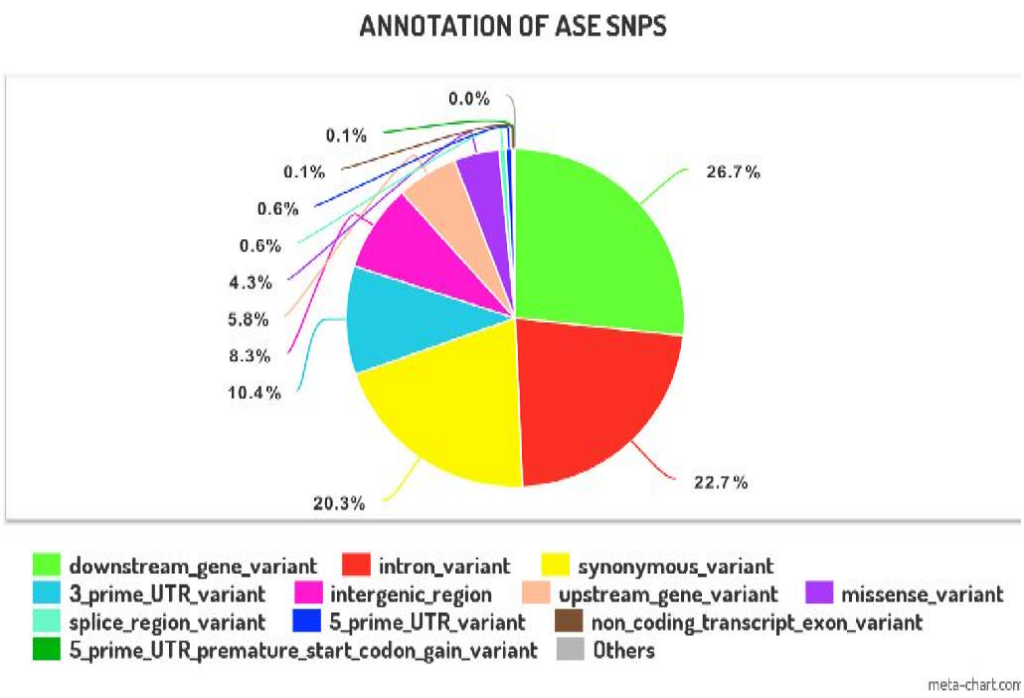


Figure 2.4 - Pie plot showing the annotation of ASE SNPs. Source: Author's

2.4.3 aseQTLs were associated with the allelic imbalance of ASE SNPs

The aseQTL analysis searched for potential *cis*-regulatory SNPs associated with the allelic imbalance observed in the ASE SNPs. Thus, we searched for aseQTLs located in a window of 200 kb around ASE SNPs that have more than ten heterozygous samples, encompassing 14,384 ASE SNPs. We performed 7,266,276 tests between these ASE SNPs and 1,286,188 SNPs to be tested as aseQTLs. In total, 26,956 tests were significant, discovering 21,543 aseQTLs for 430 ASE SNPs (FDR ≤ 0.05) (Fig. 2.5A; Supplementary Table 5). A total of 21,448 aseQTLs were distributed in 1174 LD blocks.

The average of aseQTLs found for each ASE SNP was 62.68, ranging from one to 420. The ASE SNP associated with more aseQTLs was the chr16:43573041G>A, located in the *NMNAT1* gene. Another ASE SNP within the same gene (chr16:43573035G>A) was associated with 414 aseQTLs, 413 in common to the aseQTLs associated with chr16:43573041G>A, and located in five LD blocks were the sum of SNPs from two blocks corresponded to 84 % of the

overlapping aseQTLs. The second ASE SNP with more aseQTLs was the chr13:66887795A>G, within the *TTI1* gene, with 417 associated aseQTLs. AseQTLs can regulate more than one ASE SNP. The average number of ASE SNPs per aseQTLs was 1.25, ranging from one to six. Forty aseQTLs located in the chr3:106229134–106243323 window were associated with six ASE SNPs in the same chromosome within *NT5C1A* and *HEYL* ASE genes. The majority of the aseQTLs were located in intronic regions (50.4 %) (Supplementary Table 5), and the following aseQTLs were found in high-impact regions: chr3:109965793T>C (splice acceptor); chr3:14913466T>A (start lost), and chr14:441812G>A (stop gained). In addition, we observed that 673 aseQTLs were also identified as ASE SNPs, *i.e.*, chr3:109965793T>C (splice acceptor variant), located in the gene *AGO4*, associated with the ASE SNP chr3:109879430T>C, from *AGO1* gene.

2.4.4 ASE SNPs and aseQTLs can affect transcripts abundance

To search for ASE SNPs and aseQTLs with regulatory potential on variants and transcripts of ASE genes. We found 3333 significant associations (FDR ≤ 0.05), where 3107 *cis*-eQTLs (*i.e.*, 2098 ASE SNPs and 1075 aseQTLs, with 66 identified as both ASE SNPs and aseQTLs) were associated with the expression of 758 transcripts. These transcripts correspond to 696 ASE genes, 9.34 % of all ASE genes identified in this study (Supplementary Table 6). The *cis*-eQTLs were mainly predicted as intron variants (47.15 %), and downstream gene variants (23.49 %), being 141 *cis*-eQTLs predicted to cause missense variation, 15 *cis*-eQTLs located in splice regions, six *cis*-eQTLs predicted to be related to a premature start codon gain, three *cis*-eQTLs were stop gained variants, and one was annotated as a splice donor variant.

Table 2.1 presents the annotation of *cis*-eQTLs, with the total number of *cis*-eQTLs per effect and the respective classification as ASE SNPs or aseQTLs. For 2294 *cis*-eQTLs, the minor allele was predicted to upregulate the expression of the associated transcript, and for 1039, the minor allele downregulates the transcript expression. *Cis*-eQTLs with the larger absolute effects sizes were mainly classified as ASE SNPs.

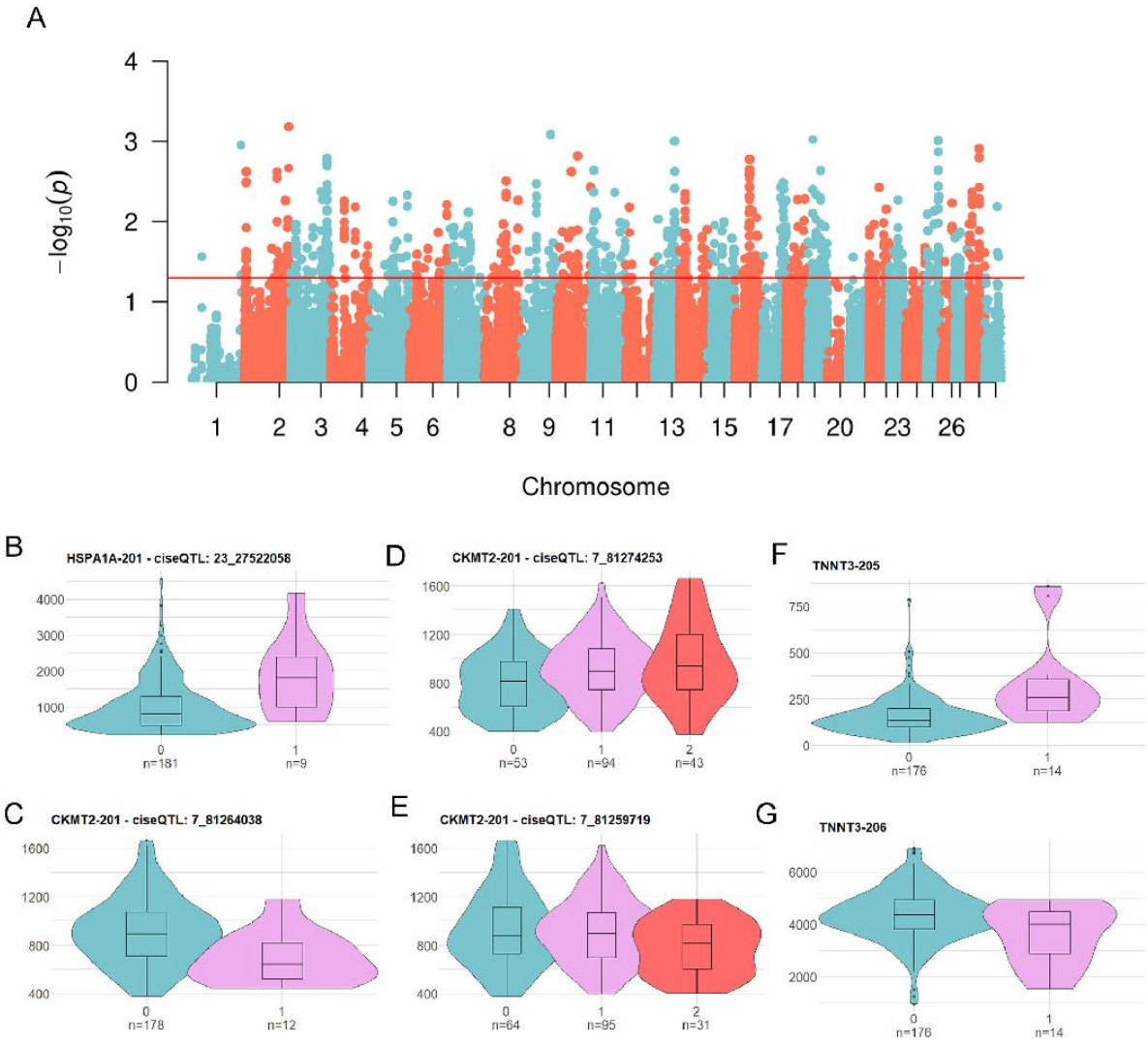


Figure 2.5 - AseQTL results and *cis*-eQTL effects on selected transcripts. A) Manhattan plot with aseQTL analysis results. The points correspond to the executed tests, the x-axis corresponds to chromosomes and genomic position, and the y-axis corresponds to FDR-adjusted *p*-values. Above the red line were the tests corresponding to significant aseQTLs (FDR ≤ 0.05). B–G) Violin plots representing the transcripts abundance statistics according to the samples' genotypes for selected *cis*-eQTLs, being “0” the major homozygotes, “1” the heterozygotes, and “2” the minor homozygotes. B) Violin plot comparing *HSPA1A-201* abundance according to chr23:27522058G>A genotypes. C) Violin plot comparing *CKMT2-201* abundance according to chr7:81264038G>A genotypes. D) Violin plot comparing *CKMT2-201* abundance according to chr7:81274253T>C genotypes. E) Violin plot comparing *CKMT2-201* abundance according to chr7:81259719G>A genotypes. F) Violin plot comparing *TNNT3-205* abundance according to each of the 13 significantly associated *cis*-eQTLs*. G) Violin plot comparing *TNNT3-206* abundance according to each of the 13 significantly associated *cis*-eQTLs* (the same associated to *TNNT3-205* abundance).*all *cis*-eQTLs have the same genotype in the sample, producing the same graph and results.

Table 2.1 - Predicted effects of *cis-eQTLs* found as ASE SNPs and aseQTLs.

Effect	ASE SNP	aseQTL	Total
3′ UTR	465	29	494
5′ UTR premature start codon gain (start gained)	3	1	4
5′ UTR	30	4	34
Downstream	451	122	573
Intergenic region	1	224	225
Intron	330	478	808
Missense	113	8	121
Missense and splice region	2	0	2
Non-coding transcript exon	6	1	7
Splice donor and intron	1	0	1
Splice region	2	0	2
Splice region and intron	4	0	4
Splice region variant and synonymous	7	0	7
Stop gained	3	0	3
Synonymous	549	58	607
Upstream gene	131	200	331

Some *cis*-eQTLs were associated with more than one transcript. The chr3:33707890A>G, chr3:33708844G>A, and chr3:33708874C>T *cis*-eQTLs were associated with four transcripts of the *GSTM1*, located in the 3′UTR region of this ASE gene. The chr7:14983829T>G, chr2:107492306C>G, chr26:23086765C>T, and chr9:67736789A>G *cis*-eQTLs were associated with three transcripts each of *PDE4A*, *OBSL1*, *MFSD13A*, and *ARHGAP18* genes, respectively. A total of 209 *cis*-eQTLs was associated with different transcripts pairs. The average number of *cis*-eQTLs identified per transcript was 4.39, ranging from one to 421 *cis*-eQTLs. The transcript *NMNAT1-201* had more *cis*-eQTLs, all of them distributed in five LD blocks. *KLC1-201* had 184 *cis*-eQTLs in six LD blocks and *UBE3B-202* with 116 *cis*-eQTLs in 12 LD blocks. For these three transcripts, the *cis*-eQTLs were mainly identified as aseQTLs. The chr23:27522058G>A *cis*-eQTL minor allele had the higher positive effect size (β = 862.80), associated with the transcript *HSPA1A-201* (Fig. 2.5B), followed by *TNNT3-205*, *CKMT2-201*, *RPLP0-201*, and *MSRB1- 201*. The *CKMT2-201* abundance had *cis*-eQTLs with negative effects associated with the minor allele. The minor allele of chr7:81264038G>A *cis*-eQTL showed

association with a notable decrease of *CKMT2-201* abundance (*β* = - 204.56) (Fig. 2.5C). This same transcript was positively regulated by the minor allele of chr7:81274253T>C *cis*-eQTL (*β* = 98.88) (Fig. 2.5D). Thirteen *cis*-eQTLs with the same phased genotypes were negatively associated with *CKMT2-201* transcript levels (*β* = - 69.63), as represented in Fig. 2.5E for the *cis*-eQTL chr7:81259719G>A.

Thirteen *cis*-eQTLs with identical genotypes located between chr29:49576764–49594787 were associated with two different transcripts of the *TNNT3* gene, with its minor allele upregulating *TNNT3-205* (*β* = 111.44) and downregulating *TNNT3–206* (*β* = - 823.08) (Fig. 2.5F and G, respectively). From these 13 *cis*-eQTLs, five were within the same LD block, being intronic and synonymous variants. The remaining were not located in any LD block. Fig. 2.6 shows the distribution of transcripts of the gene *TNNT3*. This reverse-strand gene has *cis*-eQTLs classified as downstream gene variants before the 49.58 Mb position. The two transcripts have similar exon positions, except for the 1st, 2nd, 3rd, and 5th exons of *TNNT3-206*, absent in the *TNNT3-205* transcript. The mentioned LD block containing five *cis*-eQTLs is located near the 49.59 position, encompassing the 2nd exon of the *TNNT3-205* and between the 4th and 5th exons of the *TNNT3-206*.

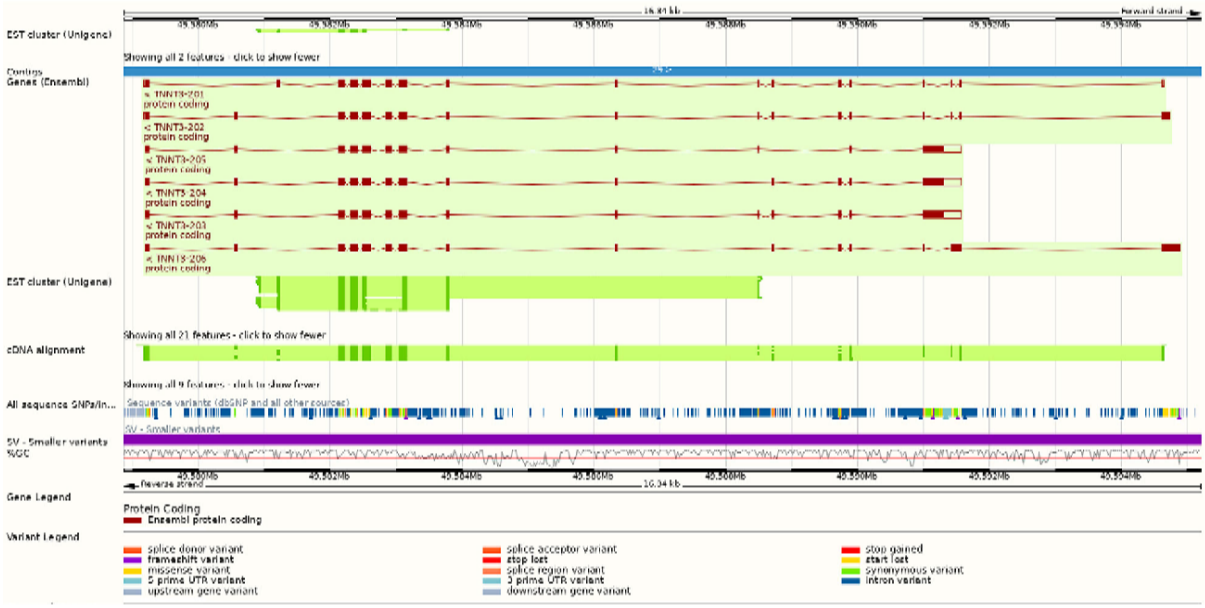


Figure 2.6 - Genome view of *TNNT3* transcripts. Source: Author's.

2.4.5. ASE data showed consistency with the literature

A previous study within the same Nelore population [14] identified 820 ASE SNPs from Illumina BovineHD BeadChip (770 K), which overlapped with 479 ASE SNPs identified in this study. The overlapping ASE SNPs corresponded to 58.4 % of those identified by De Souza et al. [14] and represent 1.25 % of the ASE SNPs identified here. ASE SNPs identified in the muscle of a Limousin (*Bos taurus*) population [7] overlapped with 1020 of our ASE SNPs, including 75 reported by De Souza et al. [14] as well. Eleven ASE SNPs were previously identified as aseQTLs, from the chip dataset [16]. Regarding the ASE genes, 482 overlaps were found with data obtained by De Souza et al. [14], corresponding to 90.94 % of the ASE genes identified in that study. Thus, with the use of the imputed dataset, we identified 6966 novel ASE genes in Nelore muscle compared to De Souza's previous work Furthermore, 1699 ASE genes identified here were also identified by Guillocheau et al. [7] in the muscle of Limousin cattle, corresponding to 80.17 % of their ASE genes. As 229 ASE genes were found in common for these three studies of ASE in bovine muscle (Supplementary Table 2), this work identified 5496 novel ASE genes which were not found in the previous studies in Nelore and Limousin cattle.

Fig. 2.7 shows the number of overlaps between different studies about ASE data obtained in bovine muscle. As allelic imbalance can result from genomic imprinting, we compared the ASE genes with those predicted or experimentally validated showing imprinting in various species. We identified 51 ASE genes potentially affected by imprinting; seven confirmed in *Bos taurus*: *SGCE*, *PEG10*, *MEST*, *IGF2R*, *GNAS*, *NNAT*, and *TSSC4* (Supplementary Table 2). The *IGF2R* and *NNAT* ASE genes showed 72.7 % and 68.6 % of significant samples showing monoallelic expression, respectively. The other five bovine imprinted genes overlapping our ASE genes showed monoallelic expression for all the corresponding ASE SNPs. Conversely, from the 4354 monoallelic genes in our study, 4303 were not in the Gene Imprint database [38]. We also integrated ASE genes and differentially expressed genes (the DEGs data) to evaluate their biological relevance for important livestock traits.

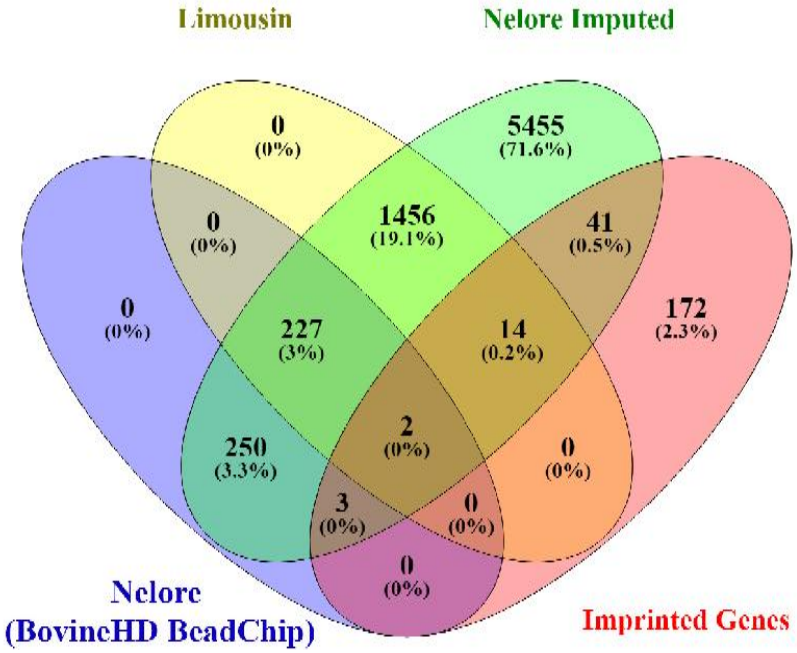


Figure 2.7 - Venn diagram representing the integration of ASE genes studied in bovine muscle. Source: Author's.

We found 1823 (1450 unique) ASE genes integrating with DEGs previously identified in the muscle of the same Nelore population and related to 26 traits across eight studies (Supplementary Table 2). Of these, 1158 ASE genes overlapped with DEGs between contrasting samples for four fatty acid content traits: CLA-c9t11, oleic, eicosapentaenoic, and palmitic acids [44]; 399 for Fe [41], Cu, Mn, P, Na, Se, Ca, Mg, K, Zn and S [40] minerals' concentration; 93 ASE genes integrated with DEGs for three feed efficiency traits: residual feed intake, average daily gain and dry matter intake [44]; 79 were common to DEGs for ribeye area (REA) and backfat thickness (BFT) traits [42]; 45 ASE genes were identified as DEGs for intramuscular fat (IMF), and seven for tenderness measured 14 days after slaughter (WBSF14) [43]. In addition, 2043 ASE SNPs were located within CNV regions identified in the same population in previous work, [17], corresponding to 5.35 % of all ASE SNPs (Supplementary Table 1). The CNV data used for this comparison was predicted using the UMD 3.1 bovine genome assembly [17], thus, the genomic coordinates were converted with LiftOver to the ARS-UCD1.2 version. As this process may have

some error rates, we choose not to discard the ASE SNPs within these regions. The overlapping between the CNV regions and the ASE SNPs is shown in Supplementary Table 1.

2.4.6 Integration with functional data

The effect of the ASE-related variants on ASE or total gene expression can result from epigenetic regulation. Studies exploring potential regulatory mechanisms affecting the gene expression in the same population have been conducted [16,47,50]. By integrating the aseQTLs found here using the imputation dataset and our previous aseQTL data obtained from the Illumina Bovine*HD* BeadChip [16], we found eight overlaps, being these aseQTLs associated with the allelic imbalance of six ASE SNPs (Supplementary Table 5).

De Souza et al. [48] used reduced representation bisulfite sequencing (RRBS) to determine the differential methylation cytosines (DMCs) and regions (DMRs) between two groups contrasting for the tenderness phenotype, with six muscle samples each. We identified 33 ASE genes containing DMCs and DMRs, indicating that the ASE genes can be regulated by DNA methylation and probably underlie beef tenderness. In contrasting muscle samples for mineral content phenotypes, Afonso et al. (personal communication) reported a list of genes predicted to be regulatory for these phenotypes using a methodology based on the predicted influence of the H3K27me3 over them.

Comparing the list of genes potentially regulating muscle mineral concentration, we found 396 overlaps, affecting 116 genes and the Cu, Mg, S, Zn, K, P, Na, Fe, Se, and Ca content (Supplementary Table 2). Considering the ASE genes integrating differential methylation data, the percentage of monoallelic samples per gene ranged from 78.4 % to 100 %, with 23 genes with 100 % of significant samples showing monoallelic expression (Supplementary Table 2). The ASE genes overlapping the putative H3K27me3 data showed a monoallelic expression percentage ranging from 50% to 100%, from which 68 genes had monoallelic expression in all significant samples. We found evidence of regulatory variants co-located with our ASE-related variants (ASE SNPs, *cis*-eQTLs, and

aseQTLs). Regarding the comparisons with *cis*-eQTLs of bovine muscle from the cGTEx database, we found 11,493 ASE SNPs, 5559 aseQTLs, and 1045 of our *cis*-eQTLs co-located with the same genomic location of these published *cis*-eQTLs (Supplementary Tables 1, 5 and 6, respectively).

Our findings were also compared to epigenetics data from a pilot project from the FAANG consortium [46]. The number and proportion of SNPs overlapping accessible chromatin regions or ChIP-seq data are presented in Table 2.2. We evaluated the significance of this integration to each set of epigenetic data using a permutation test using a resampling approach with 1000 randomizations. The overlaps between our data and these datasets were more frequent than by chance, with a permutation *p*-value of 0.001, and the following *Z*-scores: 72.856 (ATAC-seq), 33.422 (CTCF), 53.423 (H3K4me1), 61.883 (H3K4me3), 89.573 (H3K27ac) and 38.742 (H3K27me3). The permutation with the lowest and the highest Z-scores are shown in Fig. 2.8A and 2.8B, respectively. We selected the more promising regulatory aseQTLs based on the predicted effect, filtering out SNPs located in low effect regions and those not overlapping epigenetics marks in Table 2.3. Table 2.4 shows the *cis*-eQTLs with the larger effect size (slope coefficient) that overlapped with the epigenetic data. All integration results are presented in Supplementary Tables 1, 5, and 6, for ASE SNPs, aseQTLs, and *cis*-eQTLs, respectively.

Table 2.2 - Number of ASE-related SNPs overlapping functional regions from Functional Annotation of ANimal Genomes (FAANG) consortium.

ASE-related SNP	**ACR**	**CTCF**	**H3K27ac**	**H3K27me3**	**H3K4me1**	**H3K4me3**
ASE SNPs	4714 (12%)	1021 (2.6%)	2899 (7.6%)	995 (2.6%)	797 (2%)	1250 (3.2%)
aseQTLs	1503 (6.9%)	359 (1.6%)	825 (3.8%)	433 (2%)	254 (1.1%)	330 (1.5%)
cis-eQTLs	356 (10.6%)	87 (2.6%)	226 (6.7%)	112 (3.3%)	93 (2.7%)	118 (3.5%)

Table 2.3 - aseQTLs with moderate and high predicted effects that overlapped epigenetics data from Annotation of ANimal Genomes (FAANG) consortium.

aseQTL	Effect	Associated ASE SNPs	ATAC-Seq	Target_ChIPseq
chr15:45219900	missense_variant	chr15:45052156G>A	ACR	CTCF, H3K27ac
chr15:45219923	missense_variant	chr15:45052156G>A	ACR	CTCF, H3K27ac
chr15:45220100	missense_variant	chr15:45052156G>A	ACR	CTCF, H3K27ac
chr23:28407910	5_prime_UTR_premature_start_codon_gain_variant	chr23:28402954G>C	ACR	CTCF, H3K27ac, H3K4me1, H3K4me3
chr21:68308892	missense_variant	chr21:68298154A>G	NA	H3K27ac
chr3:16468730	splice_region_variant&synonymous_variant	chr3:16332848A>G	ACR	H3K27ac
chr3:16468730	splice_region_variant&synonymous_variant	chr3:16446714A>C	ACR	H3K27ac
chr3:16468730	splice_region_variant&synonymous_variant	chr3:16344888C>T	ACR	H3K27ac
chr3:16468730	splice_region_variant&synonymous_variant	chr3:16497319C>T	ACR	H3K27ac
chr22:17777164	splice_region_variant&intron_variant	chr22:17926204G>A	NA	H3K27ac, H3K27me3, H3K4me1, H3K4me3
chr13:46434080	missense_variant	chr13:46480257T>C	ACR1	H3K27ac, H3K4me3
chr13:46434083	missense_variant	chr13:46480257T>C	ACR	H3K27ac, H3K4me3
chr14:441812	stop_gained	chr14:313854C>G	ACR	H3K27ac, H3K4me3
chr13:51985282	splice_region_variant&intron_variant	chr13:519994006C>T	ACR	NA
chr17:51788227	5_prime_UTR_premature_start_codon_gain_variant	chr17:519666674G>T	ACR	NA
chr17:55726121	splice_region_variant&intron_variant	chr17:557833791G>T	ACR	NA

chr17:55726121	splice_region_variant&intron_variant	chr17:55783782C>T	ACR	NA
chr19:55342674	splice_region_variant&intron_variant	cht19:55194508T>C	ACR	NA
chr3:49790464	splice_region_variant&intron_variant	chr3:49612851A>G	ACR	NA
chr5:116603911	missense_variant	chr5:116508632A>G	ACR	NA
chr7:5726429	5_prime_UTR_premature_start_codon_gain_variant	chr7:5868232T>C	ACR	NA
chr8:84596628	splice_region_variant&intron_variant	chr8:84622947G>A	ACR	NA

Table 2.4 - *cis*-eQTLs with effect sizes higher than |20| overlapping epigenetics data from Annotation of ANimal Genomes consortium.

***cis*-eQTL**	**SNP_type**	**Beta**	**Transcript**	**ATAC-seq**	**ChIP-seq**
15_17267662	ASE SNP	−311,1428661	*SLN-201*	ACR	NA
17_56161474, 17_56162192, 17_56162417, 17_56162443, 17_56162457, 17_56167643, 17_56168683	ASE SNP	−114,2005503	*HSPB8-201*	ACR	H3K27ac
18_49102944	ASE SNP	28,55763824	*ZFP36-201*	ACR	H3K27ac, H3K27me3, H3K4me1, H3K4me3
18_49103134	ASE SNP	28,55763824	*ZFP36-201*	NA	H3K27ac, H3K27me3, H3K4me1, H3K4me3
19_51260226	ASE SNP	28,93826906	*ACTG1-201*	ACR	H3K27ac, H3K27me3, H3K4me1, H3K4me3

2_107407598	ASE SNP	66,1226443	*SPEG-201*	ACR	NA
22_48945156	ASE SNP	−39,5816066	*RPL29-201*	ACR	CTCF, H3K27ac, H3K27me3, H3K4me1, H3K4me3
23_27522058	ASE SNP	862,8080538	*HSPA1A-201*	ACR	H3K27ac, H3K27me3, H3K4me1, H3K4me3
25_1500564, 25_1500749, 25_1500883	ASE SNP	67,80534341	*MSRB1-201*	ACR	CTCF, H3K27me3, H3K4me3
26_13209693	ASE SNP	35,78627411	*PPP1R3C-201*	ACR	H3K27ac, H3K4me3
29_49580452, 29_49580473	ASE SNP	111,4468683	*TNNT3-205*	ACR	NA
29_49580452, 29_49580473	ASE SNP	−823,0812458	*TNNT3-206*	ACR	NA
29_49594787	ASE SNP	111,4468683	*TNNT3-205*	ACR	NA
29_49594787	ASE SNP	−823,0812458	*TNNT3-206*	ACR	NA
3_16768178, 3_16768207, 3_16768337	ASE SNP	22,36054315	*S100A14-201*	NA	H3K27ac
5_103581376	ASE SNP	57,93593094	*TPI1-201*	NA	CTCF, H3K27ac
5_25510849	ASE SNP	−44,18357512	*PPP1R1A-202*	NA	H3K27ac
5_25515097	ASE SNP	−46,64203616	*PPP1R1A-202*	ACR	CTCF
5_57163724	ASE SNP	29,89763368	*MYL6-202*	NA	H3K27ac, H3K27me3, H3K4me1, H3K4me3
7_43816056	ASE SNP	28,44291619	*NDUFS7-202*	ACR	NA
7_81,259,719	ASE SNP	−69,63008916	*CKMT2-201*	ACR	NA

7_81269881	ASE SNP	85,73973385	*CKMT2-201*	ACR	NA

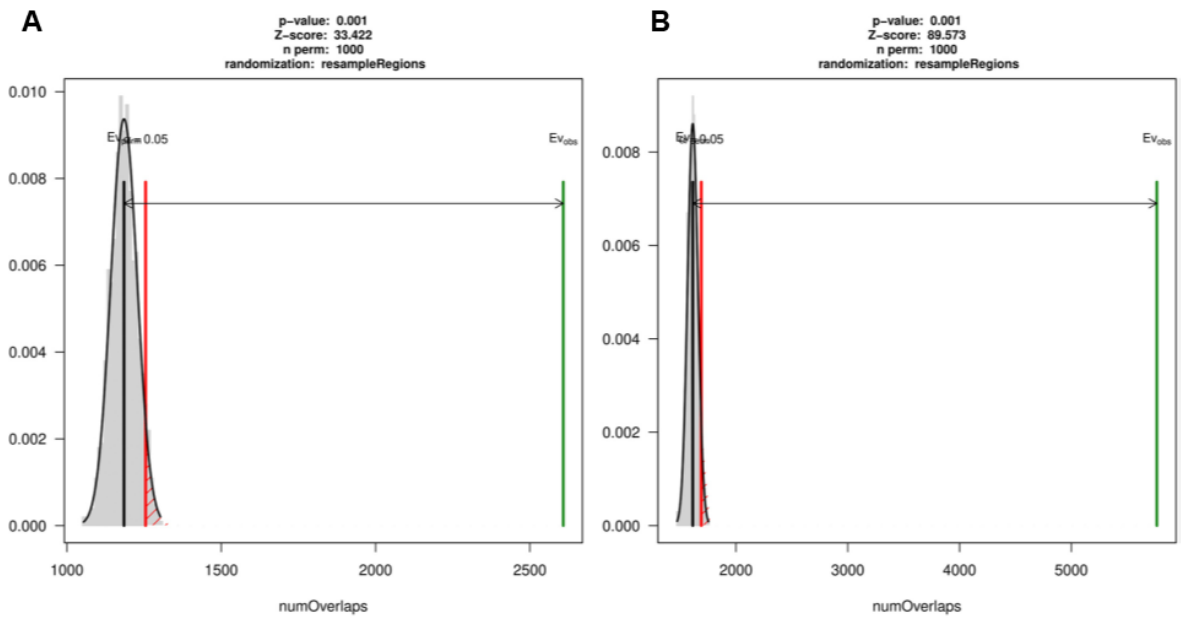


Figure 2.8 - Permutation tests plots. The black vertical line represents the average number of overlaps from randomized regions. The red vertical line represents the significance threshold and the green vertical line represents the number of overlaps observed with our data. A) Permutation test with CTCF regions. B) Permutation test with H3K27ac regions. Source: Author's.

2.5 Discussion

In this study we aimed to improve the ASE SNP discovery in bovine muscle samples using imputed genotype data. We identified a total of 38,177 ASE SNPs, 96.27% of these being novels. In addition, we were able to identify 21,543 aseQTL and 7448 ASE genes, many of them for the first time associated with ASE. Here, we improved the accuracy of the results without losing too much discovery power by using the WASP pipeline to filter out reads containing mapping bias [10,23] due to the inflated presence of the major allele [10,33]. This filtering, together with the requirement of 10 counts and the need for heterozygous SNPs for ASE analysis, resulted in 40,969 variants eligible to be tested. Despite this number being a small part of all imputed genotypes (0.85 %), it was considerably larger than the previous

study in the same population, in which De Souza et al. [14] tested 1085 SNPs from the Illumina BovineHD chip and identified 820 ASE SNPs. From these, 479 SNPs overlapped with the present study from imputation data. The remaining 341 ASE SNPs identified by De Souza et al. [14], corresponding to 41.6 % of their ASE SNPs, were not tested in any animal in the current study because, as we used a different method, they were filtered out by the WASP algorithm (see Methods section).

Almost all ASE SNPs identified here showed monoallelic expression in at least one sample. Previous studies in Nelore and Holstein reported 23 % and 4.3 % of monoallelic ASE SNPs, respectively. The lower rate observed in Holstein can be related to the sample size of the study, which focused on the ASE comparison between different tissues of a single cow [5]. The discovery of ASE genes overlapping imprinted genes in mammals, together with the monoallelic expression of these genes and the epigenetic data integration, evidenced the potential of our results helping investigate new imprinted genes in bovine.

The integration of our results with known ASE genes in bovine muscle [7,14] indicated conservation of the ASE pattern in cattle. Using a more extensive dataset, we presented 5496 novel ASE regions which were not identified in these previous studies, increasing the discovery depth of ASE regions in bovine muscle. We showed consistency with literature, encompassing large proportions of the previously identified ASE genes from the same population with less genotype density [14] and of the reported for *Limousin* cattle [7].

ASE genes presented interesting evidence of being related to metabolic pathways essential for muscle homeostasis such as mitogen-activated protein kinases (MAPKs). MAPKs regulate skeletal muscle development [51] and transcription in response to stress in muscle [52], which in turn can trigger autophagy [53], a lysosome-mediated degradative process involved with meat maturation [54]. Focal adhesion is a pathway involved in multiple muscle processes, improving muscle cells growth by inactivating anti-apoptotic pathways [55]. TNF pathway is involved with apoptosis and meat tenderness [56]. In addition, proteolysis mediated by the proteasome is a known mechanism involved in muscle wasting and meat tenderness [56–58].

Some ASE genes showed a great biological relevance in regulation and meat-related traits. The *MYOM2* (Myomesin 2) showed a higher number of ASE SNPs (*N* = 166). This gene was reported as an ASE gene by Guillocheau et al. and as putatively differentially affected by H3K27me3 [47]. The *MYOM2* gene encodes an isoform of MYOM, which is a protein that links the thick filaments in the middle of the sarcomere [59]. *MYOM2* was identified in Nelore muscle showing differential alternative splicing between extreme groups of ribeye area [60], also being its protein more abundant in the bovine muscle of lower adiposity group [61], evidencing the importance of this gene on the muscularity of beef cattle. The *MEST* (mesoderm specific transcript) and *GNAS* (GNAS complex locus) ASE genes figure among the imprinted genes in bovine [38]. Both genes were differentially methylated concerning the tenderness phenotype and *GNAS* was potentially differentially regulated by H3K27me3 between contrasting groups of Fe content [47]. The *MEST* ASE gene has been related to adipocyte differentiation and hypertrophy [62,63]. Moreover, the maternal imprinting of this gene is also known [64]. Here, *MEST* showed monoallelic expression in all significant samples for its unique ASE SNP. *GNAS* ASE gene also presented all the samples with monoallelic expression. This gene was previously identified as differentially expressed between contrasting samples for CLA-c9t11 [44], and associated with growth and carcass traits in pigs [65]. Parent-specific methylation and H3K27me3 modification are known mechanisms regulating genomic imprinting [66,67]. *IGF2R* (insulin-like growth factor 2 receptor) regulates the levels of IGF2, a peptide that functions as an embryonic growth factor. This ASE gene modulates muscle growth [68] and was previously identified as downregulated in the low group for CLA-c9t11 [44] and the low group for REA in our Nelore population [42], highlighting its relationship with muscle development. In addition, it was regulated by H3K27me3 histone modification [47] and differentially methylated for tenderness phenotype in our population [50]. This gene is within various enriched metabolic pathways, such as growth hormone synthesis, secretion and action, and thermogenesis.

Many genes affected by *cis*-eQTLs showed relevance for muscle homeostasis. For example, the *TNNT3* (Troponin 3) ASE gene encodes an isoform

of the troponin proteins family expressed in fast skeletal muscle, being related to muscle contraction [69] and negatively associated with tenderness in pig's [70]. We found two transcripts of *TNNT3* being regulated by the same *cis*-eQTLs in an opposite manner. The exon absent in *TNNT3-205* has SNPs annotated as splice acceptors, splice donors, splice regions, and missense variations. The last SNP of the LD block is located at 49.594 Mb, considerably close to the exon containing the start lost and stop gained regions, within the same LD block. Thus, the *cis*-eQTLs of this LD block are probably in LD with SNPs with high impact be observed for *TNNT3-205* perhaps because of the absence of some exons containing the start lost/stop gain regions that are present in the *TNNT3-206* transcript. Thus, our results suggest this region is probably under the effect of multiple regulatory mechanisms splice-variant dependent. We also found various *cis*-eQTLs regulating transcripts of the *CKMT2* gene. This gene encodes a sarcomeric mitochondrial creatine kinase responsible for energy metabolism and is involved in muscle contraction [71]. A previous study found an upregulation of this gene in pigs with better meat quality characteristics [71]. The *HSPA1A-201* transcript of the *HSPA1A* (Heat Shock Protein Family A (Hsp70) Member 1A) was associated with the *cis*-eQTL with the larger effect size. *HSPA1A* encodes to the Hsp70 protein, a biomarker of heat stress in animals [72]. This protein is upregulated under heat stress conditions [73], protecting cells from oxidative damage and apoptosis [72]. Chronic and acute heat stress can affect meat quality traits [74]. Previous studies in Nelore muscle showed that Hsp70 was downregulated in tender muscle samples [75] and *HSPA1A* SNPs were associated with ribeye area, water holding capacity, and cooking loss [19]. Intense regulatory and transcription activity mediated by motifs for TF binding is observed within accessible chromatin regions, in which chromatin structure is controlled by histones modifications resulting from acetylation and methylation events [11].

Exploring the regulatory potential of the ASE-related variants, we found considerable overlap with functional regions obtained experimentally from bovine muscle [46], mostly for ASE SNPs. This evidence is conflicting with the concept of ASE regions being only markers of allelic imbalance useful for discovering *cis*-regulatory variants (*cis*-eQTLs and aseQTLs). Still, it is reinforced by the

identification of 2098 ASE SNPs being identified as *cis*-eQTLs of the corresponding transcripts and the larger effect sizes observed for *cis*-eQTLs identified as ASE SNPs.

The integration with cGTEx data was consistent with these results. Our *cis*-eQTLs overlapped with >33% of *cis*-eQTLs obtained from the cGTEx data. In addition, 30% of our ASE SNPs and 25% of our aseQTLs overlapped these published *cis*-eQTLs. Disregarding divergences in methodology and statistical power inherent in the analysis in each of these studies, SNPs resulting from our study that did not overlap *cis*-eQTLs [76] or ASE SNPs [7] previously obtained in *Bos taurus* muscle may be specific to the Nelore breed, once there are genomic regions exclusive to indicine or taurine sub-species, as reported by Koufariotis et al. [77].

Additionally, the ASE can reflect allele-specific CTCF binding [78] or histone modifications [79]. In mammals, conserved histone modifications are less stable and less frequent in intergenic regions [80]. Only 0.6 % of the identified ASE SNPs were located in intergenic regions, contrasting with the proportion of 25 % of the aseQTLs. ASE SNPs in conserved histone modification regions could be a possible explanation for the ASE SNPs presenting more epigenetic regulation than the aseQTLs. More studies in larger sample sizes must be performed to explore the regulatory effect of ASE SNPs. Still, the potential regulatory effect of *cis*-eQTLs and aseQTLs is relevant.

A commonly accepted mechanism that may explain the allelic imbalance is the allele-specific binding of TFs within *cis*-regulatory regions [4,78]. We found 1075 aseQTLs identified as *cis*-eQTLs, which increase the confidence of these SNPs having a *cis*-regulatory potential [12]. aseQTLs seemingly have small effects individually but may act cumulatively throughout the transcript [10] or be found in large LD blocks, as observed by multiple aseQTLs affecting the same ASE gene, as for *NMNAT1*, the gene with more associated aseQTLs and *cis*-eQTLs. We found 2.043 ASE SNPs overlapping CNV regions [17], corresponding to 5.35 % of all ASE SNPs. Assuming the role of CNVs, any potential evidence of ASE SNPs in these regions warrants further investigation. For these cases, the allelic imbalance marked

by ASE SNPs may not be a result of a *cis*-regulatory mechanism alone but may also result from structural variation in their chromosomic region, but for the remaining 94.65 % the ASE pattern is likely due to transcriptional regulation.

The considerable overlap of ASE-related SNPs and epigenetic data from FAANG [46] drafts potential mechanisms acting in transcript regulation, as the *cis*-regulation of enhancers activation mediated by CTCFs [81], recruited through permissive chromatin states. However, within-population *in vitro* experiments would be necessary for more reliable interpretations.

2.6 Conclusion

This study identified a large ASE dataset in cattle, consistent with the previously published data in bovine muscle, and presented a considerable number of novel ASE regions and potentially regulatory variants (aseQTLs and *cis*-eQTLs). We show evidence that the identified ASE genes may be involved with metabolic pathways related to relevant livestock traits. The putative regulatory variants presented here may be affected by epigenetic regulation and influence muscle-related bovine phenotypes.

References

[1] L.-Y. Chang, S. Toghiani, A. Ling, S.E. Aggrey, R. Rekaya, High density marker panels, SNPs prioritizing and accuracy of genomic selection, BMC Genet. 19 (2018), https://doi.org/10.1186/s12863-017-0595-2, 44.

[2] D.J. Balding, A tutorial on statistical methods for population association studies, Nat. Rev. Genet. 7 (2006) 781–791, https://doi.org/10.1038/nrg1916.

[3] D.J. Schaid, W. Chen, N.B. Larson, From genome-wide associations to candidate causal variants by statistical fine-mapping, Nat. Rev. Genet. 19 (2018) 491–504, https://doi.org/10.1038/s41576-018-0016-z.

[4] S.E. Castel, A. Levy-Moonshine, P. Mohammadi, E. Banks, T. Lappalainen, Tools and best practices for data processing in allelic expression analysis, Genome Biol. 16 (2015), https://doi.org/10.1186/s13059-015-0762-6.

[5] A.J. Chamberlain, C.J. Vander Jagt, B.J. Hayes, M. Khansefid, L.C. Marett, C. A. Millen, et al., Extensive variation between tissues in allele specific expression in an outbred mammal, BMC Genomics 16 (2015) 1–20, https://doi.org/10.1186/ s12864-015-2174-0.

[6] H.A. Lawson, J.M. Cheverud, J.B. Wolf, Genomic imprinting and parent-of-origin effects on complex traits, Nat. Rev. Genet. 14 (2013) 609–617, https://doi.org/ 10.1038/nrg3543.

[7] G.M. Guillocheau, A. El Hou, C. Meersseman, D. Esquerr´e, E. Rebours, R. Letaief, et al., Survey of allele specific expression in bovine muscle, Sci. Rep. 9 (2019), https://doi.org/10.1038/s41598-019-40781-6.

[8] J. Zou, F. Hormozdiari, B. Jew, S.E. Castel, T. Lappalainen, J. Ernst, et al., Leveraging allelic imbalance to refine fine-mapping for eQTL studies, PLoS Genet. (2019) 15, https://doi.org/10.1371/journal.pgen.1008481.

[9] M. Khansefid, J.E. Pryce, S. Bolormaa, Y. Chen, C.A. Millen, A.J. Chamberlain, et al., Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle, BMC Genomics 19 (2018) 1–18, https://doi.org/10.1186/s12864-018-5181-0.

[10] S.E. Castel, F. Aguet, P. Mohammadi, F. Aguet, S. Anand, K.G. Ardlie, et al., A vast resource of allelic expression data spanning human tissues, Genome Biol. 21 (2020) 1–12, https://doi.org/10.1186/s13059-020-02122-z.

[11] E. Calo, J. Wysocka, Modification of enhancer chromatin: what, how, and Why? Mol. Cell 49 (2013) 825–837, https://doi.org/10.1016/j.molcel.2013.01.038.

[12] A. Battle, S. Mostafavi, X. Zhu, J.B. Potash, M.M. Weissman, C. McCormick, et al., Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals, Genome Res. 24 (2014) 14–24, https://doi.org/ 10.1101/gr.155192.113.

[13] Y. Gilad, S.A. Rifkin, J.K. Pritchard, Revealing the architecture of gene regulation: the promise of eQTL studies, Trends Genet. 24 (2008) 408–415, https://doi.org/ 10.1016/j.tig.2008.06.001.

[14] M.M. de Souza, A. Zerlotini, Rocha MIP, J.J. Bruscadin, T.F. Cardoso, , et al.S. Diniz WJ da, S. Diniz WJ da, Allele-specific expression is widespread in bos indicus muscle and affects meat quality candidate genes, Scientific Reports 10 (2020), https://doi.org/10.1038/s41598-020-67089-0, 10204–10204.

[15] S. Varkoohi, M.H. Banabazi, M. Ghsemi-Siab, Allele specific expression (Ase) analysis between bos taurus and bos indicus cows using rna-seq data at snp level and gene level, Anais Da Academia Brasileira de Ciencias 93 (2021) 1–9, https:// doi.org/10.1590/0001-3765202120191453.

[16] J.J. Bruscadin, M.M. de Souza, K.S. de Oliveira, M.I.P. Rocha, J. Afonso, T. F. Cardoso, et al., Muscle allele-specific expression QTLs may affect meat quality traits in bos indicus, Sci. Rep. 11 (2021), https://doi.org/10.1038/s41598-021- 86782-2, 7321-7321.

[17] L. Geistlinger, F. P´ertille, P.F. Giachetto, R.A. Brassaloti, et al., Genome-wide detection of CNVs and their association with meat tenderness in nelore cattle, PLoS ONE 11 (2016), https://doi.org/10.1371/journal.pone.0157711 e0157711.

[18] P.F. Przytycki, M. Singh, Differential allele-specific expression uncovers breast cancer genes dysregulated by cis noncoding mutations, Cell Systems 10 (2020) 193–203.e4, https://doi.org/10.1016/j.cels.2020.01.002.

[19] P.C. Tizioto, J.E. Decker, J.F. Taylor, R.D. Schnabel, M.A. Mudadu, F.L. Silva, et al., Genome scan for meat quality traits in nelore beef cattle, Physiol. Genomics 45 (2013) 1012–1020, https://doi.org/10.1152/physiolgenomics.00066.2013.

[20] A.S.M. Cesar, L.C.A. Regitano, J.M. Reecy, M.D. Poleti, P.S.N. Oliveira, G.B. de Oliveira, et al., Identification of putative regulatory regions and transcription factors associated with intramuscular fat content traits, BMC Genomics 19 (2018), https://doi.org/10.1186/s12864-018-4871-y, 499-499.

[21] B.D. Rosen, D.M. Bickhart, R.D. Schnabel, S. Koren, C.G. Elsik, E. Tseng, et al., De novo assembly of the cattle reference genome with single-molecule sequencing, GigaScience (2020) 9, https://doi.org/10.1093/gigascience/giaa021.

[22] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, et al., STAR: ultrafast universal RNA-seq aligner, Bioinformatics 29 (2013) 15–21, https://doi. org/10.1093/bioinformatics/bts635.

[23] B. Van De Geijn, G. Mcvicker, Y. Gilad, J.K. Pritchard, WASP: allele-specific software for robust molecular quantitative trait locus discovery, Nat. Methods 12 (2015) 1061–1063, https://doi.org/10.1038/nmeth.3582.

[24] P.C. Tizioto, S.L. Meirelles, G.B. Veneroni, R.R. Tullio, A.N. Rosa, M.M. Alencar, et al., A SNP in ASAP1 gene is associated with meat quality and production traits in

nelore breed, Meat Sci. 92 (2012) 855–857, https://doi.org/10.1016/j. meatsci.2012.05.018.

[25] Loh P-R, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics 2016;48:1443–8. doi:10.1038/ng.3679.

[26] S. Das, L. Forer, S. Sch¨onherr, C. Sidore, A.E. Locke, A. Kwong, et al., Next-generation genotype imputation service and methods, Nat. Genet. 48 (2016) 1284–1287, https://doi.org/10.1038/ng.3656.

[27] T.F. Cardoso, L.L. Coutinho, J.J. Bruscadin, W.J. da Silva Diniz, J. Petrini, B.G. N. Andrade, et al., Multi-omics approach reveals miR-SNPs affecting muscle fatty acids profile in nelore cattle, Genes 12 (2021), https://doi.org/10.3390/ genes12010067, 67-67.

[28] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A.R. Ferreira, D. Bender, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (2007) 559–575, https://doi.org/10.1086/ 519795.

[29] B.L. Browning, S.R. Browning, A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals, Am. J. Hum. Genet. 84 (2008) 210–223, https://doi.org/10.1016/j. ajhg.2009.01.005.

[30] H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, Bioinformatics 27 (2011) 2987–2993, https://doi.org/10.1093/bioinformatics/ btr509.

[31] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, et al., The variant call format and VCFtools, Bioinformatics 27 (2011) 2156–2158, https:// doi.org/10.1093/bioinformatics/btr330.

[32] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, et al., The structure of haplotype blocks in the human genome, Science 296 (2002) 2225–2229, https://doi.org/10.1126/science.1069424.

[33] J.F. Degner, J.C. Marioni, A.A. Pai, J.K. Pickrell, E. Nkadori, Y. Gilad, et al., Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data, Bioinformatics 25 (2009) 3207–3212, https://doi.org/10.1093/ bioinformatics/btp579.

[34] R. Patro, G. Duggal, M.I. Love, R.A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression, Nat. Methods 14 (2017) 417–419, https://doi.org/10.1038/nmeth.4197.

[35] A.A. Shabalin, Matrix eQTL: ultra fast eQTL analysis via large matrix operations, Bioinformatics 28 (2012) 1353–1358, https://doi.org/10.1093/bioinformatics/ bts163.

[36] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, et al., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6 (2012) 80–92, https://doi.org/10.4161/fly.19695.

[37] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, et al., ClueGO: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks, Bioinformatics 25 (2009) 1091–1093, https://doi.org/10.1093/bioinformatics/btp101.

[38] GeneImprint 2012. www.geneimprint.com (accessed July 20, 2021).

[39] UCSC Lift Genome Annotations n.d. http://genome.ucsc.edu/cgi-bin/hgLiftOver (accessed July 20, 2021).

[40] J. Afonso, L.L. Coutinho, P.C. Tizioto, W.J. da Silva Diniz, A.O. de Lima, M.I. P. Rocha, et al., Muscle transcriptome analysis reveals genes and metabolic pathways related to mineral concentration in bos indicus, Sci. Rep. 9 (2019), https://doi.org/10.1038/s41598-019-49089-x.

[41] L.L. Coutinho, P.C. Tizioto, Cesar ASM, C.F. Gromboni, Nogueira ARA, et al., Iron Content Affects Lipogenic Gene Expression in the Muscle of Nelore Beef Cattle, PLOS ONE 11 (2016), https://doi.org/10.1371/journal.pone.0161160 e0161160–e0161160.

[42] B. Silva-Vignato, L.L. Coutinho, A.S.M. Cesar, M.D. Poleti, L.C.A. Regitano, J.C. C. Balieiro, Comparative muscle transcriptome associated with carcass traits of Nellore cattle, BMC Genomics 18 (2017), https://doi.org/10.1186/s12864-017- 3897-x.

[43] T.M. Gonçalves, L.C. de Almeida Regitano, J.E. Koltes, A.S.M. Cesar, S.C. da Silva Andrade, G.B. Mour˜ao, et al., Gene co-expression analysis indicates potential pathways and regulators of beef tenderness in Nellore cattle, Front. Genet. (2018) 9, https://doi.org/10.3389/fgene.2018.00441.

[44] A.S.M. Cesar, L.C.A. Regitano, M.D. Poleti, S.C.S. Andrade, P.C. Tizioto, P.S. N. Oliveira, et al., Differences in the skeletal muscle transcriptome profile associated with extreme values of fatty acids content, BMC Genomics 17 (2016), https://doi.org/10.1186/s12864-016-3306-x.

[45] A.S.M. Cesar, L.C.A. Regitano, J.E. Koltes, E.R. Fritz-Waters, D.P.D. Lanna, G. Gasparin, et al., Putative regulatory factors associated with intramuscular fat content, PLOS ONE 10 (2015), https://doi.org/10.1371/journal.pone.0128350.

[46] C. Kern, Y. Wang, X. Xu, Z. Pan, M. Halstead, G. Chanthavixay, et al., Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research, Nat. Commun. 12 (2021), https://doi.org/10.1038/s41467-021-22100-8.

[47] Afonso J, Shim WJ, Boden M, Palpant N, Fortes MRS. H3K27me3 predicted regulation as an indicator of Nelore muscle gene expression modulation related to mineral quantification. Personal Communication n.d.

[48] M.M. de Souza, S.C.M. Niciura, M.I.P. Rocha, Z. Pan, H. Zhou, J.J. Bruscadin, et al., DNA methylation may affect beef tenderness through signal transduction in

bos indicus, Epigenetics Chromatin 15 (2022) 15, https://doi.org/10.1186/s13072-022-00449-4.

[49] B. Gel, A. Díez-Villanueva, E. Serra, M. Buschbeck, M.A. Peinado, R. Malinverni, regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests, Bioinformatics (2015) btv562, https://doi.org/ 10.1093/bioinformatics/btv562.

[50] M. de Souza, S.C.M. Niciura, M.I.P. Rocha, Z. Pan, H. Zhou, J.J. Bruscadin, et al., DNA methylation may affect beef tenderness through signal transduction in bos indicus, In Review (2022), https://doi.org/10.21203/rs.3.rs-1415533/v1.

[51] A. Keren, Y. Tamir, E. Bengal, The p38 MAPK signaling pathway: a major regulator of skeletal muscle development, Mol. Cell. Endocrinol. 252 (2006) 224–230, https://doi.org/10.1016/j.mce.2006.03.017.

[52] H.F. Kramer, L.J. Goodyear, Exercise, MAPK, and NF-κB signaling in skeletal muscle, J. Appl. Physiol. 103 (2007) 388–395, https://doi.org/10.1152/japplphysiol.00085.2007.

[53] G. Kroemer, G. Mari˜no, B. Levine, Autophagy and the integrated stress response, Mol. Cell 40 (2010) 280–293, https://doi.org/10.1016/j.molcel.2010.09.023.

[54] M. García-Macia, V. Sierra, A. Palanca, I. Vega-Naredo, D. de Gonzalo-Calvo, S. Rodríguez-Gonz´alez, et al., Autophagy during beef aging, Autophagy 10 (2014) 137–143, https://doi.org/10.4161/auto.26659.

[55] Z.A. Graham, P.M. Gallagher, C.P. Cardozo, Focal adhesion kinase and its role in skeletal muscle, J. Muscle Res. Cell Motil. 36 (2015) 305–315, https://doi.org/10.1007/s10974-015-9415-3.

[56] C.M. Kemp, P.L. Sensky, R.G. Bardsley, P.J. Buttery, T. Parr, Tenderness – an enzymatic view, Meat Sci. 84 (2010) 248–256, https://doi.org/10.1016/j.meatsci.2009.06.008.

[57] M. Altun, H.C. Besche, H.S. Overkleeft, R. Piccirillo, M.J. Edelmann, B.M. Kessler, et al., Muscle wasting in aged, sarcopenic rats is associated with enhanced activity of the ubiquitin proteasome pathway, J. Biol. Chem. 285 (2010) 39597–39608, https://doi.org/10.1074/jbc.M110.129718.

[58] A.R. Thomas, H. Gondoza, L.C. Hoffman, V. Oosthuizen, R.J. Naud´e, The roles of the proteasome, and cathepsins B, L, H and D, in ostrich meat tenderisation, Meat Sci. 67 (2004) 113–120, https://doi.org/10.1016/j.meatsci.2003.10.001.

[59] L. Tskhovrebova, J. Trinick, Making muscle elastic: the structural basis of myomesin stretching, PLoS Biol. (2012) 10, https://doi.org/10.1371/journal.pbio.1001264.

[60] D.B.S. Silva, L.F.S. Fonseca, D.G. Pinheiro, A.F.B. Magalh˜aes, M.M.M. Muniz, J. A. Ferro, et al., Spliced genes in muscle from nelore cattle and their association

with carcass and meat quality, Sci. Rep. 10 (2020), https://doi.org/10.1038/ s41598-020-71783-4.

[61] J. Bazile, B. Picard, C. Chambon, A. Valais, M. Bonnet, Pathways and biomarkers of marbling and carcass fat deposition in bovine revealed by a combination of gel-based and gel-free proteomic analyses, Meat Sci. 156 (2019) 146–155, https://doi. org/10.1016/j.meatsci.2019.05.018.

[62] Y. Kadota, M. Yanagawa, T. Nakaya, T. Kawakami, M. Sato, S. Suzuki, Gene expression of mesoderm-specific transcript is upregulated as preadipocytes differentiate to adipocytes in vitro, J. Physiol. Sci. 62 (2012) 403–411, https://doi. org/10.1007/s12576-012-0217-8.

[63] I. Prudovsky, R.P. Anunciado-Koza, C.G. Jacobs, D. Kacer, M.E. Siviski, R.A. Koza, Mesoderm-specific transcript localization in the ER and ER-lipid droplet interface supports a role in adipocyte hypertrophy, J. Cell. Biochem. 119 (2018) 2636–2645, https://doi.org/10.1002/jcb.26429.

[64] L. Lefebvre, S. Viville, S.C. Barton, F. Ishino, E.B. Keverne, M.A. Surani, Abnormal maternal behaviour and growth retardation associated with loss of the imprinted gene mest, Nat. Genet. 20 (1998) 163–169, https://doi.org/10.1038/2464.

[65] M. Oczkowicz, K. Ropka-Molik, M. Tyra, Analysis of the associations between polymorphisms in GNAS complex locus and growth, carcass and meat quality traits in pigs, Mol. Biol. Rep. 40 (2013) 6419–6427, https://doi.org/10.1007/s11033- 013-2756-3.

[66] M.S. Bartolomei, R.J. Oakey, A. Wutz, Genomic imprinting: an epigenetic regulatory system, PLoS Genet. 16 (2020), https://doi.org/10.1371/journal. pgen.1008970.

[67] M.W.D. Raas, D.W. Zijlmans, M. Vermeulen, H. Marks, There is another: H3K27me3-mediated genomic imprinting, Trends Genet. (2021), https://doi.org/ 10.1016/j.tig.2021.06.017.

[68] Y. Torrente, P. Bella, L. Tripodi, C. Villa, A. Farini, Role of insulin-like growth factor receptor 2 across muscle homeostasis: implications for treating muscular dystrophy, Cells 9 (2020), https://doi.org/10.3390/cells9020441, 441-441.

[69] B. Wei, J.-P. Jin, TNNT1, TNNT2, and TNNT3: isoform genes, regulation, and structure–function relationships, Gene 582 (2016) 1–13, https://doi.org/10.1016/ j.gene.2016.01.006.

[70] V. Lobjois, L. Liaubet, M. SanCristobal, J. Gl´enisson, K. F`eve, J. Ralli`eres, et al., A muscle transcriptome analysis identifies positional candidate genes for a complex trait in pig, Anim. Genet. 39 (2008) 147–162, https://doi.org/10.1111/j.1365-2052.2007.01695.x.

[71] M. Damon, J. Wyszynska-Koko, A. Vincent, F. H´erault, B. Lebret, Comparison of muscle transcriptome between pigs with divergent meat quality phenotypes

identifies genes related to muscle metabolism and structure, PLoS ONE 7 (2012), https://doi.org/10.1371/journal.pone.0033763 e33763-e33763.

[72] A. Pr, Role of heat shock proteins in livestock adaptation to heat stress, JDVAR (2017) 5, https://doi.org/10.15406/jdvar.2017.05.00127.

[73] I. Hyder, M. Pasumarti, P.R. Reddy, C.S. Prasad, K.A. Kumar, V. Sejian, Thermotolerance in Domestic Ruminants: A HSP70 Perspective, in: Asea AAA, P. Kaur (Eds.), Heat Shock Proteins in Veterinary Medicine and Sciences 12, Springer International Publishing, Cham, 2017, pp. 3–35, https://doi.org/ 10.1007/978-3-319-73377-7_1.

[74] P.A. Gonzalez-Rivas, S.S. Chauhan, M. Ha, N. Fegan, F.R. Dunshea, R.D. Warner, Effects of heat stress on animal physiology, metabolism, and meat quality: a review, Meat Sci. 162 (2020), 108025, https://doi.org/10.1016/j.meatsci.2019.108025.

[75] J.M. Malheiros, C.E. Enríquez-Valencia, C.P. Braga, J.C.S. Vieira, D.S. Vieira, G. L. Pereira, et al., Application of proteomic to investigate the different degrees of meat tenderness in Nellore breed, J. Proteome 248 (2021), 104331, https://doi.org/10.1016/j.jprot.2021.104331.

[76] S. Liu, Y. Gao, O. Canela-Xandri, S. Wang, Y. Yu, W. Cai, et al., A comprehensive catalogue of regulatory variants in the cattle transcriptome, Genomics (2020), https://doi.org/10.1101/2020.12.01.406280.

[77] L. Koufariotis, B.J. Hayes, M. Kelly, B.M. Burns, R. Lyons, P. Stothard, et al., Sequencing the mosaic genome of Brahman cattle identifies historic and recent introgression including polled, Sci. Rep. 8 (2018) 17761, https://doi.org/10.1038/s41598-018-35698-5.

[78] A.O. Degtyareva, E.V. Antontseva, T.I. Merkulova, Regulatory SNPs: altered transcription factor binding sites implicated in complex traits and diseases, IJMS 22 (2021) 6454, https://doi.org/10.3390/ijms22126454.

[79] U. Gaur, K. Li, S. Mei, G. Liu, Research progress in allele-specific expression and its regulatory mechanisms, J. Appl. Genetics 54 (2013) 271–283, https://doi.org/10.1007/s13353-013-0148-y.

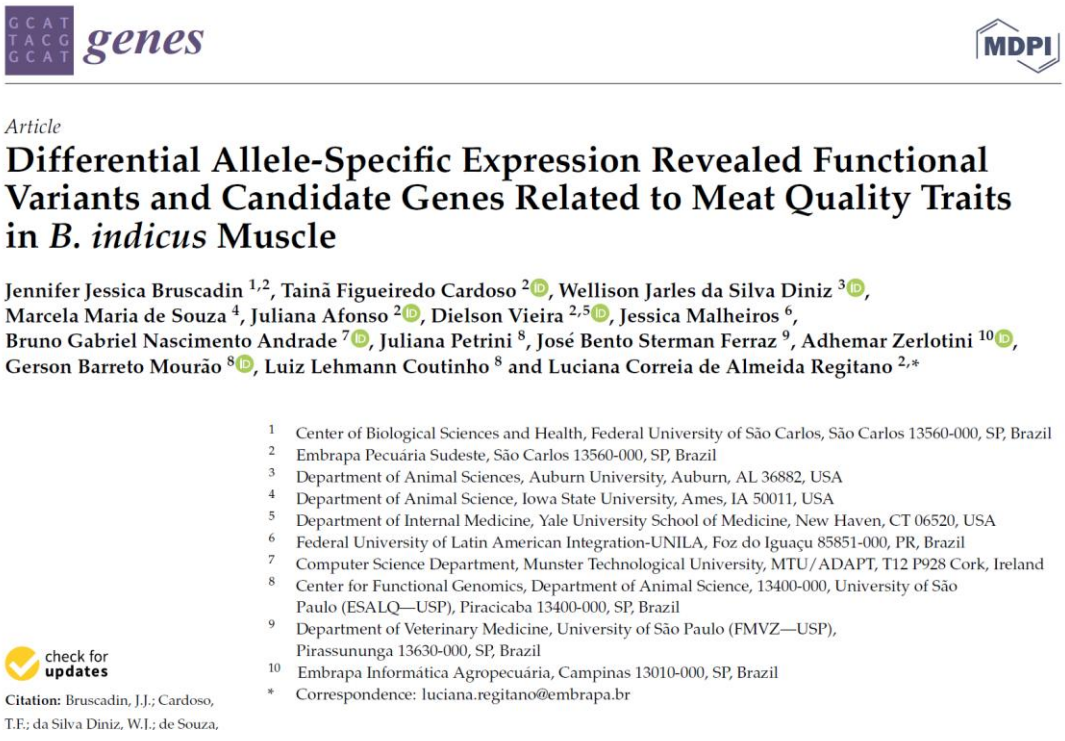
[80] Y.H. Woo, W.-H. Li, Evolutionary conservation of histone modifications in mammals, Mol. Biol. Evol. 29 (2012) 1757–1767, https://doi.org/10.1093/molbev/mss022.

[81] G. Ren, W. Jin, K. Cui, J. Rodrigez, G. Hu, Z. Zhang, et al., CTCF-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression, Mol. Cell 67 (2017) 1049–1058.e6, https://doi.org/10.1016/j.molcel.2017.08.026.

3 PHENOTYPE-DEPENDENT ALLELE-SPECIFIC EXPRESSION

This section contains transcriptions from the paper “Differential Allele-Specific Expression Revealed Functional Variants and Candidate Genes Related to Meat Quality Traits in *B. indicus* Muscle”, published in the Journal Genes on December 11th, 2022. The following results are based on differentially allele-specific expressed SNPs (DASE SNPs), for which the allelic imbalance diverged between samples from contrasting groups of meat-related traits, showing a potential relevance of ASE SNPs on phenotypic variation. The manuscript header is shown on Figure 3.1.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/genes13122336/s1



genes

MDPI

Article

Differential Allele-Specific Expression Revealed Functional Variants and Candidate Genes Related to Meat Quality Traits in *B. indicus* Muscle

Jennifer Jessica Bruscadin [1,2], Tainã Figueiredo Cardoso [2], Wellison Jarles da Silva Diniz [3], Marcela Maria de Souza [4], Juliana Afonso [2], Dielson Vieira [2,5], Jessica Malheiros [6], Bruno Gabriel Nascimento Andrade [7], Juliana Petrini [8], José Bento Sterman Ferraz [9], Adhemar Zerlotini [10], Gerson Barreto Mourão [8], Luiz Lehmann Coutinho [8] and Luciana Correia de Almeida Regitano [2,*]

1 Center of Biological Sciences and Health, Federal University of São Carlos, São Carlos 13560-000, SP, Brazil
2 Embrapa Pecuária Sudeste, São Carlos 13560-000, SP, Brazil
3 Department of Animal Sciences, Auburn University, Auburn, AL 36882, USA
4 Department of Animal Science, Iowa State University, Ames, IA 50011, USA
5 Department of Internal Medicine, Yale University School of Medicine, New Haven, CT 06520, USA
6 Federal University of Latin American Integration-UNILA, Foz do Iguaçu 85851-000, PR, Brazil
7 Computer Science Department, Munster Technological University, MTU/ADAPT, T12 P928 Cork, Ireland
8 Center for Functional Genomics, Department of Animal Science, 13400-000, University of São Paulo (ESALQ—USP), Piracicaba 13400-000, SP, Brazil
9 Department of Veterinary Medicine, University of São Paulo (FMVZ—USP), Pirassununga 13630-000, SP, Brazil
10 Embrapa Informática Agropecuária, Campinas 13010-000, SP, Brazil
* Correspondence: luciana.regitano@embrapa.br

check for updates

Citation: Bruscadin, J.J.; Cardoso, T.F.; da Silva Diniz, W.I.; de Souza,

Figure 3.1 - Header of the published manuscript: “Differential Allele-Specific Expression Revealed Functional Variants and Candidate Genes Related to Meat Quality Traits in *B. indicus* Muscle”. Source: Author’s.

3.1 Abstract

Traditional transcriptomics approaches have been used to identify candidate genes affecting economically important livestock traits. Regulatory variants affecting these traits, however, remain under covered. Genomic regions showing allele-specific expression (ASE) are under the effect of *cis*-regulatory variants, being useful for improving the accuracy of genomic selection models. Taking advantage of the better of these two methods, we investigated single nucleotide polymorphisms (SNPs) in regions showing differential ASE (DASE SNPs) between contrasting groups for beef quality traits. For these analyses, we used RNA sequencing data, imputed genotypes and genomic estimated breeding values of muscle-related traits from 190 Nelore (*Bos indicus*) steers. We selected 40 contrasting unrelated samples for the analysis (N = 20 animals per contrasting group) and used a beta-binomial model to identify ASE SNPs in only one group (*i.e.*, DASE SNPs). We found 1479 DASE SNPs (FDR ≤ 0.05) associated with 55 beef-quality traits. Most DASE genes were involved with tenderness and muscle homeostasis, presenting a co-expression module enriched for the protein ubiquitination process. The results overlapped with epigenetics and phenotype-associated data, suggesting that DASE SNPs are potentially linked to *cis*-regulatory variants affecting simultaneously the transcription and phenotype through chromatin state modulation.

Keywords: phenotype; beef; cis-regulation; ASE

3.2 Introduction

Allele-specific expression (ASE) analysis is a helpful approach to searching for *cis*-regulatory variants [1,2]. ASE consists of the imbalanced ratio between reference and alternative alleles' counts in a heterozygous locus, being a common pattern in mammals [3–5]. As ASE is determined per individual, it reduces the effect of technical factors found in total expression studies [1,6]. Thus, from the knowledge of the transcribed or repressed alleles it is possible to infer *cis*-regulatory variants related to this expression pattern [1], allowing increased accuracy in predictive models and improving functional variants discovery [7].

As our research group [8,9] and others reported, ASE is widespread in bovine [4,6,8, 10,11]. Previously, we described a high coverage list of single nucleotide polymorphisms (SNPs) located in transcripts with ASE (ASE SNPs) distributed in Nelore (*Bos indicus*) muscle and the candidate *cis*-regulatory variants potentially affecting the allelic expression. Additionally, we have shown evidence that the ASE SNPs can regulate transcription, as they are located in functional regions [9]. ASE analysis benefits the search for functional variants affecting important economic traits, since genes containing ASE SNPs (ASE genes) may play a role in muscle maintenance and are related to meat quality traits [8,10].

A traditional transcriptomics analysis compares the total gene expression between phenotypically contrasting animals [12], resulting in differentially expressed genes (DEGs). Other techniques are needed to investigate variants associated with the DEGs expression (expression quantitative trait loci - eQTLs, for example) and associated with the phenotype (quantitative trait loci - QTLs). Thus, these data layers need to be integrated to draft more consistent potential mechanisms. Facing the benefits of conducting differential expression and ASE analyses, a differential allele-specific expression (DASE) approach can identify regions where the ASE level differs between extreme samples for a given trait.

The DASE regions are likely affected by *cis*-regulatory variants affecting the expression pattern because of the observed ASE. In addition, the same *cis*-regulatory variants may be related to the phenotype of interest due to allelic imbalance being trait-dependent. Thus, this approach can reduce analytic steps and

avoid false positives compared to integrating results from multiple consecutive analyses. There are only a few studies with global DASE analysis, probably due to the requirement of high density genotypic, transcriptomic and phenotypic data from the same sample. Genome-wide DASE regions were related to breast cancer phenotypes [7,13], but no study could be found in livestock so far.

In this study, we searched for DASE SNPs related to 55 meat quality, carcass quality, mineral content, and fatty acid composition traits in the muscle of a Nelore population. We conducted gene enrichment and co-expression analyses to investigate functional interactions between DASE genes. We integrated the DASE SNPs with results obtained previously in the same population, such as *cis*-eQTLs (*i.e.*, SNPs associated with the transcript abundance) and aseQTLs (*i.e.*, SNPs associated with the allelic imbalance). These variants may affect the traits by the cumulative regulatory effect in co-expressed genes related to these muscle-related traits.

3.3 Methods

3.3.1 Animal Production and Sample Collection

The experimental population used in this study was part of a large project with all experimental approaches approved by the Institutional Animal Care and Use Committee Guidelines of Embrapa Pecuária Sudeste ethics committee (São Carlos, São Paulo, Brazil. Protocol CEUA 01/2013). Details regarding the animal's production can be found in previous studies [14–16]. For the animal production, 34 sires were selected to have the lesser kinship and to represent the main lineages of Nelore cattle in Brazil. Semen samples of these animals were stored at -80°C for later analysis. Artificial inseminations were performed in purebred Nelore dams, avoiding the same parental pair [15], generating a 386 Nelore steers population [16]. The animals were raised under the same feeding and management as detailed described previously [15]. The animals were slaughtered once they reached 5 mm of backfat thickness (BFT), measured by ultrasound. Blood samples were collected in vacuum tubes containing potassium EDTA (K3) and muscle samples were immediately collected between the 11th and 13th ribs of the *Longissimus thoracis*

muscle [14–16]. Here, we used the data of 190 steers from this experimental population for which phenotype, genotype, and RNA-Seq data were available, being the same samples used in previous studies of our group [8,9], to ensure comparability.

3.3.2 Phenotype Measurements

Phenotypes' description and genomic estimated breeding values (GEBVs)' summary are described in Table S2. Tenderness, Backfat Thickness (BFT), and Ribeye Area (REA) traits were collected as described previously [15] Briefly, tenderness phenotypes were measured based on the shear force obtained from the Warner-Bratzler equipment at 24 h (WBSF0), seven days (WBSF7), and 14 days after slaughter (WBSF14) [15]. The phenotypes for mineral content (Ar, Ca, Cr, Co, Cu, Fe, Mg, Mn, P, K, Se, Na, S, and Zn) were obtained from 100 mg lyophilized muscle sample, being measured by a Vista Pro-CCD ICP-OES spectrometer with a radial view (Varian, Mulgrave, Australia) [17]. GEBVs for these phenotypes were calculated in GenSel software (https://github.com/austin-putz/GenSel, accessed on 10 April 2020), using birth, feedlot locations, and breeding seasons as fixed effects, and age at slaughter was included as a covariate in the statistical model [15,17].

The intramuscular fat (IMF) and fatty acid composition traits (described in Table S2) were previously reported elsewhere [16]. The measures were conducted in the Animal Nutrition and Growth Laboratory at ESALQ, Piracicaba, São Paulo, Brazil. IMF was obtained using an Ankom XT20 Fat Analyzer, following the recommended methods. Fatty acid content was extracted from 4 g of muscle samples and the profile identification was performed using gas chromatography by comparing the retention time of the obtained methyl esters with reference standards. Detailed methods are described previously [16], which also presented the model for the GEBV estimation for these traits.

The selection of animals to compose contrasting sample groups was similar to the criteria adopted in the previously reported differential expression data [18–22]. The animals were ordered and grouped according to the GEBV values for each trait, being the group with higher GEBV values classified as High and the one with minor values classified as Low. The animals were chosen to avoid steers descending from

the same sires being allocated to the same contrasting group. A Student's T-test was applied for all mentioned traits (p-value < 0.05) to determine whether the phenotypes differed between the two groups. To observe the relationship between phenotypes in our dataset, we computed the Pearson correlation between GEBVs values from the subpopulation of 190 animals in a correlogram.

3.3.3 DNA Extraction, Genotyping and Whole Genome Sequencing

DNA was extracted from frozen semen samples of the sires using the phenol-chloroform method. Extracted DNA from 26 sires was used for whole-genome sequencing, performed in an equipment Illumina HiSeq 2500 System (Illumina Inc., San Diego, CA, USA), with 8–21X coverage, in the Functional Genomic Center—ESALQ (Piracicaba, SP, Brazil). Data management was executed as suggested by the 1000 Bulls Genomes Project (http://www.1000bullgenomes.com/, accessed on 1 July 2020), as previously described [23].

Concerning the progeny, DNA was extracted from blood samples using the salting out method, as described previously [14]. For sires and steers samples, the DNA concentration was quantified in NanoDrop®, purity was analyzed by comparing the optic absorbance ratio 260/280, and the material integrity was observed by electrophoresis in agarose gel. Genotyping was conducted in Illumina BovineHD BeadChip (Illumina Inc, San Diego, CA, USA) with the extracted DNA from sires and steers, as detailed described elsewhere [24].

Variant imputation was executed with the high-density genotypes together with the 26 sires' whole-genome sequencing data, which methods were described elsewhere [23]. Concisely, the imputation accuracy was analyzed using a leave-one-out method, where the sequenced genotypes were used as a reference to observe consistent imputation results. Quality control was conducted with PLINK software [25], maintaining autosomal SNPs with minor allele frequency higher than 0.01, which showed an imputation error rate of less than 2% [23]. The final genotype file contained nearly 4.8 million SNPs.

3.3.4 RNA Extraction and Sequencing

RNA from 100 mg of 190 muscle samples from steers was extracted for transcriptomics analysis using TRIzol® (Life Technologies, Carlsbad, CA, USA), according to the manufacturer's instructions. RNA libraries were prepared with the TruSeq RNA Sample Preparation Kit (Illumina, San Diego, CA, USA) with the default protocol. Clustering and paired-end sequencing were executed in an Illumina HiSeq 2500® (Illumina, San Diego, CA, USA) equipment. The quality control removed sequencing adapters and low complexity reads with the software SeqyClean v1.4.13 [26]. The RNA sequencing was made in the Functional Genomic Center—ESALQ (Piracicaba, SP, Brazil), being detailed in a previous manuscript [27].

The sequences were mapped to the bovine reference genome (ARS-UCD1.2, Ensembl 100) using the STAR software [28] with default settings. The WASP pipeline [29] was applied to remove reads presenting mapping bias, which commonly inflates the reference allele counts [30]. Counts of reads overlapping gene regions were obtained using the Htseq-count tool [31], and allele-specific counts were obtained from heterozygous variants with GATK ASE Read Counter [32].

3.3.5 Analysis of Differential Allele-Specific Expression (DASE) between Contrasting Groups for Meat Quality Phenotypes

The DASE analysis was conducted to evaluate the relationship between the allelic expression imbalance level and the phenotype by comparing allelic ratios between contrasting groups. In this approach, tests were performed per phenotype for all SNPs within their respective genes aiming to compare results with those from differential expression analysis, co-expression, and gene functions. We used the DESeq2 R-package [33], including each sample twice in the matrix of counts (for reference and alternate alleles expression), being the design formula defined as:

design = ~phenotype + phenotype:sample + phenotype:count

where: ~phenotype corresponds to the variation between High and Low groups for the meat quality traits; phenotype:sample calculates the phenotypic variation of the animals within each group; and phenotype:count estimates the count's difference between reference and alternate alleles for the two contrasting

groups. Unlike the traditional differential expression analysis, size factor normalization is unnecessary in this approach because the sample is included in the design formula. The code and model details can be found on the tutorial page, described by the DESeq2 developer (http://rstudio-pubs-static.s3.amazonaws.com/275642:e9d578fe1f7a404aad0553f52236c0a4.html, accessed on 12 September 2022). In this analysis, each test results in three p-values. One determines the significance of the difference between allele counts within the High group, other for the Low group, and the last p-value corresponds to the ASE significance between the two contrasting groups, which is adjusted for the False Discovery Rate (FDR) implemented by the Benjamini-Hochberg method. The script also returns Log2FoldChange (FC) values, indicating which contrasting group showed allelic imbalance. The negative FC values indicate that the allelic imbalance was larger in the Low group compared to the High group. The positive ones correspond to variants in which the ASE was downregulated in the Low group compared to the High group. This analysis resulted in DASE SNPs, which corresponded to the SNPs marking the significant DASE within a gene, which were named DASE genes. We considered as significant the DASE SNPs corrected per phenotype and per gene with FDR ≤ 0.05.

3.3.6 Functional Annotation and Gene Enrichment Analyses

The DASE SNPs were annotated using SNPEff software [34] against *B. taurus* genome coordinates. Gene ontology (GO) terms related to biological processes and molecular functions involving DASE genes were obtained from BiomaRT (Ensemble release 104). The enriched KEGG metabolic pathways containing DASE genes were obtained with ClueGo Plugin for CytosCape [35], release 101, with the two-sided hypergeometric test in the enrichment analysis (FDR ≤ 0.05).

3.3.7 Co-Expression Analysis

To evaluate the correlation between DASE genes, we performed a co-expression analysis using the default parameters of the CEMITool package [36]. This software results in outputs with co-expressed modules and performs gene set

enrichment (GSEA) and over-representation (ORA) analyses. The GO biological processes terms underlying the DASE genes were used for ORA (FDR < 0.1). In addition, the phenotype with the largest number of DASE genes was chosen for GSEA. The animals for this trait were separated as High, Low, or “others” according to their GEBVs, “others” being the remaining animals excluding those belonging to contrasting groups. This analysis was conducted out to investigate the relevance of co-expressed genes in each module.

3.3.8 Data Integration

We conducted data integration analysis to examine if the SNPs found in this work were previously related to ASE and bovine phenotypes in other studies. First, we retrieved ASE data from the same Nelore population, regardless of the phenotypes [8,9]. After that, to compare the potential effect on the phenotypes, we downloaded all GWAS data obtained in the muscle of Nelore from previously published manuscripts related to meat quality [15], fatty acid content [16], and mineral concentration traits [37]. QTLs coordinates obtained from the previous bovine reference genome version (UMD3.1) were converted to the latest version (ARS-UCD1.2) using the liftOver tool from the UCSC database (https: //genome.ucsc.edu/cgi-bin/hgLiftOver, accessed on 7 June 2022).

DASE genes were compared with DEGs from the same Nelore population, identified between contrasting samples for meat quality [18,21], fatty acid composition [19], and mineral content [20,22] traits.

To identify SNPs potentially regulating the expression pattern, we considered a window of 200 kb upstream and downstream of each DASE SNP to search for aseQTLs [9], *cis*-eQTLs [9] and differentially methylated SNPs (DM) [38]. This 200 kb window was chosen based on our previous study [23] where the aseQTLs were mostly located until 200 kb of the regulated ASE SNP, and based on the high LD with the DASE SNPs expected to be present in that genomic distance. DASE SNPs and their 200 kb-distant potential regulatory SNPs were compared with ChIP-Seq and ATAC-seq data obtained in cattle muscle for the Functional Annotation of Animal Genomes (FAANG) [39]. The significance of all comparisons with genomic position

ranges was tested with 1000 permutations performed by the RegioneR package [40], with a p-value threshold of 0.05.

3.4 Results

3.4.1 The Muscle-Related Phenotypes Are Correlated

Herein, we evaluated the potential effect of ASE on 55 carcasses and meat quality, mineral content, and fatty acids composition traits obtained in Nelore steers' muscle. GEBVs and the summary by sample group used here are reported in Table S1. The average GEBVs differed significantly between the contrasting groups for all phenotypes (p-value ≤ 0.05). After selecting animals with extreme phenotypes for the different traits (N = 20 per contrasting groups), 187 from the initial 190 samples were kept for analysis. The animals could be analyzed in more than one phenotype. The average number of phenotypes tested per sample was 11.76, ranging from one to 36 traits.

We used all animals with collected phenotypes and RNA-Seq data (N = 190) to identify correlated trait clusters in our population. From the correlation matrix, we plot the correlogram in Figure 3.2. We observed two blocks of strong positive correlations, one between the Ca, S, Zn, Na, Mg, and K minerals and the other between the C14:1 cis-9, C16:1 cis-9, C20:2, C18:3n3, PUFA, C20:5, C20:4, C22:5, and n3 fatty acids, including weak positive correlations with Cr and Cu in the same block. Negative correlation blocks can be observed, being the more consistent involving C10:0, C12:0, C14:0, C15:0, C18:1 trans -10,11,12, C17:0, C18:3n6, C18:0, C16:0 and SFA traits against WBSF0, WBSF7, WBSF14, C18:1 cis-13, C18:1 cis-12, C18:1 cis-11, C17:1, C20:1, C18:1 cis-9, and C18:1 total traits. The correlation matrix is presented in Table S2.

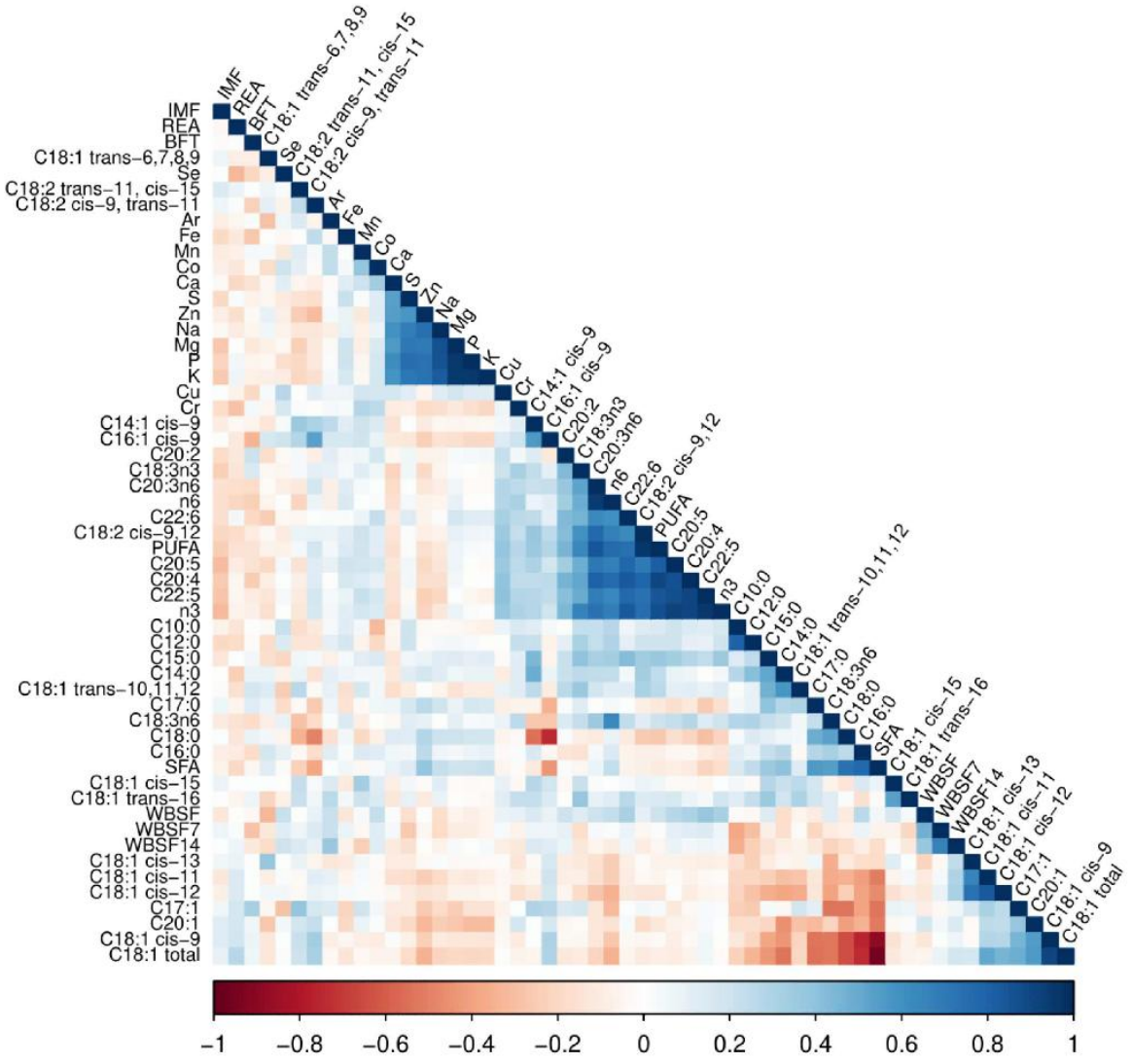


Figure 3.2 - Correlogram of mineral and fatty acid content phenotypes, carcass and meat quality traits. Red squares represent negative correlations, blue squares represent positive correlations and the white squares represent phenotypes without correlation in this population.

3.4.2 Differential Allelic Imbalance Related to Livestock Traits Was Identified in Bovine Muscle

We used several filters to increase the DASE results' reliability. A sample could be tested if it exhibited a heterozygous candidate SNP with at least ten counts. Therefore, we set a threshold of at least three animals for which the SNP satisfied our requirements in each contrasting group. Thus, for all SNPs, the sample number

that passed these filters was less than the original 20 samples selected to compose the contrasting groups. The average number of samples per tested SNP was 5.88 in the Low and 5.82 in the High groups, ranging from three to 18 animals in both extreme groups (Table S3).

For the DASE analysis, 14,429 SNPs were analyzed in 387,919 tests, being 1479 significant tests corresponding to 937 unique DASE SNPs (FDR ≤ 0.05). Focusing on each trait, the number of tests ranged from 5871 to 8141, with a mean of 7053 per phenotype. The average of DASE SNPs identified per phenotype was 26.89, ranging from 11 (for the C16:0 trait) to 73 (for WBSF0). All tests and the corresponding significance are represented in Figure 3.3A and described in Table S4. Figure 3.3B represents the significant DASE SNPs number per phenotype.

FCs ranged from -28.00 to 27.97, with an average of -0.41. The three DASE SNPs with the most negative FC values were the rs524639701 (FC = -28.00; trait: C18:C15), the rs441798208 (FC = -27.93; trait: C18:C15), and the rs523288898 (FC = -26.53; trait: Na). The three DASE SNPs with the most positive values were the rs517071070 (FC = 27.97; trait: BFT), the rs520889234 (FC = 25.93; trait: C18:1:C12), and the rs443152461 (FC = 20.49; trait: C22:6). For 742 DASE SNPs the ASE was larger in the High contrasting sample group, and for 737 DASE SNPs this pattern was found in the Low group. Figure 3.3B represents the number of DASE SNPs according to which phenotype extreme group exhibited ASE.

We used SNPEff to annotate DASE SNPs. We found that DASE SNPs are mostly located in synonymous variation regions (n = 486 occurrences), followed by 3’ UTR regions (n = 318), downstream regions (n = 248), and 152 DASE SNPs were predicted as missense variations (Table S4).

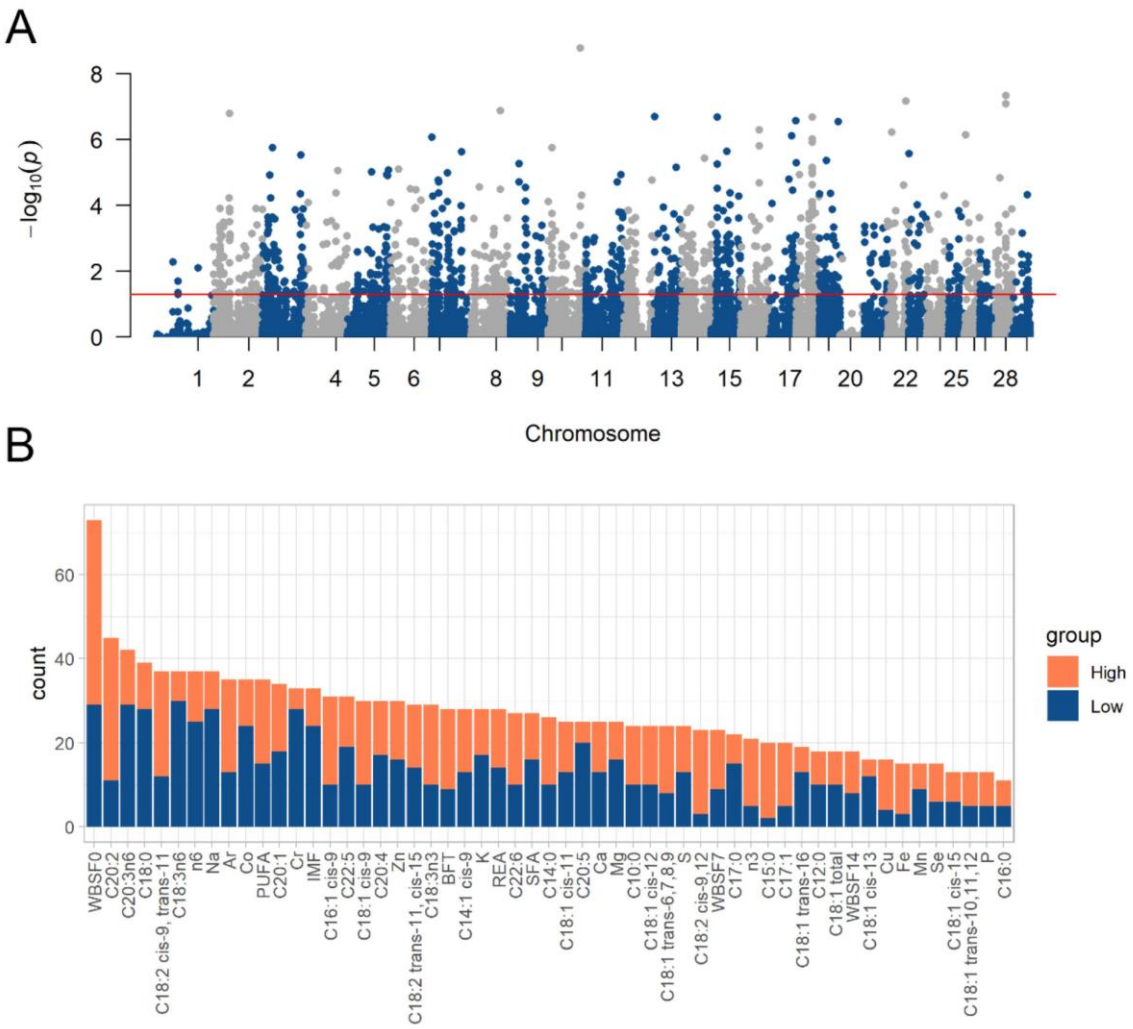


Figure 3.3 - Differential allele-specific expression analysis results. (**A**) Manhattan Plot with all DASE analysis results. The points above the red line were the significant DASE SNPs (FDR ≤0.05). (**B**) Number of DASE SNPs per trait. The colors correspond to the sample group that presented more allelic imbalance: Orange color represent the number of DASE SNPs with more ASE in the high group (positive Log2FoldChange values), and the blue color represents the ones with more ASE in the low group (negative Log2FoldChange values).

3.4.3. Genes with Differential Allele-Specific Expression Were Related to Muscle Homeostasis

The 937 unique DASE SNPs are within 656 genes. The average of DASE SNPs per gene was 2.25. The genes showing more DASE SNPs included Heat

shock protein family B—small—member 6 (HSPB6), ENSBTAG00000052709 (novel), and ENSBTAG00000048585 (novel), with 36, 30, and 22 DASE SNPs, respectively.

To investigate the biological relevance of the DASE genes, we analyzed the gene enrichment of KEGG metabolic pathways, biological processes, and molecular functions (Table S4). The DASE genes are related to 2246 biological processes, such as regulation of transcription by RNA polymerase II (with 5.64% of the DASE genes), regulation of transcription, DNA-templated (5.48% of the DASE genes), positive regulation of transcription by RNA polymerase II (5.03%), protein phosphorylation (4.11%), positive regulation of transcription, DNA-templated (4.11%), and protein ubiquitination (3.65%) (Table S5).

Regarding molecular functions, 22.25% of DASE genes were identified in protein binding, 16.15% metal ion binding, identical protein binding (10.36%), ATP binding (9.14%), nucleotide binding (8.84%), and other 758 molecular functions (Table S5). We used ClueGo to perform the gene enrichment analysis of the KEGG metabolic pathways. Based on the number of genes in each pathway, the largest number was found for pathways in neurodegeneration (26 DASE genes), Alzheimer disease (24), Parkinson disease (21), Prion disease (19), chemical carcinogenesis (19), and diabetic cardiomyopathy (19). Considering the percentage of all genes belonging to the pathways that overlapped our DASE genes, 16.13% of genes from the citrate cycle (TCA cycle) pathway were found in this study, followed by 13.46% of the genes from amino sugar and nucleotide sugar metabolism pathway, and 9.78% from the hypertrophic cardiomyopathy pathway. Figure 3.4A shows the enriched KEGG pathways from all DASE genes (Table S6) regarding the number and proportion of DASE genes in the pathways. The interaction of the pathways is represented in the network in Figure 3.4B.

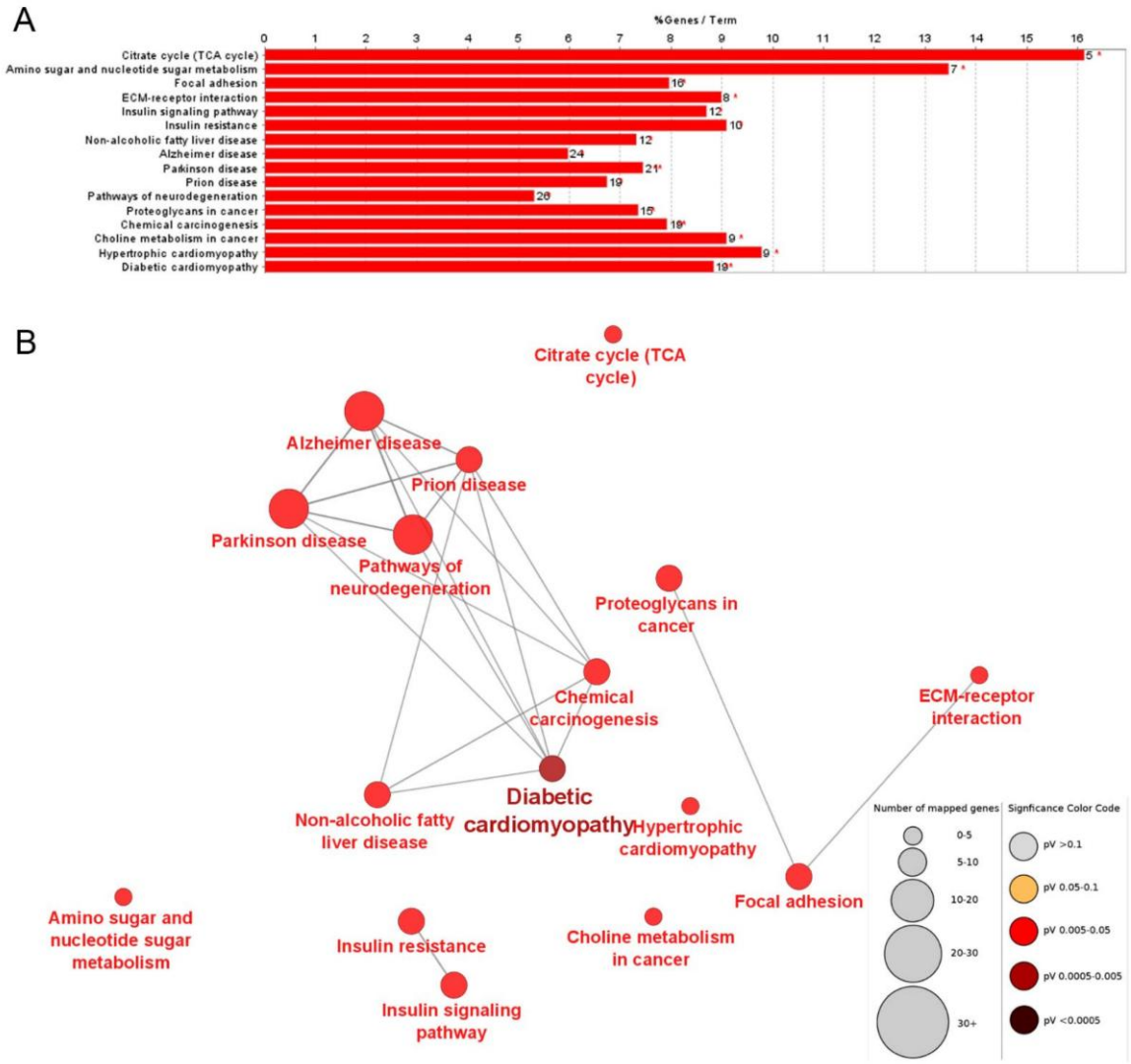


Figure 3.4 - KEGG metabolic pathways enriched in the DASE genes. (A) Enrichment of KEGG biological pathways showing the number and proportion of DASE genes in the given pathways. (B) Interaction network of the enriched KEGG pathways.

3.4.4. Co-Expression between DASE Genes

We conducted a co-expression analysis to identify modules of correlated genes that can affect the studied phenotypes. We found five co-expressed modules (M1, M2, M3, M4, and M5). The M1 module gathered 225 DASE genes, with *NFAT5, UQCR11, UTRN, NCOA3*, and *KDM7A* as hub genes; M2 had 98 DASE genes, with the *SMAD5, UBE2W, OXR1, EDEM3*, and *ARID4A* hubs; M3 had 69 DASE genes, with *MMADHC, NDUFA5, LRRC39, FASTKD2*, and *RTN4* hub genes; 60 DASE

genes are within the module M4, with *FANCG, CCNL2, RTEL1, STARD3,* and *CXXC1* hubs. M5 had 42 co-expressed DASE genes, being VIM, ANXA5, SPARC, ANXA2, and *COL3A1* its hub genes. A total of 145 DASE genes were not correlated. Modules and paired interacting genes are in Table S7.

To examine the biological relevance of the co-expressed genes, we analyzed the gene ontology biological processes (BPs) of the DASE genes, corresponding to the ORA (Table S8).

This analysis was not significant for M1, M3, M4, and M5. For M2, the protein ubiquitination process was enriched with FDR = 0.07, with ten DASE genes. Due to the greatest number of DASE genes (n = 56), we selected the WBSF0 trait for GSEA analysis. The GSEA was performed to explore if modules are upregulated or downregulated in response to this phenotype based on the extreme sample groups. WBSF0 showed significantly enriched gene sets for the modules M1 and M2 in the High and Low groups (Figure 3.5 and Table S9). These modules presented opposite effects in each sample group, with the M2 module induced in the High group and repressed in the Low group. Similarly, the M1 module showed repression in the High group and induction in the Low group.

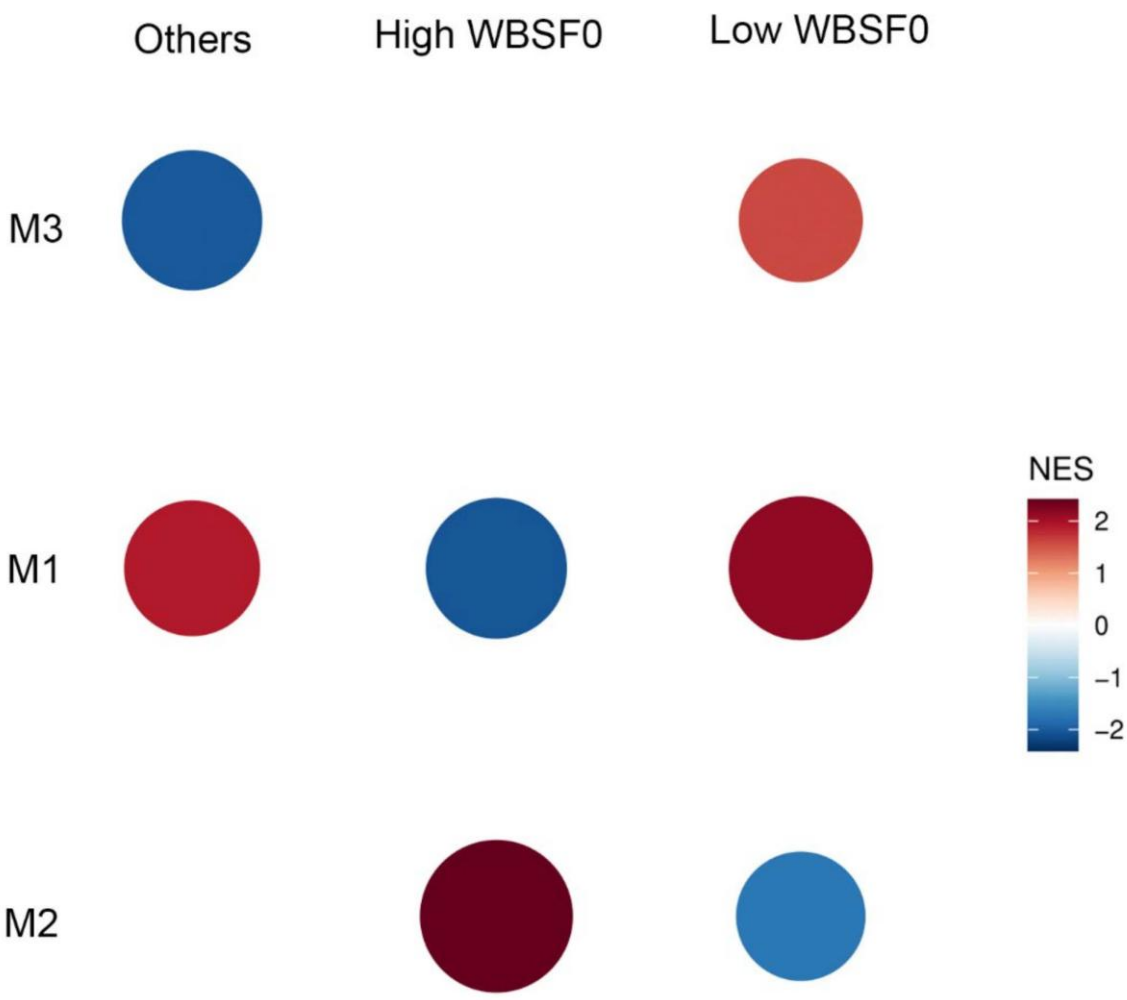


Figure 3.5 - Gene set enrichment analysis of co-expression modules involving DASE genes for WBSF0 phenotype.

3.4.5 Integration Was Found between DASE SNPs and QTLs for the Related Phenotypes

Integrating GWAS data obtained in the same Nelore population and the DASE SNPs, retrieved 353 overlaps, involving 22 traits and 129 DASE SNPs (Table S10). The top three QTLs with more overlap with DASE SNPs were REA (n = 86), LFAT (n = 49), and C18:0 (n = 41). Furthermore, the rs134535828 DASE SNP was identified as a *cis*-eQTL in the same population. Likewise, the rs136209194 DASE SNP was previously found as an aseQTL [9] (Table S10). Four variants were identified as DASE SNPs and QTLs for C18:0, three for REA, two for WBSF7, and

two DASE SNPs for WBSF0 were identified as QTLs for WBSF7, as described in Table 3.1.

Table 3.1 - DASE SNPs overlapping QTLs associated with the same phenotypes in both analyses.

DASE SNP	Trait (DASE Analysis)	Trait (QTL Integration)
rs720456892, rs43664623, chr11:63468724 G>A, chr11:69714613 G>A	C18:0	C18:0
rs716062365, rs41721088, rs109947761	REA	REA
rs109976566, rs444293703	WBSF7	WBSF7
rs208417619, rs720858445	WBSF0	WBSF7

DEGs were also found with the identified DASE SNPs. We found 159 overlaps, between 16 traits and 136 genes containing 196 DASE SNPs (Table S10). Three DASE genes, SAM and SH3 domain containing 1 (*SASH1*), Solute carrier family 25 member 4 (*SLC25A4*), and Diacylglycerol kinase delta (*DGKD*), identified for C18:1 cis-9 were differentially expressed for oleic acid content (*i.e.*, the same fatty acid). Kelch-like family member 40 (*KLHL40*) was a DEG for palmitic acid (C16:0) and a DASE gene for the same trait. Striatin Interacting Protein 2 (*STRIP2*) was differentially expressed for REA [21] and has the rs137477165 DASE SNP identified here for REA.

3.4.6 Regulatory Variants Related to the DASE Pattern

We compared DASE SNPs with data from our previous study [9]. The ASE SNPs were tested as *cis*-eQTLs, and SNPs until 200 kb distants of the ASE SNPs were tested as aseQTLs. We found all DASE SNPs overlapping ASE SNPs from our previous study [9], 23 of them classified as aseQTLs and 47 as *cis*-eQTLs (Table S10). Since regulatory mechanisms can act in variants neighboring the region with ASE, we searched for potential regulatory variants until 200 kb from the DASE SNPs. We found 9857 aseQTLs and 1274 *cis*-eQTLs in this genomic window, around 302 and 289 DASE SNPs, respectively, from our recent study [9]. De Souza et al. (2022) used a subset of 12 animals from the same population to obtain the methylation profile in the muscle, identifying regions with differentially methylated (DM) SNPs

across contrasting groups for the tenderness phenotype. We found 50 DASE SNPs located until 200 kb of distance from these DM regions (Table S10).

The DASE SNPs and potentially regulatory SNPs until 200 kb around them (DM SNPs, *cis*-eQTLs, and aseQTLs) were compared with FAANG's ChIP-seq and ATAQ-seq data [39]. We found 182 DASE SNPs overlapping functional regions (Table S10). From the overlapped DASE SNPs, 21 were previously identified as QTLs in the same population, affecting traits such as REA (15 SNPs), BFAT, C18:0 and LFAT (five SNPs), WBSF7 and WHC (four SNPs). From all potential regulatory variants until 200 kb of distance from the DASE SNPs, 1145 aseQTLs, 212 cis-eQTLs and 14 DM SNPs were located within functional regions from FAANG (Table S10). The permutation tests conducted with RegioneR indicated that the data integration was significantly more frequent than by chance (Supplementary Data S1).

We compared all features overlapping DASE SNPs to evaluate the evidence of the regulatory potential on the transcription and phenotype (Table S11). Some SNPs were found in multiple datasets (Table S11). The rs519474617, identified as aseQTLs and *cis*-eQTLs [9], is located in a region presenting all epigenetic marks from FAANG's study obtained in bovine muscle [39]. The rs717173356, found as a *cis*-eQTL and aseQTL in our previous study [9], is located in a CTCF binding region from Kern et al. (2021). The rs470985689 DASE SNP, which showed DASE for the Ar content, is located in accessible chromatin and CTCF regions [39], integrated with aseQTLs [9] and was previously identified as a QTL for REA [15].

DASE SNPs of DEGs overlapping FAANG's epigenetic data are also integrated with potentially regulatory SNPs (Tables S10 and S11). The main SNPs in this context are: five variants overlapping meat-quality QTLs, five overlapping aseQTLs and seven with *cis*-eQTLs, as described in Table 3.2.

Table 3.2 - DASE SNPs located in DEGs which overlapped epigenetic data and putative regulatory variants.

DASE SNP	Variant Annotation / Gene	Trait (DASE Analysis)	Trait (DE)	FAANG Trait (QTL)	aseQTL	ciseQTL
rs457578905	synonymous_variant / THBS4	Cr	P, Cu, Na, K, Mg, Average daily gain	ATAC-seq LFAT	No	No
rs718541482	3_prime_UTR_variant / WDR48	PUFA	CLA-c9t11, PA	ATAC-seq REA	No	No
rs211442363	3_prime_UTR_variant / SPARC	C20:1	IMF	H3K27ac REA	No	No
rs715527852	downstream_gene_variant / SUGT1	P, S	CLA-c9t11	ATAC-seq REA	No	No
rs1115255230	synonymous_variant / SPARC	n6, C20:1, C10:0	IMF	H3K27ac REA	No	No
rs209388096	synonymous_variant / AMFR	C22:6	REA	ATAC-seq	No	Yes
rs719946630	synonymous_variant / PGD	WBSF14	CLA-c9t11	ATAC-seq	No	Yes
rs1117381943	upstream_gene_variant / CAVIN2	C18_1_T6_T7_T8_T9	AO	H3K4me3, H3K27ac, H3K4me1, H3K4me3, ATAC-seq, CTCF	No	Yes
rs516592412	synonymous_variant / PDG	C20:1	CLA-c9t11	ATAC-seq	No	Yes
rs1117068355	upstream_gene_variant / CAVIN2	C18_1_T6_T7_T8_T9	AO	H3K4me3, H3K4me1, H3K27me3, H3K27ac, ATAC-seq, CTCF	No	Yes
rs109090536	5_prime_UTR_variant / DCAF11	C20:2	AO	H3K4me3, H3K4me1, H3K27ac	No	Yes
rs443738741	synonymous_variant / LNX2	C20:2	REA	ATAC-seq	No	Yes

rs379719524	synonymous_variant / NRAP	Cr	AO	ATAC-seq	No	Yes
rs715652252	synonymous_variant / LNX2	C20:2	REA	ATAC-seq, CTCF	No	Yes
rs468833876	upstream_gene_variant / NUTF2	C18_2_C9_C12	CLA-c9t11	H3K27ac, H3K27me3, H3K4me3	No	Yes
rs135451771	synonymous_variant / NRAP	Cr	AO	ATAC-seq	No	Yes
chr2:5447956 G>A	synonymous_variant / BIN1	C18_1_T6_T7_T8_T9	AO	H3K27ac	No	Yes

3.5 Discussion

The information on correlated phenotypes or the traits belonging to the same cluster is important to investigate the pleiotropic effect of allelic imbalance and understand DASE SNPs eventually identified for multiple meat quality traits. By correlating all phenotypes in our population, we found phenotypic interactions making biological sense. The Ca, Mg, and Zn minerals showed a positive correlation here and were associated with human muscle mass [41]. Zinc is an essential mineral for lean body mass synthesis, but relatively large amounts of Zn are needed for new tissue synthesis [42]. Moreover, we found positive correlations between unsaturated fatty acids and Cu and Cr. Chromium methionine chelate supplementation in Holstein steers increased the C20:4 (p-value = 0.07) and PUFA (p-value = 0.04) in beef during late fattening period [43].

Negative correlations were found between saturated fatty acid (SFA) composition and WBSF traits, which means that tender samples had larger concentrations of SFA. SFA can be related to initial and sustained juiciness in beef [44]. Although the excessive ingestion of SFA can be detrimental to the consumer's health, they are still related to tenderness and juiciness in the bovine muscle [44,45]. To bring a balance to the production aspects, selecting a fatty acid composition in the herd must help the beef quality without increasing undesired fatty acids, which harm the consumers' health.

The selection of contrasting samples for DASE analysis used almost all initially available animals (187 samples out of 190). Nevertheless, the DASE analysis was still conservative due to the limited number of animals in each contrasting group presenting collected phenotypes and a transcribed heterozygous SNP. The number of animals tested for DASE ranged from three to 18 per contrasting group. The statistical power of DASE analyses would be improved in larger experimental populations, in which it is possible to test larger contrasting groups because of the increased probability of heterozygous SNPs able to be tested. However, requiring at least three animals in each group may have guaranteed confidence in the obtained results without losing too much information due to our sample size.

Comparing the number of significant results with the ones from our previous work [9], we found that only 2.45% of the ASE SNPs were identified here as DASE SNPs. The proportion of tested SNPs in DASE analysis was 1.06% of the SNPs from the original dataset (~4 M SNPs), compared with the 3.02% tested for ASE [9]. Firstly, we did not expect all ASE SNPs to be classified as DASE SNPs because it is improbable that they all were located in genes related to the studied phenotypes. Moreover, the ASE results are obtained per animal, with multiple ASE SNPs being significant in a small proportion of the population, thus resulting in a larger number of significant results compared to the DASE analysis. Our analysis depends on the phenotypic variation, so it was performed in a selected sample rather than the whole population, resulting in a reduced number of significant results. The annotation of DASE and ASE SNPs was similar, mostly located in synonymous variation, downstream, intronic, and 3’ UTR regions.

DASE genes identified as hubs of the co-expression analysis were underlying metabolic pathways. ORA enriched for the protein ubiquitination process in M2. The ubiquitination process marks the proteins to be degraded in the cell cycle, requiring some calcium-dependent enzymes [46]. C18:2 increases ubiquitination and has a potential role in the proteasomal degradation of tyrosinase [47]. We found 15 DASE genes associated to C18:2 traits co-expressed in M2, being seven related to C18:2 cis-9, trans-11 (*ACIN1, COIL, DGKD, MED23, SMARCA4, TMEM167A*, and *RNF20*), five DASE genes for C18:2 trans-11, cis-15 (*ITGB1, RGSS, SCAP, BMS1*,

and *CCNL2*) and three for C18:2 cis-9,12 trait (*EXOC1, NUTF2* and *TMX3*). From pig muscle, polymorphisms of ubiquitin protein ligase E3C were associated with IMF and fatty acid composition, being known as the relationship of E3 ubiquitin proteins and lipid metabolism [48]. Beef tenderness was related to proteasome proteolysis, which in turn, requires protein ubiquitination [49]. These results suggest an interplay between beef tenderness and fatty acid content in muscle, especially for C18:2, via ubiquitination-dependent proteasomal degradation.

To evaluate the enrichment of genes belonging to modules between the sample groups, the GSEA was performed. We found M1 and M2 modules with opposite effects in contrasting groups of the WBSF0 trait. The M1 module showed positive activity in the Low WBSF0 group (more tender samples). Cytochrome b-c1 complex subunit 10 (*UQCR11*), a hub gene of M1, codes to the third complex of the mitochondria's electron transport chain, and showed DASE for SFA, C18:3n6, C20:5, and PUFA. Fatty acids are related to mitochondrial membrane permeability [50]. After slaughter, under no blood circulation, mitochondria activity is impaired but still may influence biochemical processes in postmortem muscle, which in turn requires oxygen from myoglobin to produce ATP [51]. UCQCR11 converts ubiquinol in cytochrome c [52]. The release of cytochrome c induces apoptosis [51,53]. Apoptosis is a proteolytic process related to meat tenderization under oxidative stress [49,54,55], which can result from increased mitochondrial membrane permeability [50]. Utrophin, encoded by *UTRN*, the hub gene of M1, is a therapeutic target for Duchenne muscular disease [56,57]. The upregulation of this protein reduces oxidative stress and mitochondrial pathology [57], and this protein was related to the homeostasis of the sodium channel in the heart [56], attenuating the Duchenne's pathology [56]. Among the 15 genes within M1 that showed DASE for the same enriched trait (WBSF0), there are Platelet-Derived Growth Factor Receptor Alpha (*PDGFRA*), previously identified as a DEG for Ca [22], and Prune Homolog 2With BCH Domain (*PRUNE2*), differentially expressed in extremes of RFI [24].

The M2 module showed positive activity in the High WBSF0 group (less tender samples), a reverse pattern compared with the M1 module. *SMAD5* is a hub gene of the M2 module, involved in 39 BPs (see Table S4) being its product a

transcription factor part of BMP signaling [58], exhibiting a role in myogenesis and muscle growth [58,59]. Oxidation resistance 1 (*OXR1*) was associated with growth traits in Wagyu cattle [60] and was found as a hub gene in our co-expression analysis. *OXR1* has emerged as an essential antioxidant protein that controls neurons' susceptibility to oxidative stress. The OXR1 protein is involved in maintaining mitochondrial morphology in the stress response as part of its antioxidant function [61]. This protein has orthologs between several species and is conserved throughout evolution [62]. Thus, the gene set of M1 seems to be related to meat tenderness, with positive activity in the tender group and the gene set of M2 have a role in muscle growth. These genes are probable candidates for improving beef quality and production, and the causal regulatory variants can be linked to the respective DASE SNPs.

Some DASE genes were associated with the same phenotypes affected by differentially expressed genes previously identified in the same population. These findings suggest that the total expression of these DASE/DE genes can be affected by cis-regulatory mechanisms acting over the given DASE SNP region. We identified three genes, SAM And SH3 Domain Containing 1 (*SASH1*), Solute Carrier Family 25 Member 4 (*SLC25A4*), and Diacylglycerol Kinase Delta (*DGKD*) associated here and previously DEG with oleic acid content (C18:1 cis9) [19]. *SLC25A4* is a DEG up-regulated in the Low group for oleic acid content, showing ASE in the Low group in this work for its unique DASE SNP - rs109988743, located in a region with H3K4me3, H3K4me1, H3K27ac, and H3K27me3 peaks reported previously in bovine muscle [39]. *KLHL40* was found as DEG up-regulated in the Low group for palmitic acid (C16:0) [19], and its rs515929286 DASE SNP was found here with more ASE in the Low group, being located in an accessible chromatin region [39]. *STRIP2* was down-regulated in the Low group of REA [21] and showed ASE in the Low group of the same trait here for the rs137477165 DASE SNP, located in a functional region of H3K27ac histone modification [39].

The DASE SNPs pointed out here may be candidates to improve oleic acid content and REA in bovine muscle, and the variants regulating their transcription warrant further investigation. The integration of DASE results and regulatory data

revealed multiple functional variants overlapping or neighboring DASE SNPs, which might be responsible for the allelic expression pattern and phenotype. This study indicated that the DASE SNPs may not be exclusively markers for conditional-dependent ASE but also regulatory elements or in linkage disequilibrium with the causal variants. We showed DASE SNPs classified previously as cis-eQTLs and aseQTLs, such as rs519474617 and rs717173356, located in accessible chromatin regions and CTCF-binding regions, indicating CTCF-mediated transcription regulation [63]. DASE SNPs of previously described DEGs also integrated with multiple targets from FAANG's study [39] and with aseQTLs, cis-eQTLs, and QTLs obtained in the same population, such as rs1117068355 and rs1117381943 DASE SNPs/aseQTLs, rs468833876 (DASE SNP/cis-eQTL) and rs457578905 (DASE SNP/QTL). Thus, DASE regions are co-localized with functional SNPs, which are useful for discovering regulatory variants that affect gene expression and phenotypes.

Other studies can take advantage of the DASE results, highlighting which contrasting group presented ASE for the intended phenotype to track regulatory variants close to the DASE SNP or even admitting its own regulatory potential. The DASE analysis is a helpful one-step approach for identifying genomic regions simultaneously associated with gene expression and phenotypic variance, as long as a large experimental population is available. From the DASE analysis, multiple potentially functional variants were identified located in previously reported regulatory regions. Because of the predicted association between DASE genes and meat quality traits, the identified DASE SNPs can be further explored to be applied in beef improvement programs.

References

1. Castel, S.E.; Levy-Moonshine, A.; Mohammadi, P.; Banks, E.; Lappalainen, T. Tools and Best Practices for Data Processing in Allelic Expression Analysis. Genome Biol. **2015**, 16, 195. [CrossRef] [PubMed]

2. Zou, J.; Hormozdiari, F.; Jew, B.; Castel, S.E.; Lappalainen, T.; Ernst, J.; Sul, J.H.; Eskin, E. Leveraging Allelic Imbalance to Refine Fine-Mapping for EQTL Studies. PLoS Genet. **2019**, 15, e1008481. [CrossRef]

3. Pfeifer, K.; Tilghman, S.M. Allele-Specific Gene Expression in Mammals: The Curious Case of the Imprinted RNAs. Genes Dev. **1994**, 8, 1867–1874. [CrossRef]

4. Chamberlain, A.J.; Vander Jagt, C.J.; Hayes, B.J.; Khansefid, M.; Marett, L.C.; Millen, C.A.; Nguyen, T.T.T.; Goddard, M.E. Extensive Variation between Tissues in Allele Specific Expression in an Outbred Mammal. BMC Genom. **2015**, 16, 1–20. [CrossRef]

5. Huang, W.-C.; Ferris, E.; Cheng, T.; Hörndli, C.S.; Gleason, K.; Tamminga, C.; Wagner, J.D.; Boucher, K.M.; Christian, J.L.; Gregg, C. Diverse Non-Genetic, Allele-Specific Expression Effects Shape Genetic Architecture at the Cellular Level in the Mammalian Brain. Neuron **2017**, 93, 1094–1109.e7. [CrossRef] [PubMed]

6. Khansefid, M.; Pryce, J.E.; Bolormaa, S.; Chen, Y.; Millen, C.A.; Chamberlain, A.J.; Vander Jagt, C.J.; Goddard, M.E. Comparing Allele Specific Expression and Local Expression Quantitative Trait Loci and the Influence of Gene Expression on Complex Trait Variation in Cattle. BMC Genom. **2018**, 19, 1–18. [CrossRef]

7. Gao, C.; Devarajan, K.; Zhou, Y.; Slater, C.M.; Daly, M.B.; Chen, X. Identifying Breast Cancer Risk Loci by Global Differential Allele-Specific Expression (DASE) Analysis in Mammary Epithelial Transcriptome. BMC Genom. **2012**, 13, 570. [CrossRef] [PubMed]

8. de Souza, M.M.; Zerlotini, A.; Rocha, M.I.P.; Bruscadin, J.J.; Diniz, W.J.d.S.; Cardoso, T.F.; Cesar, A.S.M.; Afonso, J.; Andrade, B.G.N.; Mudadu, M.d.A.; et al. Allele-Specific Expression Is Widespread in Bos Indicus Muscle and Affects Meat Quality Candidate Genes. Sci. Rep. **2020**, 10, 10204. [CrossRef]

9. Bruscadin, J.J.; Cardoso, T.F.; da Silva Diniz, W.J.; Afonso, J.; de Souza, M.M.; Petrini, J.; de Andrade, B.G.; da Silva, V.H.; Ferraz, J.B.S.; Zerlotini, A.; et al. Allele-Specific Expression Reveals Functional SNPs Affecting Muscle-Related Genes in Bovine. Biochim. Biophys. Acta (BBA) **2022**, 194886. [CrossRef]

10. Guillocheau, G.M.; El Hou, A.; Meersseman, C.; Esquerré, D.; Rebours, E.; Letaief, R.; Simao, M.; Hypolite, N.; Bourneuf, E.; Bruneau, N.; et al. Survey of Allele Specific Expression in Bovine Muscle. Sci. Rep. **2019**, 9, 4297. [CrossRef] [PubMed]

11. Varkoohi, S.; Banabazi, M.H.; Ghsemi-Siab, M. Allele Specific Expression (Ase) Analysis between Bos Taurus and Bos Indicus Cows Using Rna-Seq Data at Snp Level and Gene Level. An. Acad. Bras. DeCienc. **2021**, 93, 1–9. [CrossRef] [PubMed]

12. McGettigan, P.A. Transcriptomics in the RNA-Seq Era. Curr. Opin. Chem. Biol. **2013**, 17, 4–11. [CrossRef] [PubMed] 13. Przytycki, P.F.; Singh, M. Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations. Cell Syst. **2020**, 10, 193–203.e4. [CrossRef] [PubMed]

14. Tizioto, P.C.; Meirelles, S.L.; Veneroni, G.B.; Tullio, R.R.; Rosa, A.N.; Alencar, M.M.; Medeiros, S.R.; Siqueira, F.; Feijó, G.L.D.; Silva, L.O.C.; et al. A SNP in ASAP1 Gene Is Associated with Meat Quality and Production Traits in Nelore Breed. Meat Sci. **2012**, 92, 855–857. [CrossRef]

15. Tizioto, P.C.; Decker, J.E.; Taylor, J.F.; Schnabel, R.D.; Mudadu, M.A.; Silva, F.L.; Mourao, G.B.; Coutinho, L.L.; Tholon, P.; Sonstegard, T.S.; et al. Genome Scan for Meat Quality Traits in Nelore Beef Cattle. Physiol. Genom. **2013**, 45, 1012–1020. [CrossRef]

16. Cesar, A.S.; Regitano, L.C.; Mourão, G.B.; Tullio, R.R.; Lanna, D.P.; Nassu, R.T.; Mudado, M.A.; Oliveira, P.S.; do Nascimento, M.L.; Chaves, A.S.; et al. Genome-Wide Association Study for Intramuscular Fat Deposition and Composition in Nellore Cattle. BMC Genet. **2014**, 15, 39. [CrossRef]

17. Tizioto, P.C.; Gromboni, C.F.; de Araujo Nogueira, A.R.; de Souza, M.M.; de Alvarenga Mudadu, M.; Tholon, P.; Rosa, A.d.N.; Tullio, R.R.; Medeiros, S.R.; Nassu, R.T.; et al. Calcium and Potassium Content in Beef: Influences on Tenderness and Associations with Molecular Markers in Nellore Cattle. Meat Sci. **2014**, 96, 436–440. [CrossRef]

18. Gonçalves, T.M.; de Almeida Regitano, L.C.; Koltes, J.E.; Cesar, A.S.M.; da Silva Andrade, S.C.; Mourão, G.B.; Gasparin, G.; Moreira, G.C.M.; Fritz-Waters, E.; Reecy, J.M.; et al. Gene Co-Expression Analysis Indicates Potential Pathways and Regulators of Beef Tenderness in Nellore Cattle. Front. Genet. **2018**, 9, 441. [CrossRef]

19. Cesar, A.S.M.; Regitano, L.C.A.; Poleti, M.D.; Andrade, S.C.S.; Tizioto, P.C.; Oliveira, P.S.N.; Felício, A.M.; do Nascimento, M.L.; Chaves, A.S.; Lanna, D.P.D.; et al. Differences in the Skeletal Muscle Transcriptome Profile Associated with Extreme Values of Fatty Acids Content. BMC Genom. **2016**, 17, 961. [CrossRef]

20. Diniz,W.J.d.S.; Coutinho, L.L.; Tizioto, P.C.; Cesar, A.S.M.; Gromboni, C.F.; Nogueira, A.R.A.; de Oliveira, P.S.N.; Souza, M.M.d.; Regitano, L.C.d.A. Iron Content Affects Lipogenic Gene Expression in the Muscle of Nelore Beef Cattle. PLoS ONE **2016**, 11, e0161160. [CrossRef]

21. Silva-Vignato, B.; Coutinho, L.L.; Cesar, A.S.M.; Poleti, M.D.; Regitano, L.C.A.; Balieiro, J.C.C. Comparative Muscle Transcriptome Associated with Carcass Traits of Nellore Cattle. BMC Genom. **2017**, 18, 506. [CrossRef] [PubMed]

22. Afonso, J.; Coutinho, L.L.; Tizioto, P.C.; da Silva Diniz, W.J.; de Lima, A.O.; Rocha, M.I.P.; Buss, C.E.; Andrade, B.G.N.; Piaya, O.; da Silva, J.V.; et al. Muscle Transcriptome Analysis Reveals Genes and Metabolic Pathways Related to Mineral Concentration in Bos Indicus. Sci. Rep. **2019**, 9, 12715. [CrossRef]

23. Cardoso, T.F.; Coutinho, L.L.; Bruscadin, J.J.; da Silva Diniz,W.J.; Petrini, J.; Andrade, B.G.N.; de Oliveira, P.S.N.; Poleti, M.D.; Cesar, A.S.M.; da Silveira, J.C.; et al. Multi-Omics Approach Reveals MiR-SNPs Affecting Muscle Fatty Acids Profile in Nelore Cattle. Genes **2021**, 12, 67. [CrossRef]

24. Mudadu, M.A.; Porto-Neto, L.R.; Mokry, F.B.; Tizioto, P.C.; Oliveira, P.S.N.; Tullio, R.R.; Nassu, R.T.; Niciura, S.C.M.; Tholon, P.; Alencar, M.M.; et al. Genomic Structure and Marker-Derived Gene Networks for Growth and Meat Quality Traits of Brazilian Nelore Beef Cattle. BMC Genom. **2016**, 17, 1–16. [CrossRef]

25. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.R.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.W.; Daly, M.J.; et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. **2007**, 81, 559–575. [CrossRef] [PubMed]

26. Zhbannikov, I.Y.; Hunter, S.S.; Foster, J.A.; Settles, M.L. SeqyClean: A Pipeline for High-Throughput Sequence Data Preprocessing. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Boston, MA, USA, 20 August 2017; pp. 407–416.

27. Cesar, A.S.M.; Regitano, L.C.A.; Reecy, J.M.; Poleti, M.D.; Oliveira, P.S.N.; de Oliveira, G.B.; Moreira, G.C.M.; Mudadu, M.A.; Tizioto, P.C.; Koltes, J.E.; et al. Identification of Putative Regulatory Regions and Transcription Factors Associated with Intramuscular Fat Content Traits. BMC Genom. **2018**, 19, 499. [CrossRef]

28. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast Universal RNA-Seq Aligner. Bioinformatics **2013**, 29, 15–21. [CrossRef] [PubMed] 29. Van De Geijn, B.; Mcvicker, G.; Gilad, Y.; Pritchard, J.K.WASP: Allele-Specific Software for Robust Molecular Quantitative Trait Locus Discovery. Nat. Methods **2015**, 12, 1061–1063. [CrossRef]

30. Degner, J.F.; Marioni, J.C.; Pai, A.A.; Pickrell, J.K.; Nkadori, E.; Gilad, Y.; Pritchard, J.K. Effect of Read-Mapping Biases on Detecting Allele-Specific Expression from RNA-Sequencing Data. Bioinformatics **2009**, 25, 3207–3212. [CrossRef]

31. Anders, S.; Pyl, P.T.; Huber,W. HTSeq–a Python Framework toWork with High-Throughput Sequencing Data. Bioinformatics **2015**, 31, 166–169. [CrossRef]

32. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data. Genome Res. **2010**, 20, 1297–1303. [CrossRef] [PubMed]

33. Love, M.I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. Genome Biol. **2014**, 15, 550. [CrossRef]

34. Cingolani, P.; Platts, A.;Wang, L.L.; Coon, M.; Nguyen, T.;Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff. Fly **2012**, 6, 80–92. [CrossRef] [PubMed]

35. Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W.-H.; Pagès, F.; Trajanoski, Z.; Galon, J. ClueGO: A Cytoscape Plug-in to Decipher Functionally Grouped Gene Ontology and Pathway Annotation Networks. Bioinformatics **2009**, 25, 1091–1093. [CrossRef] [PubMed]

36. Russo, P.S.T.; Ferreira, G.R.; Cardozo, L.E.; Bürger, M.C.; Arias-Carrasco, R.; Maruyama, S.R.; Hirata, T.D.C.; Lima, D.S.; Passos, F.M.; Fukutani, K.F.; et al. CEMiTool: A Bioconductor Package for Performing Comprehensive Modular Co-Expression Analyses. BMC Bioinform. **2018**, 19, 56. [CrossRef]

37. Tizioto, P.C.; Taylor, J.F.; Decker, J.E.; Gromboni, C.F.; Mudadu, M.A.; Schnabel, R.D.; Coutinho, L.L.; Mourão, G.B.; Oliveira, P.S.; Souza, M.M.; et al. Detection of Quantitative Trait Loci for Mineral Content of Nelore Longissimus Dorsi Muscle. Genet. Sel. Evol. **2015**, 47, 15. [CrossRef] [PubMed]

38. 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.; Mourão, G.B.; et al. DNA Methylation May Affect Beef Tenderness through Signal Transduction in Bos Indicus. Epigenetics Chromatin **2022**, 15, 15. [CrossRef]

39. Kern, C.; Wang, Y.; Xu, X.; Pan, Z.; Halstead, M.; Chanthavixay, G.; Saelao, P.; Waters, S.; Xiang, R.; Chamberlain, A.; et al. Functional Annotations of Three Domestic Animal Genomes Provide Vital Resources for Comparative and Agricultural Research. Nat. Commun. **2021**, 12, 1821. [CrossRef]

40. Gel, B.; Díez-Villanueva, A.; Serra, E.; Buschbeck, M.; Peinado, M.A.; Malinverni, R. RegioneR: An R/Bioconductor Package for the Association Analysis of Genomic Regions Based on Permutation Tests. Bioinformatics **2015**, btv562. [CrossRef] [PubMed]

41. van Dronkelaar, C.; van Velzen, A.; Abdelrazek, M.; van der Steen, A.;Weijs, P.J.M.; Tieland, M. Minerals and Sarcopenia; The Role of Calcium, Iron, Magnesium, Phosphorus, Potassium, Selenium, Sodium, and Zinc on Muscle Mass, Muscle Strength, and Physical Performance in Older Adults: A Systematic Review. J. Am. Med Dir. Assoc. **2018**, 19, 6–11.e3. [CrossRef]

42. Cunha, T.A.; Vermeulen-Serpa, K.M.; Grilo, E.C.; Leite-Lais, L.; Brandão-Neto, J.; Vale, S.H.L. Association between Zinc and Body Composition: An Integrative Review. J. Trace Elem. Med. Biol. **2022**, 71, 126940. [CrossRef] [PubMed]

43. Ghassemi Nejad, J.; Lee, B.-H.; Kim, B.-W.; Ohh, S.-J.; Sung, K.I. Effects of Chromium Methionine Supplementation on Blood Metabolites and Fatty Acid Profile of Beef during Late Fattening Period in Holstein Steers. Asian-Australas. J. Anim. Sci. **2016**, 29, 378–383. [CrossRef]

44. Garmyn, A.J.; Hilton, G.G.; Mateescu, R.G.; Morgan, J.B.; Reecy, J.M.; Tait, R.G.; Beitz, D.C.; Duan, Q.; Schoonmaker, J.P.; Mayes, M.S.; et al. Estimation of Relationships between Mineral Concentration and Fatty Acid Composition of Longissimus Muscle and Beef Palatability Traits1. J. Anim. Sci. **2011**, 89, 2849–2858. [CrossRef] [PubMed]

45. Listrat, A.; Gagaoua, M.; Andueza, D.; Gruffat, D.; Normand, J.; Mairesse, G.; Picard, B.; Hocquette, J.-F. What Are the Drivers of Beef Sensory Quality Using Metadata of Intramuscular Connective Tissue, Fatty Acids and Muscle Fiber Characteristics? Livest. Sci. **2020**, 240, 104209. [CrossRef]

46. Mukherjee, R.; Das, A.; Chakrabarti, S.; Chakrabarti, O. Calcium Dependent Regulation of Protein Ubiquitination–Interplay between E3 Ligases and Calcium Binding Proteins. BBA-Mol. Cell Res. **2017**, 1864, 1227–1235. [CrossRef]

47. Ando, H.; Watabe, H.; Valencia, J.C.; Yasumoto, K.; Furumura, M.; Funasaka, Y.; Oka, M.; Ichihashi, M.; Hearing, V.J. Fatty Acids Regulate Pigmentation via Proteasomal Degradation of Tyrosinase. J. Biol. Chem. **2004**, 279, 15427–15433. [CrossRef] [PubMed]

48. Supakankul, P.; Mekchay, S. Effect of UBE3C Polymorphisms on Intramuscular Fat Content and Fatty Acid Composition in Duroc Pigs. Genet. Mol. Res. **2016**, 15, 3. [CrossRef] [PubMed]

49. Kemp, C.M.; Sensky, P.L.; Bardsley, R.G.; Buttery, P.J.; Parr, T. Tenderness – An Enzymatic View. Meat Sci. **2010**, 84, 248–256. [CrossRef]

50. Furuno, T.; Kanno, T.; Arita, K.; Asami, M.; Utsumi, T.; Doi, Y.; Inoue, M.; Utsumi, K. Roles of Long Chain Fatty Acids and Carnitine in Mitochondrial Membrane Permeability Transition11Abbreviations: CsA, Cyclosporin A; Cyt. c, Cytochrome c;

FBS, Fetal Bovine Serum; Pal-CoA, Palmitoyl-CoA; Pi, Inorganic Phosphate; LCFA, Long Chain Fatty Acid; MPT, Membrane Permeability Transition; PLA2, Phospholipase A2; RCR, Respiratory Control Ratio; and T3, 3,30,5-Triiodothyronine. Biochem. Pharmacol. **2001**, 62, 1037–1046. [CrossRef]

51. Matarneh, S.K.; England, E.M.; Scheffler, T.L.; Gerrard, D.E. The Conversion of Muscle to Meat. In Lawrie's Meat Science; Elsevier: Amsterdam, The Netherlands, 2017; pp. 159–185. ISBN 978-0-08-100694-8. 52. Letts, J.A.; Sazanov, L.A. Clarifying the Supercomplex: The Higher-Order Organization of the Mitochondrial Electron Transport Chain. Nat. Struct. Mol. Biol. **2017**, 24, 800–808. [CrossRef]

53. Kalpage, H.A.; Wan, J.; Morse, P.T.; Zurek, M.P.; Turner, A.A.; Khobeir, A.; Yazdi, N.; Hakim, L.; Liu, J.; Vaishnav, A.; et al. Cytochrome c Phosphorylation: Control of Mitochondrial Electron Transport Chain Flux and Apoptosis. Int. J. Biochem. Cell Biol. **2020**, 121, 105704. [CrossRef] [PubMed]

54. Wang, L.-L.; Yu, Q.-L.; Han, L.; Ma, X.-L.; Song, R.-D.; Zhao, S.-N.; Zhang, W.-H. Study on the Effect of Reactive Oxygen Species-Mediated Oxidative Stress on the Activation of Mitochondrial Apoptosis and the Tenderness of Yak Meat. Food Chem. **2018**, 244, 394–402. [CrossRef]

55. Liu, C.;Wei, Q.; Li, X.; Han, D.; Liu, J.; Huang, F.; Zhang, C. Proteomic Analyses of Mitochondrial Damage in Postmortem Beef Muscles. J. Sci. Food Agric. **2022**, jsfa.11767. [CrossRef] [PubMed]

56. Albesa, M.; Ogrodnik, J.; Rougier, J.-S.; Abriel, H. Regulation of the Cardiac Sodium Channel Nav1.5 by Utrophin in Dystrophin-Deficient Mice. Cardiovasc. Res. **2011**, 89, 320–328. [CrossRef] [PubMed]

57. Kennedy, T.L.; Moir, L.; Hemming, S.; Edwards, B.; Squire, S.; Davies, K.; Guiraud, S. Utrophin Influences Mitochondrial Pathology and Oxidative Stress in Dystrophic Muscle. Skelet. Muscle **2017**, 7, 22. [CrossRef]

58. Sartori, R.; Schirwis, E.; Blaauw, B.; Bortolanza, S.; Zhao, J.; Enzo, E.; Stantzou, A.; Mouisel, E.; Toniolo, L.; Ferry, A.; et al. BMP Signaling Controls Muscle Mass. Nat. Genet. **2013**, 45, 1309–1318. [CrossRef] [PubMed]

59. Ning, Y.; Zhang, L.; Wang, W.; Wu, S. Effect of Genetic Variants in the SMAD1 and SMAD5 Genes Promoter on Growth and Beef Quality Traits in Cattle. Gene **2022**, 819, 146220. [CrossRef]

60. Zhang, R.; Miao, J.; Song, Y.; Zhang, W.; Xu, L.; Chen, Y.; Zhang, L.; Gao, H.; Zhu, B.; Li, J.; et al. Genome-Wide Association Study Identifies the PLAG1-OXR1 Region on BTA14 for Carcass Meat Yield in Cattle. Physiol. Genom. **2019**, 51, 137–144. [CrossRef]

61. Wu, Y.; Davies, K.E.; Oliver, P.L. The Antioxidant Protein Oxr1 Influences Aspects of Mitochondrial Morphology. Free Radic. Biol. Med. **2016**, 95, 255–267. [CrossRef]

62. Durand, M.; Kolpak, A.; Farrell, T.; Elliott, N.A.; Shao, W.; Brown, M.; Volkert, .R. The OXR Domain Defines a Conserved Family of Eukaryotic Oxidation Resistance Proteins. BMC Cell Biol. **2007**, 8, 13. [CrossRef]

63. Xu, B.; Wang, H.; Wright, S.; Hyle, J.; Zhang, Y.; Shao, Y.; Niu, M.; Fan, Y.; Rosikiewicz, W.; Djekidel, M.N.; et al. Acute Depletion of CTCF Rewires Genome-Wide Chromatin Accessibility. Genome Biol. **2021**, 22, 244. [CrossRef] [PubMed]

4 PARENT-OF-ORIGIN QTLS AND ALLELE-SPECIFIC METHYLATION REGIONS IDENTIFICATION

This section contains transcriptions from the paper under development, **“Parent-of-origin and allele-specific methylation effects on gene expression and muscle traits in *Bos indicus* cattle”**. The following results are based on allele-specific methylation (ASM) and POEs affecting muscle-related traits.

Supplementary Materials: The following supporting information can be downloaded at:

https://drive.google.com/drive/folders/1W3GrAI1Got4NDTtkyf8uvsxcTXa-gjPM?usp=sharing

4.1 Abstract

Parent-of-origin effects (POEs) occur when the impact of an allele depends on its parental origin; if the POE influences gene expression by silencing one parental allele, it can be classified as genomic imprinting. Allele-specific methylation (ASM) is an epigenetic mechanism that drives allele-specific expression (ASE) and genomic imprinting, which, in turn, can affect gene expression and contribute to phenotypic variance. Our previous studies identified SNPs within ASE regions (ASE SNPs) and predicted their *cis*-regulatory variants in muscle from 190 Nelore (*Bos indicus*) steers. This study expands these findings by investigating POEs and ASM regions in Nelore muscle to reveal the potential mechanisms impacting economically important traits. By applying a Robust Omnibus Test model to identify POEs from six groups of muscle-related traits, we were able to identify 1,817 parent-of-origin quantitative trait *loci* (PO-QTLs), mostly associated with polyunsaturated fatty acids concentration traits. The lack of overlapping PO-QTLs across different trait clusters suggests that PO-QTLs are phenotype-set-specific. PO-QTLs overlapping ASE SNPs and *cis*-eQTLs, were annotated on genes imprinted in bovine (*e.g.*, *C1S*) and human (*e.g.*, *RASGRF1*). To profile ASM regions genome-wide, we used reduced representation bisulfite sequence datasets from twelve steers of the same population. We identified 8,074 ASM regions, of which 269 were identified in all samples. ASM was mainly located on promoters, and many of the annotated genes overlapped imprinted genes in bovine and humans, such as *GNAS* and *MEST*, presenting more candidates for validation. A total of 111 ASM regions impacted genes with monoallelic expression determined in our previous studies, reaffirming the connection between ASM, ASE, and genomic imprinting. PO-QTLs and ASM were predicted to be affected by transcription factors and miRNA binding, emphasizing their potential regulatory implications. From predicted functional variants and the evidence of novel genomic imprinting *loci* in the bovine genome, this study offered new insights into allele-specific methylation and POEs mechanisms in cattle. It is deepening our understanding of how epigenetic modifications contribute to important genes and traits in Nelore cattle.

4.2 Introduction

Parent-of-origin effects (POE) consist of asymmetric and parental-dependent transmission and in its impact of alleles on offspring. This pattern does not follow the Mendelian inheritance, where one allele is transferred from each parent to their diploid progeny (GUILMATRE; SHARP, 2012). Genomic imprinting is the most well-known and studied POE, comprising the silenced expression of a gene in one allele inherited from a specific parent (GUILMATRE; SHARP, 2012; LAWSON; CHEVERUD; WOLF, 2013). Epigenetic mechanisms, either through parental-dependent histone modification or DNA methylation in CpG-rich areas, are correlated with *imprinting control regions* and dictate this pattern during the gametogenesis (JIRTLE; WEIDMAN, 2007; NEUGEBAUER et al., 2010).

Genome-wide association studies (GWAS) have been developed to identify quantitative trait locus (QTLs) associated with phenotypic variance; however, complex traits may be dependent on multiple loci and can be affected by POEs (GUILMATRE; SHARP, 2012), such as genomic imprinting (LAWSON; CHEVERUD; WOLF, 2013). The most traditional GWAS approaches do not account for POE asymmetry on phenotypic variance (LAWSON; CHEVERUD; WOLF, 2013), missing this significant aspect of genetic inheritance. For example, with differing maternal and paternal contributions, genomic imprinting accounted for 8% to 25% of the additive genetic variance for ten beef traits of Simmental cattle (NEUGEBAUER et al., 2010). The incorporation of POEs in the predictive model explained 6% of the overall phenotypic variance of the total oocyte trait from Gir dairy cattle (ROCHA et al., 2024).

Beyond the genomic imprinting, POE can result in phenotypic variance without directly impacting gene expression. Trans-generational effects are very interesting for livestock because specific alleles from a parent can impact the offspring phenotypes via epistatic interactions or even by parent behavior and environmental effects that modulate the traits epigenetically (GUILMATRE; SHARP, 2012; ST. PIERRE et al., 2022). The females, in addition to transmitting its autosomal chromosomes, transmit mitochondrial DNA to their offspring, and the Y chromosome

is transmitted by the males to their male progeny, impacting phenotypic variation (GUILMATRE; SHARP, 2012).

To analyze POE and genomic imprinting is required to know the parental origin of each allele of the progeny, being more informative when using pure lines crossing, aiming to have more heterozygotes in the hybrid population (LAWSON; CHEVERUD; WOLF, 2013). This requirement is a limiting factor due to the high price to genotype founders and offspring in a sufficiently dense coverage, using a particular experimental design that is unfeasible for large livestock populations and collaborative projects. In light of this, studies have demonstrated that the presence of a POE causes heterozygotes to exhibit more phenotypic variation in comparison to homozygotes in large sample sizes with unknown familial trio genotypic information (HEAD et al., 2023; HOGGART et al., 2014). This idea was more recently applied in a novel model – POIROT (HEAD et al., 2023), which was founded on the hypothesis that pleiotropy across several linked characteristics from large GWAS may be exploited through the combined study of multiple traits. This approach uses a Robust Omnibus Test to compare the phenotypic covariance matrix of heterozygotes and homozygotes (HEAD et al., 2023).

POEs affecting gene expression are the result of strong allele-specific methylation (ASM) and allele-specific expression (ASE) events (TYCKO, 2010). Locating regions under the effect of both ASE and whole-genome methylation marks is an interesting research direction to identify potential new candidates to be further validated as imprinted *loci* (LAWSON; CHEVERUD; WOLF, 2013). RNA sequencing approaches can identify ASE patterns genome-wide, showing imbalanced transcription between the two alleles of a heterozygous SNP (CASTEL et al., 2015; ST. PIERRE et al., 2022). For ASE, imbalanced expression is based on the nucleotide variation, and for imprinting, this pattern is strong (typically expressed monoallelically) and occurs in response to the allele's parent-of-origin (ST. PIERRE et al., 2022). The allelic imbalance on gene expression consists of the most straightforward information useful to the discovery of new imprinted *loci* but requires an accurate assessment of the alleles' parental origin (LAWSON; CHEVERUD; WOLF, 2013). ASE can be the result of *cis*-regulatory mechanisms (CASTEL et al.,

2015; TYCKO, 2010) acting in only one heterozygous allele, as the transcription factor binding sites (TFBS) presence in the gene promoters (ABRAMOV et al., 2021; TYCKO, 2010), miRNA binding sites mainly in the 3'end (as reviewed by VOHRA et al., 2020), and epigenetic mechanisms, such as allele-specific DNA methylation (ASM) (LAWSON; CHEVERUD; WOLF, 2013; ST. PIERRE et al., 2022; VOHRA et al., 2020).

In our previous studies, we identified ASE SNPs from muscle samples of 190 Nelore steers, predicting neighboring regulatory variants impacting gene expression (aseQTLs) and the transcript abundance (*cis*-eQTLs) to increase knowledge on the regulatory machinery controlling the haplotype expression (BRUSCADIN et al., 2022a). In a subsequent paper, due to the evidence of ASE being shown to have a role in the phenotypic variation due to *cis*-regulation (LIU et al., 2020; PRZYTYCKI; SINGH, 2020; ST. PIERRE et al., 2022), the allelic imbalance was compared between animals contrasting for muscle-related traits groups, identifying differential (context-dependent) ASE SNPs (DASE SNPs) (BRUSCADIN et al., 2022b). These findings provided a basis for the potential regulatory dynamics of the allelic imbalance and pointed out regions where this biased expression can impact phenotypes. However, our previous efforts did not explore POEs or genomic imprinting aspects. Thus, herein, we investigated the POEs affecting muscle-related traits and analyzed the genome-wide distribution of ASM regions in the bovine muscle. Our hypothesis is that the parent-of-origin effects can influence the phenotypic variation of beef-related traits obtained in cattle muscle. Also, we believe that the ASM can be useful in integrative analyses to uncover the allelic regulation in Nelore muscle, pointing out new candidates for genomic imprinting in cattle.

4.3 METHODS

4.3.1 Animal production and data collection: Phenotypes, DNA and RNA sequencing

Animal production: All experimental protocols were approved by the Institutional Animal Care and Use Committee of Embrapa Pecuária Sudeste (São Carlos, São Paulo, Brazil, Protocol CEUA 01/2013). Briefly, semen samples from 34

sires that represented the main Nelore cattle lineages in Brazil were used for artificial inseminations with purebred Nelore dams to produce a population of 386 half-sib Nelore steers (CESAR et al., 2014; TIZIOTO et al., 2013). The animals were grown in consistent management and feeding circumstances (TIZIOTO et al., 2013) and were slaughtered when their backfat thickness (BFT), measured by ultrasonography, reached 5 mm. *Longissimus thoracis* muscle samples were extracted between the 11th and 13th ribs, and blood samples were collected in vacuum tubes with potassium EDTA (K3) (CESAR et al., 2014; TIZIOTO et al., 2012, 2013), stored frozen in an -80 ºC freezer until the next steps.

Phenotype measurements: Backfat thickness (BFT) and Rib eye area (REA) were measured from fresh steaks 2.5 cm thick. Using Warner-Bratzler equipment, tenderness traits were assessed as a shear force at 24 hours (WBSF0), seven days (WBSF7), and fourteen days after slaughter (WBSF14) (TIZIOTO et al., 2013). Breeding season, slaughter group, birth and feedlot sites, and animal age at slaughter were all fixed factors in the model to obtain the Genomic Estimated Breeding Values (GEBVs) for these phenotypes, as described previously (TIZIOTO et al., 2013, 2014).

A Vista Pro-CCD ICP-OES spectrometer (Varian, Australia) was used to examine 100 mg of lyophilized muscle samples to determine the mineral content (e.g., Ar, Ca, Cr, Co, Cu, Fe, Mg, Mn, P, K, Se, Na, S, and Zn). In the Animal Nutrition and Growth Laboratory at ESALQ, Piracicaba, São Paulo, Brazil, the measurements of intramuscular fat (IMF) traits were carried out with Ankom XT20 Fat Analyzer in accordance with the suggested procedures (CESAR et al., 2014). From 4 g of muscle samples the fatty acid (FA) content was extracted and profiled with gas chromatography by comparing the retention time of the resulting methyl esters with reference standards. With slaughter age as a covariate and fixed variables (birth, feedlot location, and breeding season) taken into consideration, GEBVs for the fat-related traits were computed using GenSel software (TIZIOTO et al., 2013, 2014).

DNA sequencing and imputation: The phenol-chloroform procedure was used to extract DNA from frozen semen samples from the sires. The Illumina HiSeq 2500 System (Illumina Inc., San Diego, CA, USA) at the Functional Genomic Center—ESALQ (Piracicaba, SP, Brazil) was used to perform whole-genome sequencing of 26 sires with a coverage of 8–21X. Data management adhered to the previously outlined procedures of the 1000 Bulls Genomes Project (accessible on July 1, 2020) (CARDOSO et al., 2021). The salting-out technique was used to recover DNA from blood samples for the offspring (TIZIOTO et al., 2012). Agarose gel electrophoresis was used to verify integrity, NanoDrop® was used to determine DNA content, and the 260/280 optical absorbance ratio was used to assess purity. As previously mentioned, the Illumina 700 K BovineHD BeadChip (Illumina Inc., San Diego, CA, USA) was used to genotype sires and steers (MUDADU et al., 2016).

Our genotypes comprise an imputed dataset that integrated whole-genome sequencing results from 26 sires and the high-density genotyping data from the progeny, as previously reported (CARDOSO et al., 2021). Autosomal SNPs with a minor allele frequency higher than 0.01 and an imputation error rate of less than 2% were kept after quality control, which was carried out using PLINK software (PURCELL et al., 2007), resulting in 4.8 million SNPs (CARDOSO et al., 2021).

RNA sequencing: TRIzol® (Life Technologies, Carlsbad, CA, USA) was used to extract total RNA from 100 mg of muscle tissue from 190 steer samples as per the manufacturer's instructions. RNA sequencing was carried out at the Functional Genomic Center—ESALQ (Piracicaba, SP, Brazil), using an Illumina HiSeq 2500® system to sequence the RNA libraries after they were produced using the TruSeq RNA Sample Preparation Kit (Illumina, San Diego, CA, USA) (CESAR et al., 2018). SeqyClean v1.4.13 was used for quality control, eliminating low-complexity reads and sequencing adapters (ZHBANNIKOV et al., 2017). STAR software (DOBIN et al., 2013) was used to align the sequences to the bovine reference genome (ARS-UCD1.2, Ensembl 100) with default settings. The WASP pipeline (VAN DE GEIJN et al., 2015) was used to mitigate inflated reference allele counts and resolve potential mapping biases. GATK ASE Read Counter (MCKENNA

et al., 2010) was used to obtain allele-specific expression counts from heterozygous variations (Bruscadin et al., 2022a)

4.3.2 Preliminary GWAS

For the parent-of-origin GWAS (PO-GWAS), we selected 185 samples for which data from all phenotypes were available. We applied the POIROT methodology (HEAD et al., 2023), whose assumptions that using correlated traits can improve the identification of parent-of-origin variants. Thus, we aimed to cluster similar phenotypes using Pearson's correlation matrix from our previous study (BRUSCADIN et al., 2022b) from the Genetic Estimated Breeding Values (GEBVs) of all 55 available phenotypes for a hierarchical clustering analysis. We used the Manhattan distance method to evaluate the trait's similarity and the K-means method to delimitate the phenotype clusters.

Based on the trait clustering results and our biology knowledge, we selected ten "tag phenotypes" to run a traditional GWAS analysis from all our available and quality-controlled SNPs to select SNPs with evidence of effect on our phenotypes. For each tag phenotype, we used GEMMA (ZHOU; STEPHENS, 2012) to apply a linear mixed model as described previously (CARDOSO et al., 2021) to associate the SNPs genome-wide to the GEBVs of the selected traits. We considered SNPs, resulting in a nominal P-value ≤ 0.001 for the subsequent analyses.

4.3.3 PO-GWAS

We constructed a panel comprising all selected SNPs from the preliminary GWAS analyses, and another SNPs obtained from our previously studies: ASE SNPs, DASE SNPs, and their neighboring *cis*-eQTLs and aseQTLs, which are SNPs associated with the transcript abundance and associated with the allele-specific expression, respectively (BRUSCADIN et al., 2022a, 2022b). Thus, we tested all selected SNPs in each PO-GWAS run to investigate if they showed parent-of-origin effects in each phenotype set using POIROT (HEAD et al., 2023), which models the phenotypic covariance matrix between heterozygotes and homozygotes using a Robust Omnibus Test. The resulting P-values were corrected using Benjamini-

Hochbergs's False Discovery Rate (FDR) method, considering significant SNPs with a FDR ≤ 0.05.

4.3.4 SNP Annotation and Gene Enrichment Analysis

The annotation of PO-QTLs and of functional effects prediction were performed using SNPEff software (CINGOLANI et al., 2012), based on the bovine reference genome assembly ARS_UCD1.2. Genes affected by PO-QTLs from each PO-GWAS analysis were used for an enrichment analysis implemented by the EnrichR package (KULESHOV et al., 2016), with data from Gene Ontology Biological Process (v2023) and the significance threshold of FDR ≤ 0.05.

4.3.5 Allele-specific methylation analysis

From 12 samples from the same experimental population, Reduced Representation Bisulfite Sequence (RRBS) data was obtained in muscle, as described elsewhere (DE SOUZA et al., 2022). The raw files were downloaded from NCBI's Gene Expression Omnibus (GEO) accession number GSE190966. The fastq files were quality-controlled using TrimGalore! (https://github.com/FelixKrueger/TrimGalore) with the –ribs flag, and the quality was checked with fastqc. The resulting reads were mapped against the reference genome ARS_UCD1.2 using Bismark software (KRUEGER; ANDREWS, 2011), which was also used to conduct the deduplication and methylation extraction by applying the default parameters.

Allele-specific methylation events were obtained by implementing the DNMTools software (SONG et al., 2013), running the amrfinder function with the parameters -g 100 -w 5 -m 3, which means that the maximum distance between CpGs is 100 bp. The sliding window (w) to the test is at least five CpGs, and -m 3 indicates the minimal coverage of reads matching a CpG to be tested. The software returns the FDR-adjusted P-values, and we considered significant all regions that showed a FDR ≤ 0.01.

The ASM regions were annotated using Homer (v5.1) (HEINZ et al., 2010), the reference genome, and annotation from the ARS-UCD1.2 assembly. Vcftools

(v0.1.16) (DANECEK et al., 2011) was used to filter the genotype file for the ASM regions, selecting SNPs overlapping regions under the control of allelic methylation.

4.3.6 Integration of variants and genes with previous studies and public data

We compared our findings with several datasets to find bibliographic consistency and to draft some hypotheses about the relevance of our findings in livestock. Genes annotated for the PO-QTLs and ASM regions were compared to the list of ASE genes and DASE genes from our previous studies (BRUSCADIN et al., 2022a, 2022b) and with imprinted genes from the GeneImprint database (http://www.geneimprint.com/), with any imprinting evidence in cow and human species. Variants within ASM regions and the PO-QTLs were overlapped with ASE SNPs, DASE SNPs, aseQTLs and *cis*-eQTLs obtained in the same experimental population (BRUSCADIN et al., 2022a, 2022b).

4.3.7 Prediction of regulatory mechanisms

Different prediction analyses were run to understand the potential regulatory machinery affecting gene expression in bovine muscle. FIMO was used to identify transcription factor motifs in regions showing ASM and PO-QTLs upstream to the transcription start site of the closest genes, using the set of transcription factors binding sites (TFBSs) from vertebrates, downloaded from Jaspar database (v2024) (RAULUSEVICIUTE et al., 2024). We considered as significant the results showing FDR-corrected P-value ≤ 0.05. Enrichment analysis was performed with the TFs with motifs predicted for PO-QTLs from each phenotypic set with enrichR (KULESHOV et al., 2016) using the same method described in Section 4.2.3.

To search for miRNA binding sites in the genomic regions of our findings, ASM/PO elements located downstream the 3'UTR regions were tested as miRNA motifs with RNAHybrid (KRUGER; REHMSMEIER, 2006), with the parameters: -s 3utr_human -c -e -18 -p 0.05 -f 2,7 and the bovine miRNAs downloaded from the mirGene database (FROMM et al., 2022). For TF and miRNA motifs discovery, the input target regions were genomic sequences extracted from the bovine reference genome (ARS_UCD1.2) according to coordinates of the ASM regions and 15 bp-flanking regions around PO-QTLs.

4.4 RESULTS AND DISCUSSION

4.4.1 Parent-of-origin effects were found for clusters of correlated traits

The phenotypes clustered in well-defined groups, as shown in Figure 4.1. Herein, we will discuss the results based on the color coding of the clusters. The black cluster consists of mineral phenotypes (Ca, S, Zn, Na, Mg, P, and K); the red cluster consists predominantly of saturated FAs as C10:0, C12:0, C14:0, C16:0, C17:0, C18:0, SFA (sum of saturated fatty acids); but also contains the monounsaturated FAs C18:1-trans16 and C18:1-trans10,11,12, and the polyunsaturated FA C18:3-n6; polyunsaturated FAs are concentrated in the green cluster (C20:2, C20:4, C20:5, C22:5, C22:6, C18:2-cis9,12, sum of polyunsaturated FAs – PUFA, C20:3 n6, n3 and n6); the blue cluster is represented mostly by monounsaturated FAs (C18:1, C18:1 cis-9, C18:1 cis-11, C18:1 cis-12, C18:1 cis-13, C18:2 trans-11 cis-15, C17:1, C20:1), but also included BFT and REA; the turquoise cluster contains shear force phenotypes (WBSF0, WBSF7, WBSF14) and arsenic concentration trait; the pink cluster consist of minerals (Cu, Se, Cr, Co, Mn, and Fe), monounsaturated FAs (C14:1 cis-9, C16:1 cis-9, C18:1 trans-6,7,8,9), one polyunsaturated FA (C18:2 cis-9 trans-11) and the IMF trait.

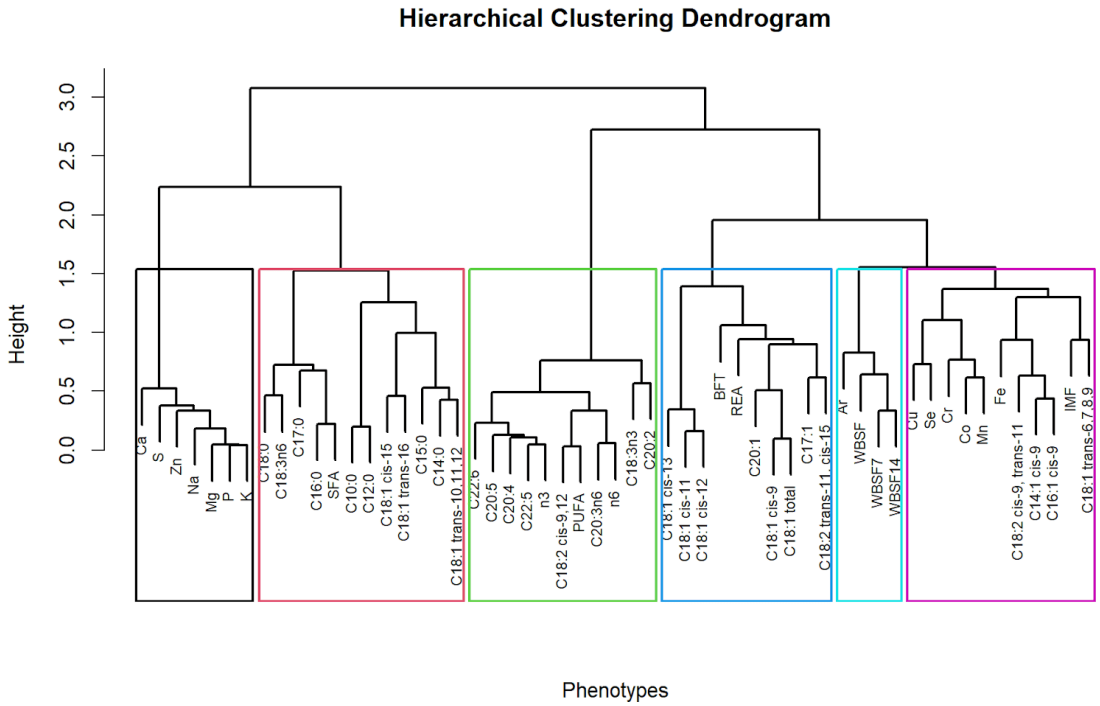


Figure 4.1 – Hierarchical Clustering Dendrogram based on GEBVs.

For the GWAS preliminary analysis, we chose ten traits (Table 4.1) based on the branches of the dendrogram of Figure 4.1. We also considered some biological knowledge about the more interesting traits of livestock to get a representative subset of variants that affect all these traits. The selected traits were sodium concentration, the sum of saturated FAs, the sum of polyunsaturated FAs (PUFAs), the sum of omega3 (n3), BFT, REA, WBSF0, WBSF7, and WBSF14, and IMF. We tested all shear force phenotype measurements because of the need to improve beef tenderness for Nelore cattle. The number of variants selected from the GWAS procedure for the PO-GWAS analyses is described in Table 4.1, as well as their minimum, average, and maximum effect (beta). All results are shown in Supplementary Table 1.

Table 4.1 – Description of variants selected from the preliminary GWAS analysis.

Phenotypes	SNPs with P-value <= 0.001	Minimum beta	Mean beta	Maximum beta
Sodium concentration	5354	-0.0259	0.0193	0.0698
Sum n3	5439	-0.0377	0.0211	0.0918
Sum Saturated FAs	6544	-0.7043	0.2229	0.8603
Sum Polyunsaturated FAs	6535	-0.1221	0.0832	0.3036
WBSF14	5414	-0.2263	0.0502	0.4834
WBSF7	3495	-0.5060	0.0089	0.5435
WBSF0	3855	-0.3993	-0.0580	0.3207
IMF	3619	-0.3735	0.1538	0.8776
BFT	5681	-0.6232	0.3257	1.1530
REA	4295	-2.0588	0.1790	3.3080

As the parent-of-origin effects depend on the parental allele, affecting phenotype variation or gene expression, we also included our variants from the allele-specific expression studies to test for the PO-GWAS model. Thus, our initial variant set consisted of 105,573 SNPs, aggregating variants from the preliminary GWAS analyses, ASE SNPs, DASE SNPs, aseQTLs, and *cis*-eQTLs (BRUSCADIN et al., 2022a, 2022b)

The number of significant variants identified in the PO-GWAS analysis and a brief description of the effect of these variants regarding the POIROT method statistic are described in Table 4.2 and represented in Figure 4.2. Supplementary Table 2 shows all variants identified in these analyses. The trait set that resulted in the most significant number of parent-of-origin quantitative trait loci (PO-QTLs) was the green cluster (PUFAs), with 1,188 significant SNPs (FDR ≤ 0.05) and the most prominent effect sizes. For the pink cluster, no variant passed our filtering criteria. We checked for overlapping variants identified in more than one trait cluster, but no

one overlapped at FDR ≤ 0.05. This observation indicates that the identified PO-QTLs are sensitive to the phenotypes selected in each group, meaning that this methodology can identify parent-of-origin pleiotropic effects from multiple phenotypes, but the results are still dependent on the phenotype set. We highlighted the variants that resulted in the larger effects for each group in Table 4.2.

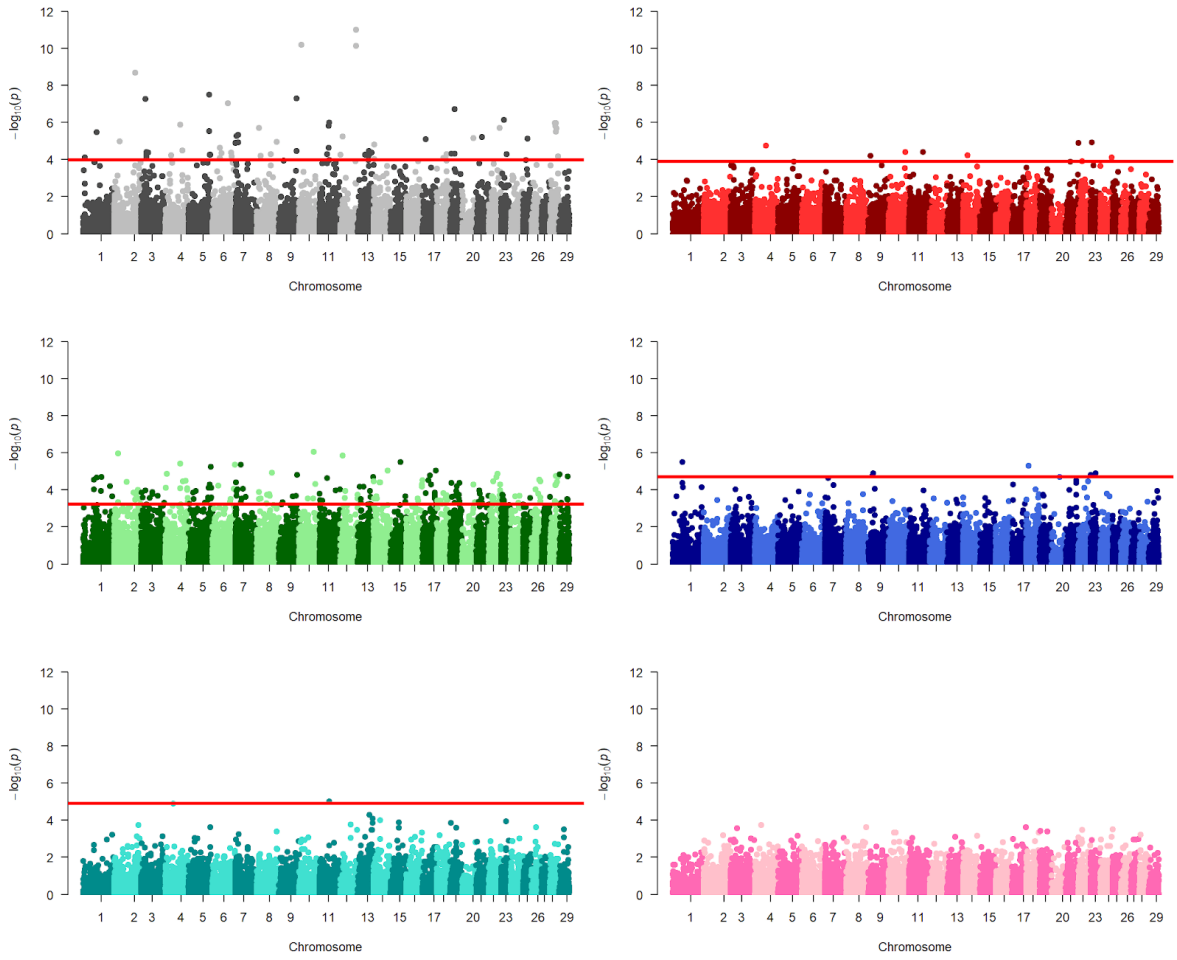


Figure 4.2 - Manhattan Plots of PO-GWAS analyses according to each phenotype set A) Black cluster (minerals); B) Red cluster (SFAs); C) Green cluster (PUFAs); D) Blue cluster (MUFAs, BFT, REA); E) Turquoise cluster (WBSF and Ar); F) Pink cluster (mixed phenotypes). The red line represents the FDR threshold of 0.05.

Table 4.2 – PO-GWAS general statistics and top PO-QTLs information per trait cluster

Cluster	Number of PO-QTLs	Min. Effect	Mean Effect	Max. Effect	Top variants	Top variant stats	Top variants FDR
Black	215	3,480	4,483	16,628	chr25:38734790	16,628	1,88E-10
					chr12:78898059	10,413	4,97E-07
					chr10:12161684, chr10:12161859, chr10:12162267, chr10:12162271	9,354	1,00E-06
Red	347	32,886	40,966	74,884	chr23:8345433-8350980 (9 SNPs)	74,884	0,021
					chr21:68116251	72,613	0,021
					chr4:59979528	64,420	0,021
Green	1188	19,395	44,712	177,733	chr10:73668411	117,733	0,004
					chr2:21003969	163,941	0,004
					chr12:11299888	151,186	0,004
Blue	40	4,897	5,089	5,862	chr1:50513220, chr1:50513307, chr1:50518601	5,862	0,044
					chr18:12939042, chr18:12950618	5,633	0,044
					chr9:26116961	5,118	0,044
Turquoise	27	5,327	5,339	5,432	chr11:48653342, chr11:48663175, chr11:48664683	5,432	0,045
					chr4:43164905-43185686 (22 SNPs)	5,327	0,045

Pink	0	.	.	.	.	.	.

We carried out the variant annotation of PO-QTLs in SNPEff (Figure 4.3A) (CINGOLANI et al., 2012). Most variants are in introns, followed by intergenic and downstream regions. Figure 4.3B shows the number of PO-QTLs identified by the trait clusters and the proportion of their annotated effects. Regarding the annotations for each trait cluster, PO-QTLs were primarily located in intronic regions, except for the black (minerals) and turquoise clusters (WBSF and Ar), which mostly have variants in intergenic regions. The blue cluster (MUFAs, BFT and REA) had the lowest proportion of intergenic variants. Downstream variants were frequently found in the PO-QTLs belonging to the green (PUFAs) and red (SFAs) clusters, and upstream variants were only found in the black, red, and green groups.

4.4.2 Genes affected by POEs: Enrichment analysis showed biological processes essential for the associated traits

We performed enrichment analyses of biological processes (BPs) with genes containing PO-QTLs identified from each trait cluster (Supplementary Table 3). No BPs were enriched at FDR ≤ 0.05 for the black (minerals) and turquoise (WBSF and Ar) groups.

Genes containing PO-QTLs from the red cluster (mostly SFAs) were enriched for 74 BPs, most related to muscular contraction, ion transport, and cell size. The most enriched BPs make sense for the PO-GWAS phenotypic set: FAs are known to regulate íon channels (ORDWAY; SINGER; WALSH, 1991); SFAs were suggested to be related to muscle loss in patients with rheumatoid arthritis (SEBE et al., 2022) and vascular contraction in rats (VORN; YOO, 2019); adipose cells size and number were correlated with a dietary increase of SFAs (GARAULET et al., 2006).

PO-QTLs from the green cluster (PUFAs), have genes enriched for Actomyosin Structure Organization (GO:0031032) and Striated Muscle Cell Development (GO:0055002) processes. Omega-3 PUFAs were previously related to muscle growth and repair (TACHTSIS; CAMERA; LACHAM-KAPLAN, 2018), being its supplementation associated with increased muscle strength (RODACKI et

al., 2012). The genes *TMOD1* (Tropomodulin 1), *LMOD1* (Leiomodin 1), *MYOZ1* (Myozenin 1) and *MYL9* (Myosin Light Chain 9) were enriched in both BPs.

For the blue cluster (MUFAs, BFT and REA), PO-QTLs are in genes enriched for 52 BPs, such as protein ubiquitination, epidermal growth factor (EGF) signaling, protein and mRNA transport processes (Supplementary Table 3). Most FAs of the blue clusters are derived from oleic acid, with an enhanced ubiquitination of a membrane scavenger receptor called FAT/CD36 that helps muscles absorb long-chain fatty acids (SMITH et al., 2008). The ubiquitin-proteasome system is a channel for intracellular protein degradation that significantly impacts muscle atrophy (PANG et al., 2023), which in turn can impact muscle growth as measured by the REA phenotype. Oleic acid also activates the epithelial growth factor signaling pathway, suggested as a target for this FA (VACARESSE et al., 1999).

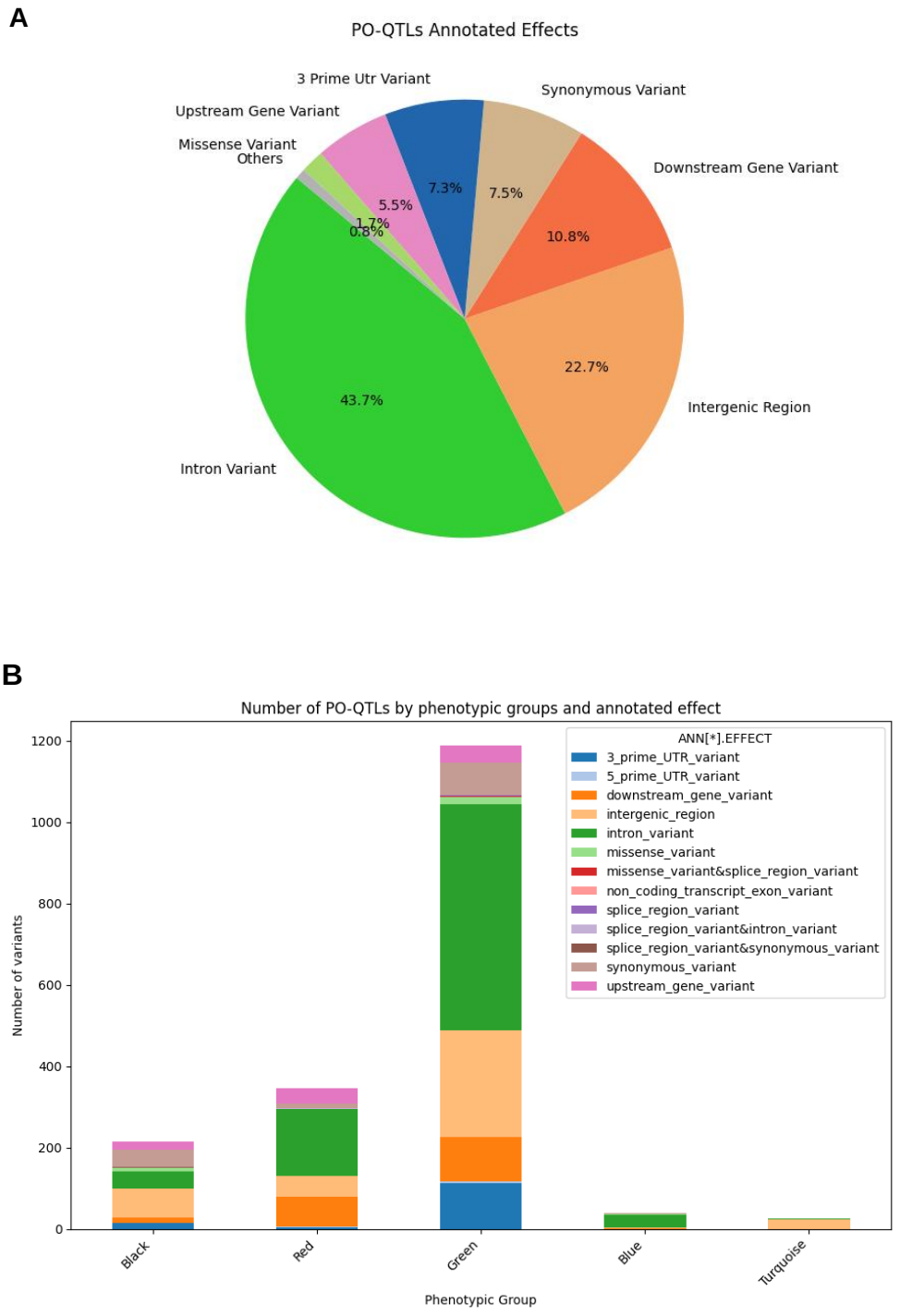


Figure 4.3 - PO-QTLs annotated effects. A) General annotation of all significant PO-QTLs. B) Annotation of PO-QTLs per trait cluster.

4.4.3 Consistent allele-specific methylation was observed in muscle of Nelore animals

Methylation processing metrics are available in Supplementary Table 4. We identified 8,074 ASM regions across all animals, ranging from 2,211 to 3,585 per sample (Supplementary Table 5). Figure 4.4A (black marks) shows the chromosomic distribution of ASM regions. We can see that these regions are well distributed across chromosomes, often with increased density in the chromosome extremities. Subtelomeric regions are typically rich in CpG islands, suggesting that the DNA methylation in these regions may have a role in the telomere length (SCHOEFTNER; BLASCO, 2009).

The chromosome with the most significant number of ASM regions was chromosome 19 (476 ASM regions), being the second chromosome with more density of ASM regions, with 7,38291E-06 ASM regions/base pair. Chromosome 25 was the one with more markers per base pair, with 8,00197E-06 ASM regions/bp and 354 ASM regions. In Figure 4.4A, we also plotted ASE SNPs obtained previously (BRUSCADIN et al., 2022a) in yellow-colored intermediate-size bars, and the identified PO-QTLs were plotted in the smallest bars, colored in blue. There is no tendency for PO-QTLs to occur in regions containing ASE SNPs or ASM, but we can visualize some regions with overlapping variant sets (chromosomes 5, 18, 21, 23, 28). It is possible to hypothesize that where ASE SNPs and ASM regions are located nearby, there is evidence for the potential of ASM to control allelic expression at the exact locations in the genome.

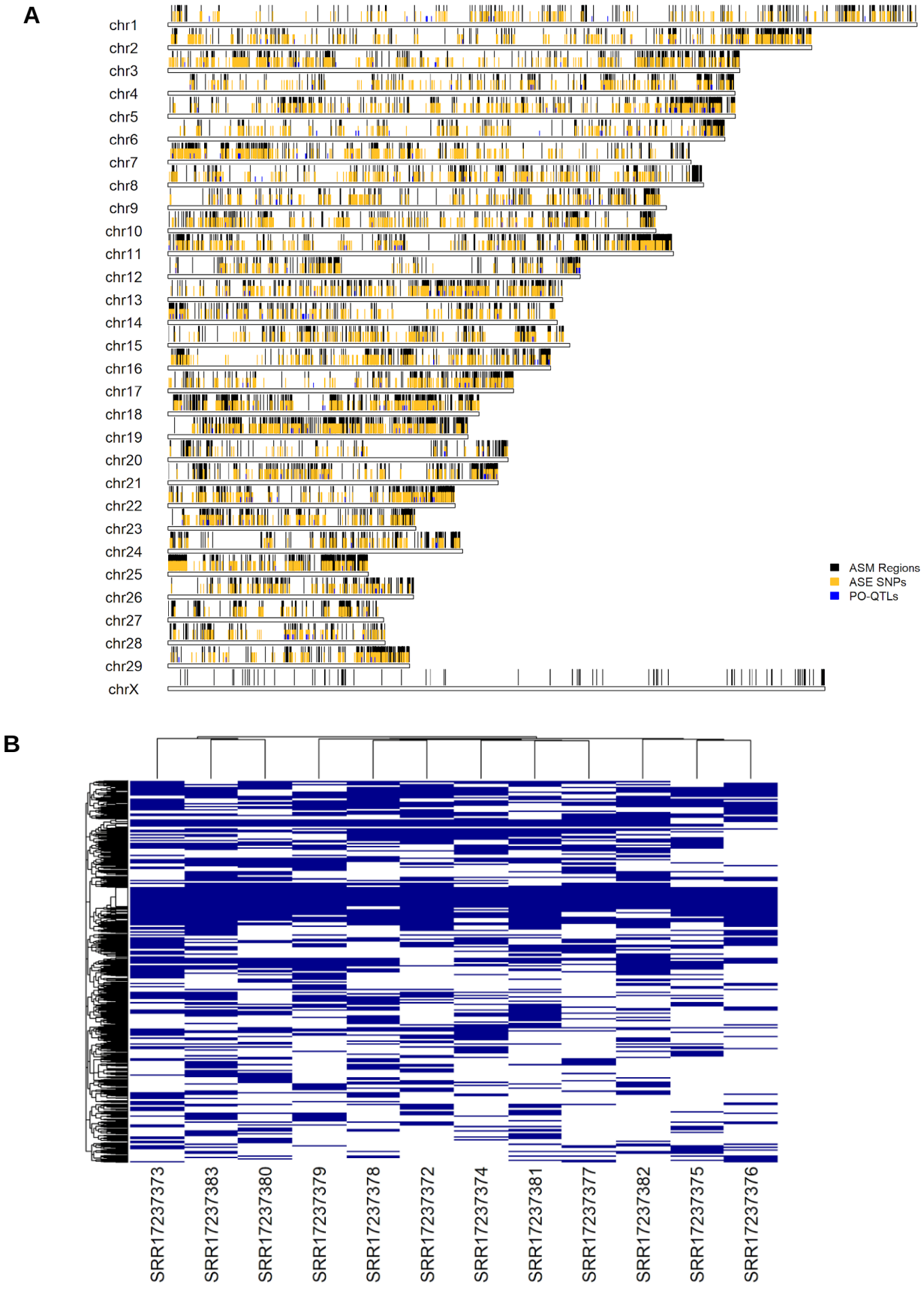

A
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chr23
chr24
chr25
chr26
chr27
chr28
chr29
chrX
ASM Regions
ASE SNPs
PO-QTLs
B
SRR17237373
SRR17237383
SRR17237380
SRR17237379
SRR17237378
SRR17237372
SRR17237374
SRR17237381
SRR17237377
SRR17237382
SRR17237375
SRR17237376

Figure 4.4 – ASM Results. A) Allele-specific methylated (ASM) regions across chromosomes. Black marks represent the genomic location of ASM regions; yellow marks represent the distribution of allele-specific expression SNPs (from Bruscadin et al., 2022a) and blue marks represent parent-of-origin QTLs for any trait cluster. B) Representation of allele-specific methylated (ASM) regions across samples. Blue color represents the presence of a given ASM region in a sample and the white color represents the absence of this ASM in a sample. Sample IDs are in the x-axis and ASM regions are in the y-axis. Dendrograms clustered samples and ASM regions based on their similarity.

When grouping the overlapping ASM regions across all samples, we identified 7,812 unique (non-overlapping) clusters, with 269 ASM regions present in all samples. Figure 4B shows the overlapping ASM regions across samples, where we can observe that the ASM can be sample-specific (regions with a few blue color cells across all samples) but also can be conserved across samples (blue color in all samples).

Annotation identified that ASM is mainly distributed in promoter-TSS regions (2,821), followed by 2,326 ASM regions located in intronic regions, 1,399 in exonic regions, 1,264 in intergenic regions, and 259 in transcription termination sites. These regions were predicted based on 5,664 genes (Supplementary Table 5). ASM regions common in all samples are primarily located in promoter-TSS regions (129 from 269 common ASM regions).

4.4.4 Genes potentially affected by ASM and PO-QTLs can impact allelic expression and genomic imprinting

ASM regions overlapped with 1,938 ASE genes and 171 DASE genes. In general, we identified 139 genes from ASM regions overlapping with any imprinting status except for “not imprinted,” 95 identified in humans, 60 in bovines, and 16 in both species. Regarding the imprinting status, 85 overlapped genes were classified as “imprinted,” 58 in bovines and 41 in humans. Four overlapping genes were found in the ASM regions identified in all samples: *CCDC71L* (ASM in an exon) imprinted

in humans, *IGF2R* (intron) classified as “Tissue-dependent” for bovines, *LMX1B* (intron) and *SALL1* (promoter) predicted in human. Our results show the importance of studying the methylation level and allelic expression to understand the genomic imprinting and allelic regulation, highlighting new candidates for imprinting in bovine. Integration with imprinting genes is available in Supplementary Table 6.

PO-QTL annotated genes overlapped with 270 ASE genes and two imprinted genes: *C1S* (imprinted in bovine) and *RASGRF1* (imprinted in humans). This result indicates that ASM has a greater impact on genomic imprinting and can be more informative when integrated with ASE than PO-QTLs, which potentially show less direct association with gene expression.

4.4.5 Transcription factor and miRNA motifs are potentially modified by PO-QTLs and ASM

In the TF motifs screening, we identified 122 TFs affecting 2,770 ASM regions (Supplementary Table 7). We identified 111 enriched BPs, mostly related to regulation of transcription by RNA polymerase II, regulation of DNA-templated Transcription, and similar processes expected for TFs; moreover, some specific processes were identified as regulation of cell differentiation and proliferation, nervous system development-related processes, regulation of miRNA metabolic process and transcription, and regulation of apoptosis, for example. The ASM region that was more often predicted to have TF motifs was chr23:42944983-42945460 (430 occurrences corresponding to 62 TFs), followed by chr17:72068021-72068256 and chr21:22397478-22398043. The most common TFs were *PATZ1* (6,179 motifs), *ZBED4* (5027 motifs) and *ZNF93* (4,857 motifs). We could observe that chromosome 13 has multiple motifs of *ZNF93* and *PATZ1* TFs, affecting *MMP24*, *RIMS4*, and *ISM1* genes.

Moreover, 45 TFBSs were predicted for 36 sequences extracted from PO-QTLs. Sequences around PO-QTLs from chromosome 24 were the most common regions affected by TFBSs. The region between positions 56816986 and 56825913 of chromosome 24 has predicted motifs for *ZNF281, NKX2-2, PRDM9, Wt1, PGR, KLF4, ZIC1, ZIC4, Tcf21, ZNF684, RREB1*, and *ZFP14*. The PO-QTL within this region was identified for the red cluster of traits (SFAs) and was predicted in the

upstream region of *FECH* (ferrochelatase) gene. FA supplementation can impact the activity of hepatic ferrochelatase (KOOLS et al., 1989). Another PO-QTL on chromosome 24 is located in an upstream region of the U6 (U6 spliceosomal RNA) gene that had a predicted motif ZNF547.

We also test the putative presence of miRNA binding sites in the ASM and PO-QTL regions (Supplementary Table 8). We identified motifs for 75 miRNAs in 121 unique ASM regions, with Bta-Mir-760_3p the most common miRNA with 18 occurrences, followed by Bta-Mir-1343_3p (17) and Bta-Mir-296_3p (12). The sequences that were predicted to have more miRNA binding sites were chr15:46454047-46454168, in the transcription termination site (TTS) of the *TAF10* gene, and chr5:27142582-27142647 (TTS of the *KRT3* gene).

The PO-QTLs regions showed predicted binding sites for 91 miRNAs for 83 unique PO-QTLs. The most common miRNA was Bta-Mir-423_5p, identified for five PO-QTLs, and Bta-Mir-154-P2-v2_5p and Bta-Mir-744_5p, identified for four O-QTLs. The PO-QTL within the region chr16:80577514 (at the downstream region of *the LMOD1* gene) has 12 predicted miRNA binding sites.

4.4.6 Data integration revealed regions potentially linked to phenotypic variance and genomic imprinting

We joined all our findings in a single table (Supplementary Table 9) to help highlight the most relevant SNPs or regions according to their regulatory potential.

We identified multiple PO-QTLs with regulatory evidence, such as chr11:69714613, identified for the red cluster (SFAs). This PO-QTL is a synonymous variant of *YPEL5* gene and was previously identified as an ASE SNP, *cis*-eQTL and showed differential ASE for the C18:0 fatty acid trait. chr22:37470306 is a PO-QTL for the green cluster (PUFAs), located in the *ATXN7* gene that was previously identified as an ASE SNP, *cis*-eQTL, and predicted to impact a motif for the Bta-Mir-122_5p miRNA. We found three variants upstream to the *CLPX* gene that were identified as PO-QTLs of the black cluster (minerals) that impact TFs: chr10:12161859 (*ZNF213*); chr10:12162267 and chr10:12162271 (Irf1). Some PO-QTLs also overlapped with QTLs of similar phenotypes, as six PO-QTLs (Red –SFAs

group) within an intergenic region of the *RASGRF1* gene between chr21:27918693 - chr21:27920870 also identified as QTLs for sum of saturated fatty acids concentration. *RASGRF1* is an imprinted gene in humans. Four PO-QTLs within the downstream region of *VIRMA* gene (chr14:69846651 - chr14:69849759) are located in binding sites of three miRNAs: Bta-Mir-103-P2_3p, Bta-Mir-103-P1_3p, Bta-Mir-103-P4_3p; these PO-QTLs were identified for the green (PUFAs) cluster and overlapped QTLs for sum of saturated FAs. chr28:30677287 is a PO-QTL of the green (PUFAs) group, an upstream variant of the gene (miRNA) bta-miR-584-3, potentially located in a motif for the *PRDM9* TF and overlapping a QTL for the sum of saturated FAs.

Several ASM regions showed interesting evidence. Some well-known imprinted genes in human and bovine genomes were identified with ASM regions in the promoter, such as *GNAS, PEG3, DIO3, CDKN1C,* and *MEST*. A total of 56 ASM regions were predicted to affect motifs for TFs and impact imprinted genes, such as the ASM chr21:65468776-65468934, which was predicted to affect multiple TFs binding sites and is located in the promoter-TSS region of the *BEGAIN* gene, imprinted in bovine; however, this gene is very low expressed in our muscle samples, probably due to the tissue-specific characteristic of the genomic imprinting (BABAK et al., 2015; BARAN et al., 2015). *PMM2* is also imprinted in bovine and has an ASM (chr25:7652863-7652934) in its promoter region also predicted to affect TFs motifs. In addition, 57 ASM regions were annotated for genes imprinted in the bovine genome, and 39 ASM regions were annotated for genes imprinted only in humans. chr19:37949212-37949275 ASM region was the only one to impact miRNA binding sites (Bta-Mir-430-P1_3p, Bta-Mir-430-P4_3p, and Bta-Mir-430-P3_3p), being located in the TTS of the *HOXB2* gene, predicted as imprinted in human.

Interesting results were found for ASM regions present in all samples, helping to draft some new candidates for imprinting in bovine for further validation. 9_96222062-96223057 is an ASM region present in all samples, being located in an intronic region of the *IGF2R* gene, that has tissue-dependent imprinting in bovine genomes and conflicting data for human; *IGF2R* showed an ASE SNP (chr9_96287564) with monoallelic expression favoring the reference allele in all

samples that have the mutation and this gene expressed (69/101 samples) (BRUSCADIN et al., 2022a) (Supplementary table 10).

18_19561422-19561497 is in the promoter-TSS region of the *SALL1* gene, showed multiple TFBSs, and the gene was predicted as imprinted in humans.17:72068021-72068256, 21:22397478-22398043 and 21:22397478-22398043, also have predicted TFBS. Of the 269 ASMs common in all samples, 111 are annotated in genes with ASE SNPs showing monoallelic expression in more than 80% of the samples from our previous study (BRUSCADIN et al., 2022a), 53 in the promoter-TSS region, 46 in intronic regions, 27 in exons and 1 in a TTS.

4.5 Conclusion

By identifying unique phenotypic clusters with clearly defined biological characteristics, our work provides information about the genetic background of these Nelore cattle phenotypes. The green cluster (PUFAs) had the most significant results and the largest impact sizes of PO-QTLs across the trait clusters. The lack of overlapping PO-QTLs between clusters is a relevant observation since it shows how sensitive these loci are to the particular traits under study, indicating the importance of this employed methodology in detecting parent-of-origin pleiotropic effects.

Our results provide additional insights by combining ASE and ASM. With the overlap of multiple known or putative imprinted genes in the human and bovine genomes, jointly with the discovered ASM regions primarily found in promoter-TSS regions, further highlighted their possible function in modulating gene expression and genomic imprinting.

Our annotation and enrichment analyses identified biological processes related to ion transport, fatty acid metabolism, and muscle growth. The complexity of the regulatory mechanisms influencing ASM and PO-QTL loci is highlighted by the discovery of TFBSs and miRNA motifs in these regions.

Collectively, these findings contribute to our knowledge of the genetic and epigenetic processes that underlie economically significant bovine traits. To improve breeding programs aimed at achieving desired phenotypes, future research should concentrate on confirming these findings from *in vitro* experiments, and investigating the functional consequences of identified regulatory regions.

References

ABRAMOV, S. et al. Landscape of allele-specific transcription factor binding in the human genome. **Nature Communications**, v. 12, n. 1, p. 2751, 12 maio 2021.

BABAK, T. et al. Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. **Nature Genetics**, v. 47, n. 5, p. 544–549, maio 2015.

BARAN, Y. et al. The landscape of genomic imprinting across diverse adult human tissues. **Genome Research**, v. 25, n. 7, p. 927–936, jul. 2015.

BRUSCADIN, J. J. et al. Allele-specific expression reveals functional SNPs affecting muscle-related genes in bovine. **Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms**, v. 1865, n. 8, p. 194886, nov. 2022a.

BRUSCADIN, J. J. et al. Differential Allele-Specific Expression Revealed Functional Variants and Candidate Genes Related to Meat Quality Traits in B. indicus Muscle. **Genes**, v. 13, n. 12, p. 2336, 11 dez. 2022b.

CARDOSO, T. F. et al. Multi-Omics Approach Reveals miR-SNPs Affecting Muscle Fatty Acids Profile in Nelore Cattle. **Genes**, v. 12, n. 1, p. 67, 6 jan. 2021.

CASTEL, S. E. et al. Tools and best practices for data processing in allelic expression analysis. **Genome Biology**, v. 16, n. 1, p. 195, dez. 2015.

CESAR, A. S. et al. Genome-wide association study for intramuscular fat deposition and composition in Nellore cattle. **BMC Genetics**, v. 15, n. 1, p. 39, dez. 2014.

CESAR, A. S. M. et al. Identification of putative regulatory regions and transcription factors associated with intramuscular fat content traits. **BMC Genomics**, v. 19, n. 1, p. 499, dez. 2018.

CINGOLANI, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w^{1118} ; iso-2; iso-3. **Fly**, v. 6, n. 2, p. 80–92, abr. 2012.

DANECEK, P. et al. The variant call format and VCFtools. **Bioinformatics**, v. 27, n. 15, p. 2156–2158, 1 ago. 2011.

DOBIN, A. et al. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, v. 29, n. 1, p. 15–21, 1 jan. 2013.

FROMM, B. et al. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. **Nucleic Acids Research**, v. 50, n. D1, p. D204–D210, 7 jan. 2022.

GARAULET, M. et al. Relationship between fat cell size and number and fatty acid composition in adipose tissue from different fat depots in overweight/obese humans. **International Journal of Obesity**, v. 30, n. 6, p. 899–905, jun. 2006.

GUILMATRE, A.; SHARP, A. Parent of origin effects. **Clinical Genetics**, v. 81, n. 3, p. 201–209, mar. 2012.

HEAD, S. T. et al. POIROT: a powerful test for parent-of-origin effects in unrelated samples leveraging multiple phenotypes. **Bioinformatics**, v. 39, n. 4, p. btad199, 3 abr. 2023.

HEINZ, S. et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. **Molecular Cell**, v. 38, n. 4, p. 576–589, maio 2010.

HOGGART, C. J. et al. Novel Approach Identifies SNPs in SLC2A10 and KCNK9 with Evidence for Parent-of-Origin Effect on Body Mass Index. **PLoS Genetics**, v. 10, n. 7, p. e1004508, 31 jul. 2014.

JIRTLE, R. L.; WEIDMAN, J. R. Imprinted and More Equal. **American Scientist**, v. 95, p. 143–149, 2007.

KOOLS, A. M. et al. Modulation of hepatic ferrochelatase activity by dietary manipulation of mitochondrial phospholipid fatty acyl groups†. **Hepatology**, v. 9, n. 4, p. 557–561, abr. 1989.

KRUEGER, F.; ANDREWS, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. **Bioinformatics**, v. 27, n. 11, p. 1571–1572, 1 jun. 2011.

KRUGER, J.; REHMSMEIER, M. RNAhybrid: microRNA target prediction easy, fast and flexible. **Nucleic Acids Research**, v. 34, n. Web Server, p. W451–W454, 1 jul. 2006.

KULESHOV, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. **Nucleic Acids Research**, v. 44, n. W1, p. W90–W97, 8 jul. 2016.

LAWSON, H. A.; CHEVERUD, J. M.; WOLF, J. B. Genomic imprinting and parent-of-origin effects on complex traits. **Nature Reviews Genetics**, v. 14, n. 9, p. 609–617, set. 2013.

LIU, Y. et al. Genome-wide analysis of expression QTL (eQTL) and allele-specific expression (ASE) in pig muscle identifies candidate genes for meat quality traits. **Genetics Selection Evolution**, v. 52, n. 1, p. 59, dez. 2020.

MCKENNA, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. **Genome Research**, v. 20, n. 9, p. 1297–1303, set. 2010.

MUDADU, M. A. et al. Genomic structure and marker-derived gene networks for growth and meat quality traits of Brazilian Nelore beef cattle. **BMC Genomics**, v. 17, n. 1, p. 235, dez. 2016.

NEUGEBAUER, N. et al. Evidence for parent-of-origin effects on genetic variability of beef traits1. **Journal of Animal Science**, v. 88, n. 2, p. 523–532, 1 fev. 2010.

ORDWAY, R. W.; SINGER, J. J.; WALSH, J. V. Direct regulation of ion channels by fatty acids. **Trends in Neurosciences**, v. 14, n. 3, p. 96–100, mar. 1991.

PANG, X. et al. Ubiquitin-proteasome pathway in skeletal muscle atrophy. **Frontiers in Physiology**, v. 14, p. 1289537, 17 nov. 2023.

PRZYTYCKI, P. F.; SINGH, M. Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations. **Cell Systems**, v. 10, n. 2, p. 193- 203.e4, fev. 2020.

PURCELL, S. et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. **The American Journal of Human Genetics**, v. 81, n. 3, p. 559–575, set. 2007.

RAULUSEVICIUTE, I. et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. **Nucleic Acids Research**, v. 52, n. D1, p. D174–D182, 5 jan. 2024.

ROCHA, R. D. F. B. et al. Parent-of-origin effects for the number of oocytes and embryos in Gir cattle. **Livestock Science**, v. 280, p. 105412, fev. 2024.

RODACKI, C. L. et al. Fish-oil supplementation enhances the effects of strength training in elderly women. **The American Journal of Clinical Nutrition**, v. 95, n. 2, p. 428–436, fev. 2012.

SCHOEFTNER, S.; BLASCO, M. A. A ‘higher order’ of telomere regulation: telomere heterochromatin and telomeric RNAs. **The EMBO Journal**, v. 28, n. 16, p. 2323–2336, 19 ago. 2009.

SEBE, M. et al. Saturated fatty acids intake is associated with muscle atrophy in rheumatoid arthritis. **JCSM Rapid Communications**, v. 5, n. 1, p. 86–101, jan. 2022.

SMITH, J. et al. Opposite Regulation of CD36 Ubiquitination by Fatty Acids and Insulin. **Journal of Biological Chemistry**, v. 283, n. 20, p. 13578–13585, maio 2008.

SONG, Q. et al. A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics. **PLoS ONE**, v. 8, n. 12, p. e81148, 6 dez. 2013.

ST. PIERRE, C. L. et al. Genetic, epigenetic, and environmental mechanisms govern allele-specific gene expression. **Genome Research**, v. 32, n. 6, p. 1042–1057, jun. 2022.

TACHTSIS, B.; CAMERA, D.; LACHAM-KAPLAN, O. Potential Roles of n-3 PUFAs during Skeletal Muscle Growth and Regeneration. **Nutrients**, v. 10, n. 3, p. 309, 5 mar. 2018.

TIZIOTO, P. C. et al. A SNP in ASAP1 gene is associated with meat quality and production traits in Nelore breed. **Meat Science**, v. 92, n. 4, p. 855–857, dez. 2012.

TIZIOTO, P. C. et al. Genome scan for meat quality traits in Nelore beef cattle. **Physiological Genomics**, v. 45, n. 21, p. 1012–1020, 1 nov. 2013.

TIZIOTO, P. C. et al. Calcium and potassium content in beef: Influences on tenderness and associations with molecular markers in Nellore cattle. **Meat Science**, v. 96, n. 1, p. 436–440, jan. 2014.

TYCKO, B. Allele-specific DNA methylation: beyond imprinting. **Human Molecular Genetics**, v. 19, n. R2, p. R210–R220, 15 out. 2010.

VACARESSE, N. et al. Activation of Epithelial Growth Factor Receptor Pathway by Unsaturated Fatty Acids. **Circulation Research**, v. 85, n. 10, p. 892–899, 12 nov. 1999.

VAN DE GEIJN, B. et al. WASP: allele-specific software for robust molecular quantitative trait locus discovery. **Nature Methods**, v. 12, n. 11, p. 1061–1063, nov. 2015.

VOHRA, M. et al. SNPs in Sites for DNA Methylation, Transcription Factor Binding, and miRNA Targets Leading to Allele-Specific Gene Expression and Contributing to Complex Disease Risk: A Systematic Review. **Public Health Genomics**, v. 23, n. 5–6, p. 155–170, 2020.

VORN, R.; YOO, H. Y. Differential effects of saturated and unsaturated fatty acids on vascular reactivity in isolated mesenteric and femoral arteries of rats. **The Korean Journal of Physiology & Pharmacology**, v. 23, n. 5, p. 403, 2019.

ZHBANNIKOV, I. Y. et al. **SeqyClean: A Pipeline for High-throughput Sequence Data Preprocessing**. Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics. **Anais**... Em: BCB '17: 8TH ACM INTERNATIONAL CONFERENCE ON BIOINFORMATICS, COMPUTATIONAL BIOLOGY, AND HEALTH INFORMATICS. Boston Massachusetts USA: ACM, 20 ago. 2017. Disponível em: <https://dl.acm.org/doi/10.1145/3107411.3107446>. Acesso em: 5 dez. 2024

ZHOU, X.; STEPHENS, M. Genome-wide efficient mixed-model analysis for association studies. **Nature Genetics**, v. 44, n. 7, p. 821–824, jul. 2012.

5 GENERAL CONCLUSIONS

Through this doctorate, thousands of new variants were identified as candidates to regulate gene expression and phenotypic variance in Nelore muscle. These are the first studies to identify aseQTLs, DASE SNPs, PO-QTLs, and ASMs in *B. indicus* muscle.

By analyzing ASE data from 190 Nelore animals using a dense imputation data, we generated the largest dataset of ASE SNPs for this tissue in cattle. This approach proved to be a powerful tool for identifying putative *cis*-regulatory variants, as we could obtain aseQTLs and *cis*-eQTLs with evidence to be regulated by multiple regulatory mechanisms. These variants showed evidence to affect genes functionally related to proteasomal proteolysis, apoptosis, and autophagy, pathways that can impact the tenderization process in the *post mortem* muscle.

Identifying DASE SNPs also leveraged evidence of variants in regions of allelic imbalance affecting muscle phenotypes. Although being very conservative, this analysis showed to be a very useful one-step approach to searching for SNPs that can directly affect gene expression and phenotypic variance. From the DASE genes, we were able to identify co-expression modules related to processes such as muscle growth and protein ubiquitination, drafting hypotheses of the dynamics of apoptosis in meat tenderness.

PO-QTLs provide new evidence of how parent-of-origin effects influence beef phenotypes. By integrating ASE and ASM data, and observing consistent ASM patterns across all tested samples, we highlighted the overlap of known imprinted genes within ASM regions located in promoters, also suggesting novel imprinting candidates in cattle. The regulatory potential of PO-QTLs and ASM regions was further demonstrated through the prediction of TF and miRNA motifs.

In summary, our work provides a comprehensive list of potentially regulatory regions associated with bovine muscle processes. These regions represent promising candidates for experimental validation and future studies aimed at improving meat quality traits in *Bos indicus* through the modulation of regulatory dynamics.

6 ONGOING RESEARCH AND MANUSCRIPTS IN PREPARATION

Despite the presented papers aligning with the objectives of the proposed PhD project, we also collaborated on other projects, whose related manuscripts are still in preparation.

During a one-year internship abroad at the University of California, Davis, CA, USA (UC Davis), we actively contributed to multiple data analyses for the Bovine FAANG project—an international consortium focused on identifying functional elements across various bovine tissues. The first manuscript from this initiative is currently under development. Briefly, for this paper, we processed multi-tissue RNA-seq and WGBS data, evaluated the tissue specificity of gene expression and methylation profiles, and predicted regulatory elements using ATAC-seq and ChIP-seq data from the same samples. These findings will significantly enhance our understanding of regulatory elements in bovine genomes and are expected to have a substantial impact in this field. Given the involvement of multiple international institutions in this manuscript, we have agreed not to include the results in the thesis to avoid potential conflicts and to prioritize the publication of this paper in a high-impact journal, emphasizing its novelty. However, we have included three abstracts summarizing the results of this study, which were presented at the Plant and Animal Genomes Conference (PAG 31) in January 2024, as shown in Attachment 1.

Additionally, as another outcome of the internship abroad, we expect to produce a highly impactful publication linked to the Bovine FAANG project, presenting the largest catalog of methylation QTLs (meQTLs) in bovine genomes. This study is based on the processing of more than 1,400 publicly available methylomes (RRBS and WGBS). The analysis is ongoing, and Jennifer will be the first author of this future publication.

This thesis is part of the Thematic Project “O Hologenoma de Nelore: implicações na qualidade de carne e em eficiência alimentar” (FAPESP number 2019/04089-2). An additional project conducted within the Thematic Project led to the development of a tool that processes metagenomes to identify host miRNAs (i.e., bovine) with binding affinity to motifs from their microbiota. The software is available on Jennifer’s GitHub (https://github.com/JBruscadin/HolomiRA), and the manuscript

is currently under coauthor review in preparation for publication. Jennifer is the first author of this publication, which is available in Attachment 2.

ATTACHMENT 1

Title: “Analysis of Multi-Tissue Bovine Epigenomes Uncovers the Tissue-Specific Mechanisms Underlying Gene Expression “

Presented as panel by Jennifer Bruscadin on PAG 31

Available at: https://pag.confex.com/pag/31/meetingapp.cgi/Paper/52930

Authors and Institutions:

Jennifer Jessica Bruscadin[1,2], Dailu Guan[1], Claire Prowse-Wilkins[1], Xiaoqin (Shally) Xu[1], Bruna Petry[3], Wenjing Yang[1], Ying Wang[1], Gonzalo Rincon[1], Sarah Corum[4], Honglin Jiang[5], Brenda M. Murdoch[6], Stephanie McKay[7], Monique Rijnkels[8], Clare A. Gill[8], Timothy Smith[9], Zhihua Jiang[10], Wansheng Liu[11], James E. Koltes[3], James M. Reecy[3], Juan F. Medrano[1], Pablo Juan Ross[12], Luciana C.A. Regitano[13] and Huaijun Zhou[12], (1)Department of Animal Science, University of California, Davis, CA, (2)Center of Biological and Health Sciences, Federal University of São Carlos, São Carlos, SP, Brazil, (3)Department of Animal Science, Iowa State University, Ames, IA, (4)Zoetis, Kalamazoo, CA, (5)Department of Animal and Poultry Sciences, VirginiaTech, Blacksburg, VA, (6)University of Idaho, Moscow, ID, (7)University of Missouri, Columbia, MO, (8)Texas A&M University, College Station, TX, (9)USDA, ARS, Clay Center, NE, (10)Washington State University, Pullman, WA, (11)Department of Animal Science, Penn State University, University Park, PA, (12)Department of Animal Science, University of California, Davis, Davis, CA, (13)Embrapa Southeast Livestock, São Carlos, SP, Brazil

Abstract

Bovine FAANG (Functional Annotation of Animal Genomes) is an international effort to functionally annotate regulatory elements in cattle. Multi-omics data integration identifies regulatory patterns as active enhancers and promoter methylation, helping to identify causal variants responsible for complex traits for precise genomic predictions. As part of this project, this work aimed to understand the tissue-specific mechanisms of gene regulation through a comprehensive and integrated analysis of

multi-omics data from various bovine tissues and cells in different biological states. We harmonized and analyzed 478 *Bos taurus* datasets encompassing 55 tissues and two development stages: H3K27ac ChIP-seq data (138 samples), ATAC-seq data (91), RNA-seq data (158) and WGBS data (91). All alignment procedures were performed according to best-practices standards used in previous FAANG publications against the ARS-UCD1.2 reference genome. From transcripts per million (TPM)-normalized gene counts, we analyzed the tissue-specificity of gene expression using the Z-score method with the cut-off value greater than 0.8 (ranging from 0 to 1). For each tissue and the respective tissue-specific genes (TSGs), gene ontology biological process (BP) enrichment analyses were obtained using the enrichR package. To identify potential tissue-specific enhancer regulation, we applied the Activity-By-Contact model from the TPM gene counts, alignment and peak files from H3K27ac and ATAC-seq data. To increase the accuracy of predictions, we only kept enhancers overlapping predicted enhancer states and affecting TSGs. We identified 14,909 TSGs from six fetal tissues that enriched for 1,326 different BPs, and 11,573 genes were specifically expressed for 23 adult tissues and enriched for 1,771 BPs. Testis and lung were the tissues with more TSGs in adult and fetal samples, respectively, followed by brain tissue for both development stages. 3,626 BPs were enriched from TSGs, ranging from 3 for thyroid tissue to 602 for fetal lung. We identified 44,070 target gene-enhancer links between 6,786 TSGs and 31,503 enhancers from 23 tissues. Fetal brain showed the largest number of TSGs identified as target genes potentially affected by enhancer regulation (2,045). The enhancer with the strongest interaction score was predicted in a hypothalamus sample for the ENSBTAG00000049135 (novel) TSG. Additionally, whole genome bisulfite sequencing (WGBS) data were used to identify differentially methylated regions (DMRs) between tissues using the software wgbs-tools (FDR ≤ 0.05). We identified 208,665 differentially methylated regions (DMRs) across tissues, with 3,370 located up to 200 bp from the transcription start site of TSGs. Thus, the methylation profile was linked to tissue-specific gene expression, which could potentially affect BPs that are relevant to the functions of the tissue. Further analysis will determine the co-regulation between DNA methylation, gene

expression and chromatin states and explore the potential effect of epigenetic regulation regarding biological functions and complex trait architecture. This work will integrate information about the regulatory landscape of bovine tissues, helping identify potential causative variants for genetic improvement in cattle**.**

Title: "Epigenomic Regulation of Gene Transcription across Tissues at Different Developmental Stages "

Orally presented by Dr. Huajun Zhou (Workshop: Functional Annotations of Animal Genomes (FAANG) on PAG 31)

Available at: https://pag.confex.com/pag/31/meetingapp.cgi/Paper/52743

Authors and Institutions:

Huaijun Zhou[1], Dailu Guan[2], Claire Prowse-Wilkins[3], **Jennifer Bruscadin[4]**, Brenda M. Murdoch[5], Honglin Jiang[6], James E. Koltes[7], Monique Rijnkels[8], Pablo Juan Ross[4], Sarah Corum[9], Timothy Smith[10], Wansheng Liu[11], Zhihua Jiang[12], Clare A. Gill[13] and Stephanie Mckay[14], (1)Department of Animal Science, University of California, Davis, Davis, CA, (2)Department of Animal Science, University of California Davis, Davis, CA, (3)Department of Animal Science, University of California, Davis, CA, (4)University of California Davis, Davis, CA, (5)University of Idaho, Moscow, ID, (6)Department of Animal and Poultry Sciences, VirginiaTech, Blacksburg, VA, (7)Iowa State University, Ames, IA, (8)Texas A&M University, College Station, TX, (9)Zoetis, Kalamazoo, CA, (10)USDA, ARS, Clay Center, NE, (11)Department of Animal Science, Penn State University, University Park, PA, (12)Washington State University, Pullman, WA, (13)Department of Animal Science, Texas A&M University, College Station, TX, (14)University of Missouri, Columbia, MO

Abstract

A comprehensive annotation of the genome enables us to better understand the mechanism of gene regulation and the genetic architecture of phenotypic traits. To achieve this, we embarked on the annotation of the bovine genome, utilizing a total of 794 newly generated datasets in combination with 353 publicly available datasets. There datasets were sampled from 41 adult tissues, 8 fetal tissues, and 6 primary cell lines, including epigenomic histone marks, (H3K4me1 (142), H3K4me3 (142), H3K27ac (138), H3K27me3 (132), H3K36me3 (59), H3K9me3 (68), CTCF (138)), DNA methylation (91); ATAC-seq (116), and RNA-Seq (158). The uniform data preprocessing and subsequent modeling using ChromHMM enabled us to annotate 10 distinct chromatin states (designated as E1-E10), which accounts for ~4.6% of the bovine genome, based on four core histone marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3) that were generated in 125 bovine epigenomes simultaneously. As expected, promoter-like states exhibited a higher enrichment around transcription start sites, while enhancers did not. Through a detailed comparison between fetal and the corresponding adult tissues, we were able to detect thousands of development stage-specific regulatory elements, shedding light on potential explanations for variations in gene expression. Similar analyses have been conducted between primary cell lines and bulk tissues. In our ongoing exploration of the contribution of regulatory elements to complex traits, we clustered genome-wide enhancers into 150 distinct modules based on their enhancer activity. Our next phase will focus on investigating the involvement of these enhancer modules in the roles of bovine complex traits. Collectively, our results demonstrated the crucial role of epigenomic information in unraveling the mechanisms of gene regulation, cellular differentiation, tissue development, and the genetics of complex traits in the bovine species.

Title: “Integrative Annotation of the Bovine Genome“

Presented as panel by Dailu Guan on PAG 31.

Available at: https://pag.confex.com/pag/31/meetingapp.cgi/Paper/53258

Authors and Institutions:

Dailu Guan[1], Claire Prowse-Wilkins[2], **Jennifer Bruscadin[3]**, Xiaoqin (Shally) Xu[3], Bruna Petry[4], Wenjing Yang[2], Ying Wang[5], Gonzalo Rincon[6], Sarah Corum[7], Honglin Jiang[8], Brenda M. Murdoch[9], Stephanie D. McKay[10], Monique Rijnkels[11], Clare A. Gill[11], Timothy P.L. Smith[12], Zhihua Jiang[13], Wansheng Liu[14], James E. Koltes[4], James M. Reecy[4], Juan F. Medrano[2], Pablo Juan Ross[3] and Huaijun Zhou[15], (1)Department of Animal Science, University of California, Davis, CA, 95616, USA, Davis, CA, (2)Department of Animal Science, University of California, Davis, CA, (3)University of California Davis, Davis, CA, (4)Iowa State University, Ames, IA, (5)UC Davis, Davis, CA, (6)University of California, Davis, Davis, CA, (7)Zoetis, Kalamazoo, CA, (8)Department of Animal and Poultry Sciences, VirginiaTech, Blacksburg, VA, (9)University of Idaho, Moscow, ID, (10)University of Missouri, Columbia, MO, (11)Texas A&M University, College Station, TX, (12)USDA, ARS, Clay Center, NE, (13)Washington State University, Pullman, WA, (14)Department of Animal Science, Penn State University, University Park, PA, (15)Department of Animal Science, University of California, Davis, Davis, CA

Abstract

A comprehensive annotation of the genome enables us to better understand the mechanism of gene regulation and the genetic architecture of phenotypic traits. To achieve this, we embarked on the annotation of the bovine genome, utilizing a total of 794 newly generated datasets in combination with 353 publicly available datasets. There datasets were sampled from 41 adult tissues, 8 fetal tissues, and 6 primary cell lines, including epigenomic histone marks, (H3K4me1 (142), H3K4me3 (142), H3K27ac (138), H3K27me3 (132), H3K36me3 (59), H3K9me3 (68), CTCF (138)), DNA methylation (91); ATAC-seq (116), and RNA-Seq (158). The uniform data preprocessing and subsequent modeling using ChromHMM enabled us to annotate

10 distinct chromatin states (designated as E1-E10), which accounts for ~4.6% of the bovine genome, based on four core histone marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3) that were generated in 125 bovine epigenomes simultaneously. As expected, promoter-like states exhibited a higher enrichment around transcription start sites, while enhancers did not. Through a detailed comparison between fetal and the corresponding adult tissues, we were able to detect thousands of development stage-specific regulatory elements, shedding light on potential explanations for variations in gene expression. Similar analyses have been conducted between primary cell lines and bulk tissues. In our ongoing exploration of the contribution of regulatory elements to complex traits, we clustered genome-wide enhancers into 150 distinct modules based on their enhancer activity. Our next phase will focus on investigating the involvement of these enhancer modules in the roles of bovine complex traits. Collectively, our results demonstrated the crucial role of epigenomic information in unraveling the mechanisms of gene regulation, cellular differentiation, tissue development, and the genetics of complex traits in the bovine species.

ATTACHMENT 2

HolomiRA: a reproducible pipeline for miRNA binding site prediction in microbial genomes

Jennifer Jessica Bruscadin[1,2†], Tainã Figueiredo Cardoso[2†*], Liliane Costa Conteville[2], Juliana Virginio da Silva[1,2], Adriana Mércia Guaratini Ibelli[2], Gabriel Alexander Colmenarez Pena[2], Thanny Porto[1,2], Priscila Silva Neubern de Oliveira[1,2], Bruno Gabriel Andrade Nascimento[3], Adhemar Zerlotini[4], Luciana Correia de Almeida Regitano[2*]

[1] Center of Biological and Health Sciences, Federal University of São Carlos, São Carlos, São Paulo, Brazil;

[2] Embrapa Southeastern Livestock , São Carlos, São Paulo, Brazil;

[3] Munster Technological University, Ireland;

[4] Embrapa Digital Agriculture , Campinas, São Paulo, Brazil

[†]These authors contributed equally to this work

Abstract

Small RNAs (sRNAs), such as microRNAs (miRNAs), are candidates to mediate communication between the host and its microbiota, regulating bacterial gene expression and influencing microbiome functions and dynamics. Here, we present HolomiRA (Holobiome miRNA Affinity Predictor), a computational pipeline developed to predict binding motifs for host miRNAs in genomes of microbiomes. HolomiRA operates within a Snakemake workflow, processing microbial genomic sequences in FASTA format by using freely available bioinformatics software along with built-in data processing methods. The pipeline first annotates protein-coding sequences from the microbial genomes using Prokka, then identifies candidate regions and tests these regions for potential host miRNA binding sites with

RNAHybrid. The predicted results that meet the quality filter parameters are summarized and then used to perform a functional analysis of the affected genes using SUPER-FOCUS and DIAMOND. In this paper, we demonstrate the use of the HolomiRA pipeline by applying it to publicly available metagenome-assembled genomes obtained from human feces, as well as from cattle feces and rumen microbiomes. This approach enabled the determination of bacterial genes and biological pathways within microbiomes that could be influenced by host miRNAs, as well as the identification of shared or unique miRNAs, genes, and taxonomies across phenotypes, environments, or host species. HolomiRA is a practical and user-friendly pipeline that facilitates studies of miRNA prediction in prokaryotic genomes, shedding light on the communication between microbiota and host cells mediated by miRNA regulation. HolomiRA is publicly available on GitHub: https://github.com/JBruscadin/HolomiRA.

1 Introduction

MicroRNAs (miRNAs), non-coding RNAs that post-transcriptionally regulate gene expression in eukaryotes (Gebert et al. 2019), have recently been recognized as essential genetic regulators in mediating communication between the host and its microbiota (Ricci et al. 2022). Through extracellular vesicles (EVs), miRNAs can be transferred between cells, constituting an important mode of communication with the potential to modulate microbial communities and host-microbe interactions (Tsatsaronis, et al., 2018).

In humans, fecal miRNAs are imbalance indicators at the host-microbe interface (Moloney et al., 2018). They may operate indirectly on the host-microbiota interaction by repressing the translation of gene transcripts targeted by specific miRNAs, thereby influencing the microbial community (Goodrich et al., 2014), and/or enter bacterial cells through endocytosis, specifically regulating bacterial gene transcripts, affecting bacterial growth (Liu et al., 2016).

In eukaryotes, gene regulation mediated by miRNAs generally occurs through complementary binding between their seed region, located between the second to the seventh nucleotide of the miRNA, and regions of the target mRNAs called miRNA recognition elements (MREs) (O'Brien et al. 2018). In addition to this interaction, pairing a region called “supplemental” at the 3’-UTR end of the miRNA makes binding to the target more specific and stable (O'Brien et al. 2018). In prokaryotes, the translation of specific mRNAs is partially controlled by sRNAs (small RNAs), which pair antisense to mRNAs near ribosome binding sites (RBS) (Desnoyers et al. 2013). These sequences can extend from twenty nucleotides upstream to fifteen nucleotides downstream of the first nucleotide of the translation initiation codon (Desnoyers et al. 2013). By pairing with mRNA, sRNAs prevent the binding of the 30S ribosomal subunit to the RBS region, promoting translation repression and degradation of the target mRNA (Ahmed et al. 2016). However, the mechanisms of miRNA regulation in prokaryotic genomes are not fully described.

The identification of putative MREs can be carried out by software such as the widely used RNAHybrid (Kruger et al., 2006), which predicts microRNA targets based on minimum free energy hybridization of a long and a short RNA. RNAHybrid finds the energetically most favorable hybridization sites of a small RNA in a large RNA (Rehmsmeier et. al, 2004). As this software can be highly customizable for different study purposes, the user can define multiple parameters applied to any target mRNA and miRNA sequences available, not only in model organisms, making it possible to analyze host miRNAs affecting the prokaryotic genomes of the microbiome.

Therefore, aiming to expand our understanding of the functional role of miRNAs in the host-microbiome communication, we developed a pipeline to predict motifs for host miRNAs in genomes from their microbiota, which can be applied to any prokaryotic genome. Holobiome miRNA Affinity Predictor (HolomiRA) uses publicly available tools for sequence annotation, miRNA motif prediction and enrichment analysis. Each bioinformatics software integrated in HolomiRA is fine-tuned for the target prokaryotic genomes. Our pipeline is the first one able to test the interaction between host miRNAs and their microbiome and can greatly contribute to the knowledge of the host regulatory interaction on the activity of the microbiome.

2 Methods

2.1 HolomiRA pipeline

HolomiRA provides an accessible and user-friendly pipeline for searching host miRNA binding sites within microbial genomes, based on the growing evidence on the regulation of microbiota by sRNAs. The workflow automates the installation and configuration of the correct software versions across different computational systems. This approach mitigates potential version conflicts and enhances usability, making it accessible for those new to sequence analysis.

In the following subsections, we describe all procedure steps of the pipeline, which are performed with a single command line in HolomiRA.

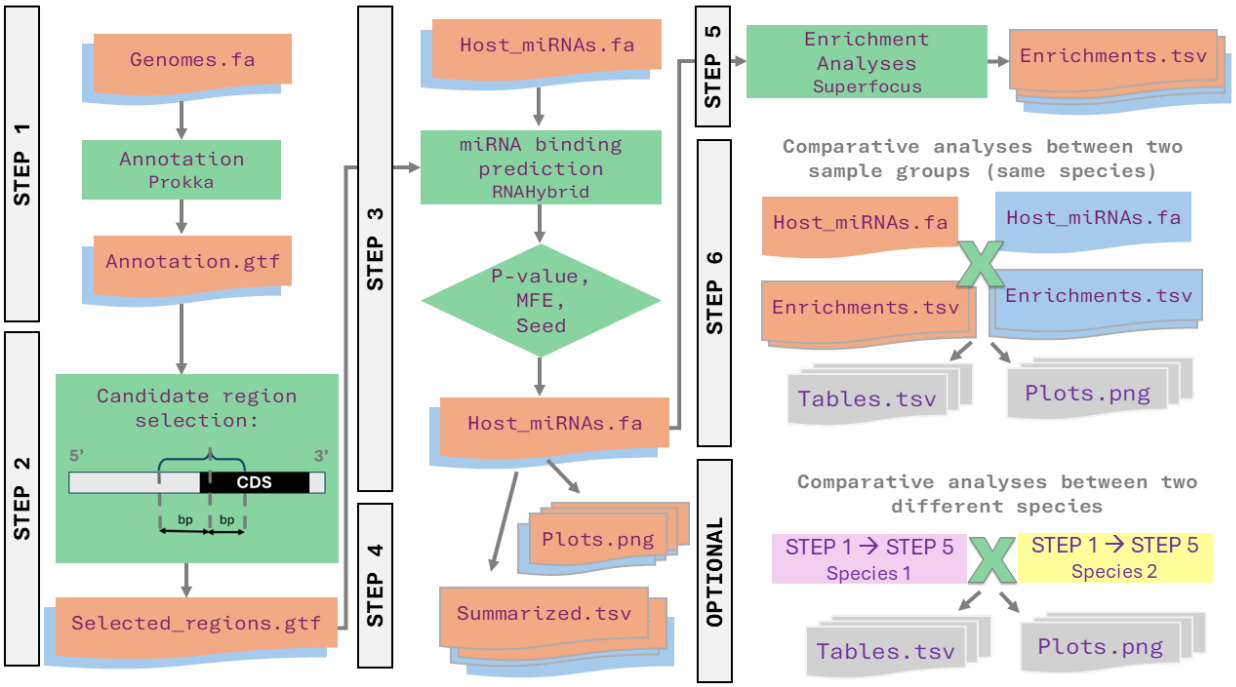


Figure 1 – Schematic representation of HolomiRA pipeline; CDS = Coding Sequence Stands; MFE = Minimum Free Energy. The “Genomes.fa” entry can be any prokaryotic genome.

Step 1 – Predict coding sequences from microbial genomes

HolomiRA expects genomic DNA sequences in FASTA format as input. These sequences are annotated using Prokka (SEEMANN, 2014), which first identifies the coordinates of protein-coding genes using Prodigal (HYATT et al., 2010), and subsequently then compares these sequences to various public databases to obtain the most accurate and reliable annotation of the given protein.

Step 2 - Select candidate target regions

The annotated coding sequences (CDS) are extracted from the annotation files (.gtf extension) generated by Prokka. The coordinate of the most 5’ nucleotides from each CDS region is considered to draft the genomic window that delimits the candidate regions for miRNA-binding sites screening (Figure 1). For this, users must

choose the appropriate upstream and downstream base pairs around the CDS start site based on their previous knowledge of the biological model. If the selected start coordinates fall outside the borders of any contig, this start coordinate is changed to zero for this given sequence. Subsequently, candidate sequences are retrieved by filtering the genomes by the selected coordinates to be used for miRNA-binding site prediction.

Step 3 – Search for host miRNAs binding sites in the candidate regions

This step relies on the software RNAHybrid (KRUGER; REHMSMEIER, 2006). It identifies the most energetically favorable hybridization sites of a small RNA in a large genomic section, avoiding pairings within the miRNA sequence or within target nucleotides (KRUGER; REHMSMEIER, 2006). This step is carried out by using as input the candidate CDS sequences obtained in step 2 and the sequences of the host miRNAs, which can be obtained from public databases such as miRDB (CHEN; WANG, 2020), miRbase (KOZOMARA; BIRGAOANU; GRIFFITHS-JONES, 2019), and MirGeneDB (FROMM et al., 2022). The user must include in the configuration file the parameters for filtering results, such as the p-value threshold and the minimum free energy (MFE) cut-offs. Optional parameters are encouraged to be included to establish the starting position and length of the miRNA seed region that must match the DNA binding region (refer to the RNAHybrid documentation for more information).

Step 4 – Get summarized metrics and plots from the results

HolomiRA creates different outputs to better organize and improve the visualization of results from the predicted host miRNA binding sites provided by the previous step. The output tables show all results, a table summarized by miRNA and a table summarized by genome. HolomiRA will generate bar plots with the top-ranked miRNAs by microbial genome and by target genes and the top-ranked microbial genome regarding the identified number of interacting miRNAs and target genes.

Step 5 – Functional Enrichment analysis

Finally, a functional analysis is performed, considering the genes affected by each miRNA and within each affected genome. To achieve this, Holomira utilizes the SUPER-FOCUS (SILVA et al., 2016) software.

Step 6 - Comparative analysis

The pipeline runs an additional analysis by comparing the results between different groups, such as phenotypes, case-control groups, or environments within the same species. Venn diagrams will be generated automatically to visualize the overlap of taxa, genes, and miRNAs between the separate sample groups analyzed with HolomiRA, as well as to compare enrichment analysis results. The user can define the groups directly in the configuration file.

Additional Step - Comparative analysis between different results sets

Additional scripts are made available to allow comparisons between different species, by comparing results obtained from independent HolomiRA analyses.

2.3 Usage

HolomiRA parameters are set in a configuration file. Its elements are: (1) the path for the microbial genomes' directory, (2) the output directory specification, (3) a file containing the sample list, (4) integer values of upstream and downstream base pairs flanking the CDS start position to indicate the candidate target region, (5) host miRNA sequences file path, (6) the parameters for the miRNA hybridization sites (seed position and length, MFE, and p-value thresholds), and (7) an additional file containing information about the microbial genomes, including their taxonomic classification and the environment they were collected (*i.e.*, any metadata as tissue type, species, phenotype, etc.).

After customizing the configuration file, just one command line is needed to run all steps of the pipeline: snakemake -s Snakefile --use-conda […]. The results are displayed in one output and summarized in two different tables based on the miRNAs or the taxonomy features of the genomes. All HolomiRA’s documentation and source code are publicly available at https://github.com/JBruscadin/HolomiRA.

2.4 Real data

We evaluated HolomiRA using two publicly available datasets of metagenome-assembled genomes (MAGs) derived from human feces (FORSTER et al., 2019) and cattle feces and ruminal content (XIE et al., 2021). The human feces dataset contained 39,913 high-quality MAGs, while the ruminant dataset comprised 10,000 high-quality MAGs. To reduce redundancy and optimize analysis time, we filtered the datasets to remove redundant MAGs at the species level and at the genus level for those classified only to the genus. After filtering, we obtained and analyzed 184 MAGs from the human feces, and 293 and 336 MAGs from the cattle feces and ruminal content, respectively. The complete human feces MAG dataset is available at ftp://ftp.ebi.ac.uk/pub/databases/metagenomics/hgg_mags.tar.gz (FORSTER et al., 2019), and the complete cattle gastrointestinal MAG dataset is accessible through the European Nucleotide Archive (ENA) database, project PRJNA657473 (XIE et al., 2021).

We downloaded human and bovine miRNAs sequences from the MirGeneDB database (FROMM et al., 2022), which contains manually curated and validated miRNAs genes. In total, 630 human miRNAs and 481 bovine miRNAs sequences were used in our analyses. We ran the analysis in all selected samples using the following parameters: 15 as the upstream base pairs, 20 as the downstream base pairs, seed region starting from the second nucleotide with eight bases length, -18 as the maximal MFE, and 0.01 as the p-value threshold. The analyses were run for each host using only procedures implemented by the HolomiRA workflow.

3. Outputs

3.1. General outputs

Different outputs and plots were generated to summarize the results found. HolomiRA allowed us to identify which miRNAs can hybridize with bacterial CDS sequences, as well as to identify which genes are being affected. The outputs also include results from Venn diagram analyses, which help identify miRNAs, genes, and taxonomies shared across or unique to each category group, grouped by

phenotype or environment defined by the user. Additionally, they provide information on the biological pathways influenced by each miRNA and highlight the pathways affected within each genome.

In our test dataset, it was possible to identify that 476 and 466 miRNAs are putative candidates for binding with 3,841 and 4,078 bacterial genes in cattle rumen and feces, respectively (Figure 2). In addition, it is possible to identify which miRNAs are impacting the largest number of genomes (Figure 3), as well as how many and on which genes.

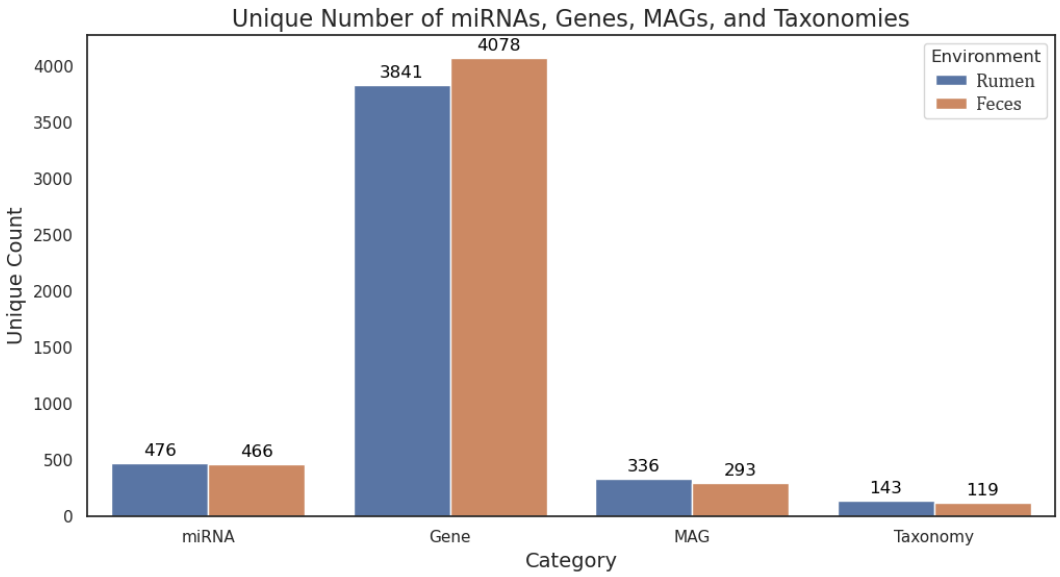


Figure 2 - Count of unique features from the predicted interactions of miRNA-microbiome in cattle, according to HolomiRA.

When more than one category group from the same host species is investigated, the pipeline also identifies lists of candidates that impact only one of the groups (Figure 4). In our example, 12 miRNAs were predicted to impact only MAGs from rumen, while 2 miRNAs are exclusive from feces. The SUPER-FOCUS program, incorporated into our pipeline, provides the user with access to information on different levels of pathways and affected biological functions.

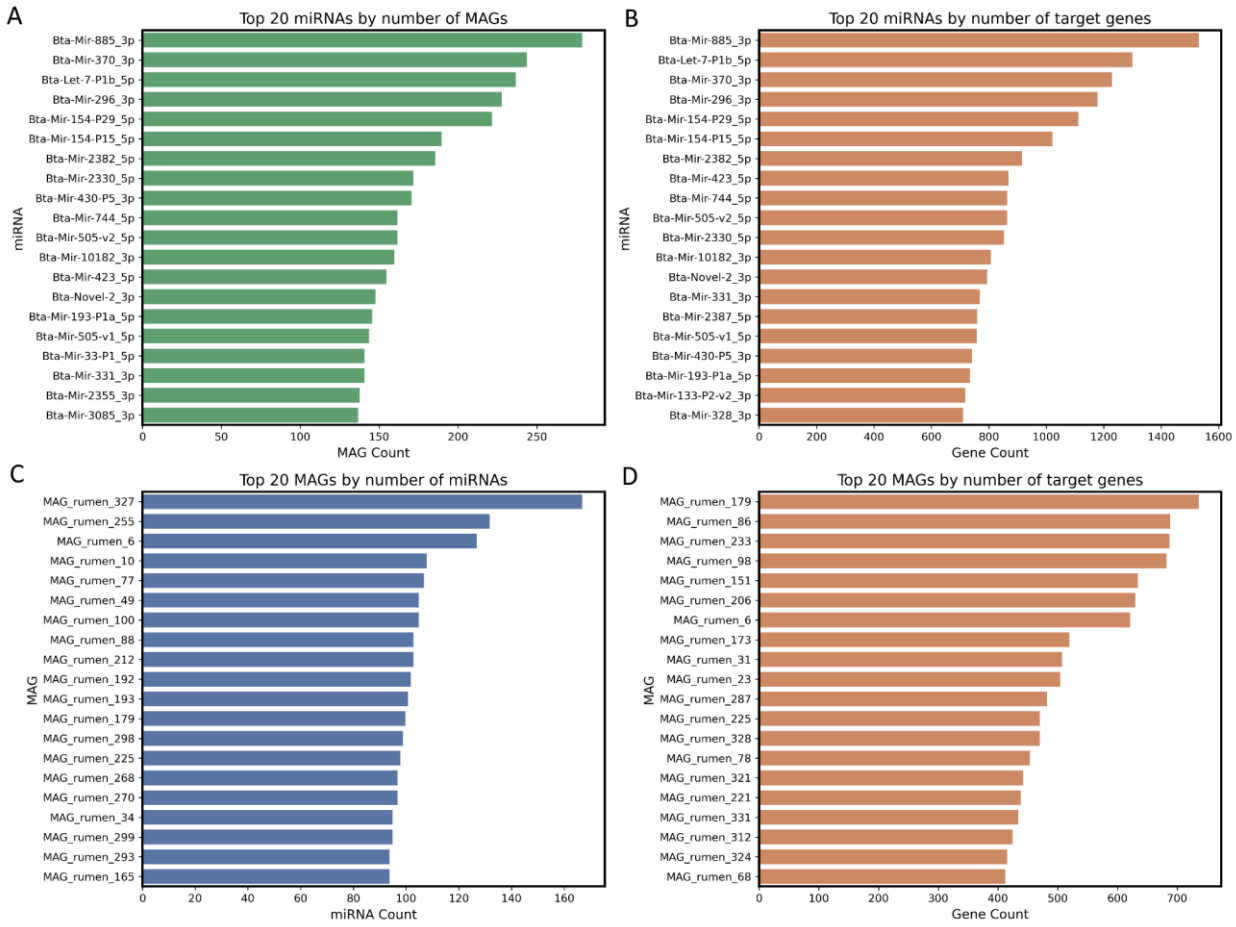


Figure 3 - Top20 features found in Rumen based on number of results. (A) Top20 miRNAs based on the number of MAGs with the related motif; (B) Top 20 miRNAs based on the number of affected target genes; (C) Top20 MAGs based on the number of miRNAs with predicted motifs; and (D) Top20 MAGs based on the number of target genes with predicted miRNA motifs.

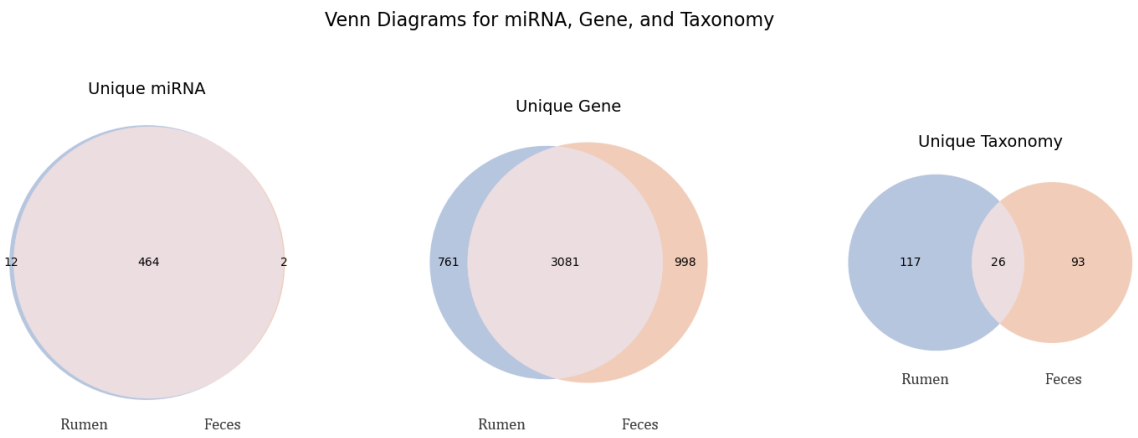


Figure 4. Venn diagrams illustrating shared and unique miRNAs, genes, and taxonomies across different sample groups of the same species. Overlapping regions represent common elements, while non-overlapping areas highlight specific features unique to each condition.

3.2. Optional outputs

To facilitate visualization of the biological processes associated with the genes potentially affected by miRNAs, supplementary scripts are provided as an attachment to the HolomiRA pipeline. These scripts allow users to perform functional analyses of biological datasets and generate plots highlighting the most abundant functions, based on user-defined thresholds (Figure 5).

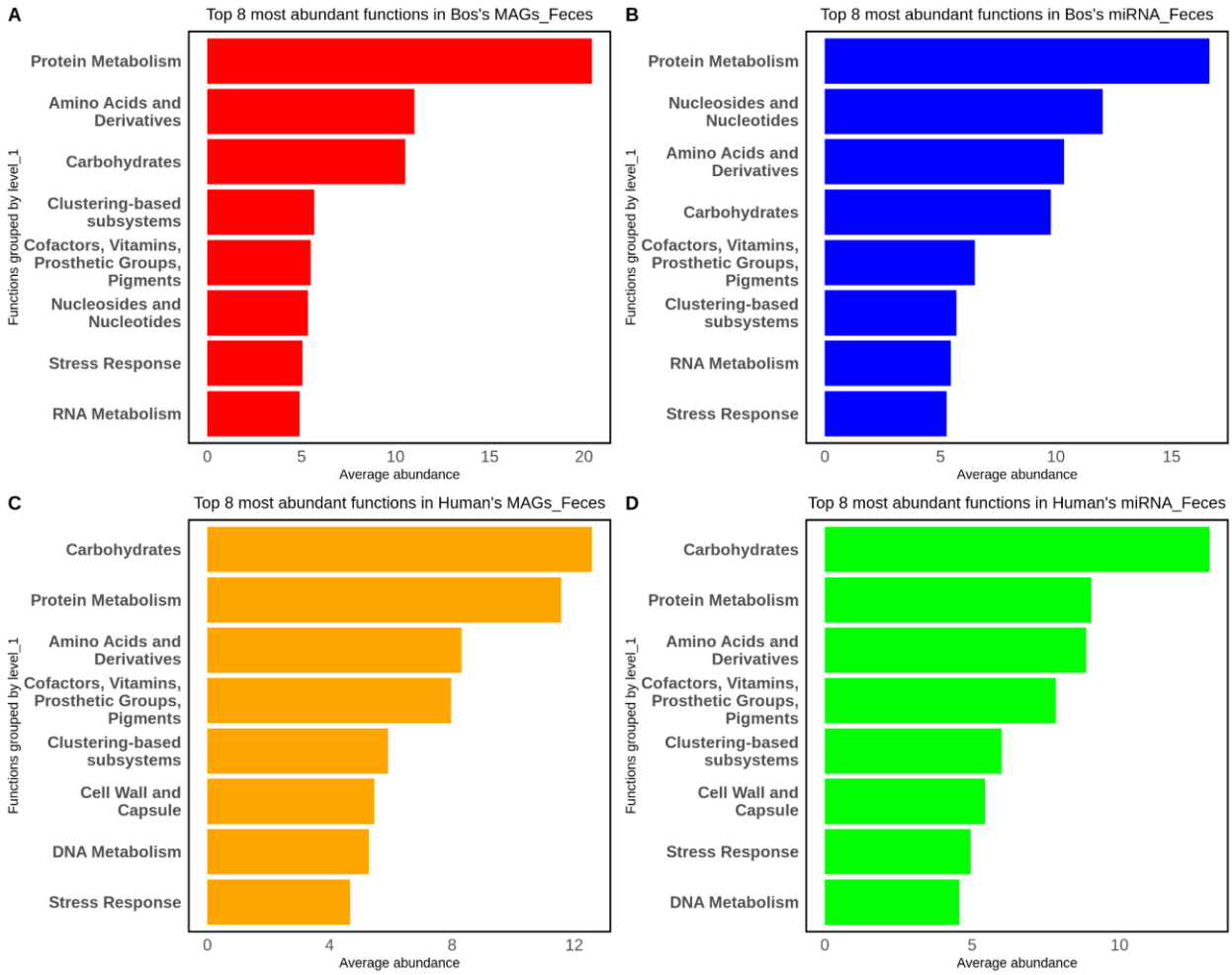


Figure 5 - The most abundant functions among the predicted targets of cattle miRNA in the microbiome. grouped by level 1 of the SUPER-FOCUS from the following biological datasets: (A) Bovine - MAGs from feces, (B) Bovine - fecal miRNA, (C) Human - MAGs from feces, and (D) Human - fecal miRNA.

Furthermore, when different sample groups are provided, even from different species, scripts were designed to perform comparative statistical analysis between two datasets, specifically focusing on the comparison of functional abundances. The script filters functions based on user-defined abundance and prevalence thresholds, checks for common functions between datasets, and performs statistical tests (Wilcoxon rank-sum) to identify significantly different functions between the groups. For example, in Figure 6, we performed a comparative analysis between the functions, grouped by SUPER-FOCUS level 2, of genes affected by each miRNA obtained in the feces of the Bovine (in purple) and of the Human (in orange). We

used 0, 0.2, 0.05 and 0.5 for abundance, prevalence thresholds, p-value and log thresholds respectively. The Venn diagram, shown in Figure 6A, revealed 52 functions shared between the two biological datasets. Furthermore, the statistical analysis conducted using the Wilcoxon test and log2 fold change revealed that 18 functions were differentially abundant between the Bovine and Human, as illustrated in Figure 6B. In this figure, the most abundant functions in the Bovine are highlighted in purple, while those associated with humans are shown in orange. The functions in gray showed no statistical differences based on the analyzed p-value and log2 fold change thresholds. Figure 6C provides a detailed view of these enriched functions, with nine being more abundant in Bovine and nine in Human. Regarding unique functions, the Human dataset exhibits 60 unique functions, while no functions are exclusive to the Bovine (Figure 6A).

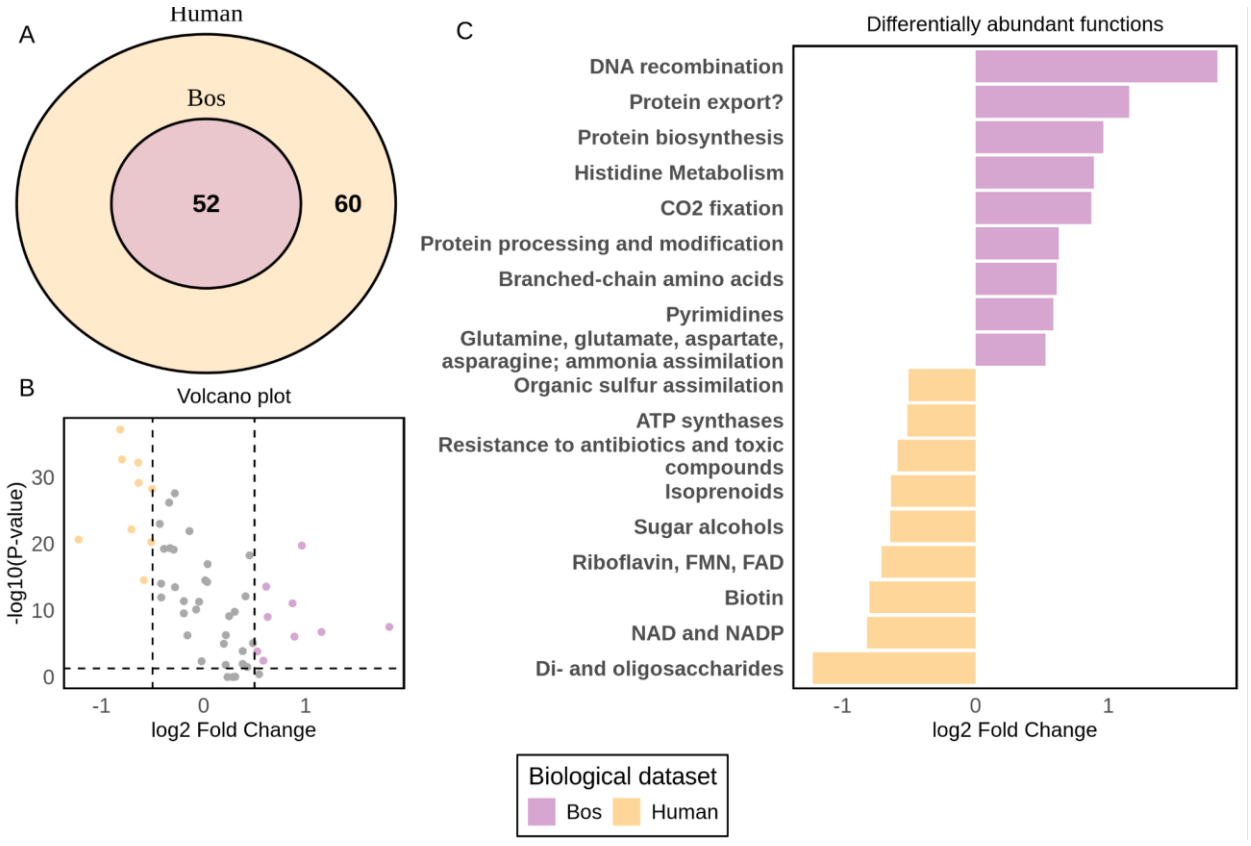


Figure 6: Functions grouped by SUPER-FOCUS level 2 of miRNA obtained from the feces of Bovine and Human. Purple bars represent the functions of miRNA from the Bovine, while orange bars represent the functions of miRNA from the Human. (A)

Venn diagram showing the functions unique and common to each organism. (B) Volcano plot showing the relationship between log2 fold change (X-axis) and adjusted p-value (-log10(p-value), Y-axis). The p-value threshold was 0.05 and the log2 fold change threshold was |0.5|. Functions in gray do not show significant differences between biological datasets for the thresholds analyzed. (C) Functions that showed statistical differences in abundance levels between Bos and Human.

Moreover, our additional scripts generate various outputs, including Venn diagrams, barplot, heatmap and files with common and exclusive functions. For example, in Figure 7 we performed a comparative analysis between the functions, grouped by SUPER-FOCUS level 3, of the ruminal MAGs (in dark blue) and fecal MAGs (in red) of Bovine targeted by the miRNAs. We used 0, 0.2, 0.05 and 0.5 for abundance, prevalence thresholds, p-value and log thresholds respectively. The Venn diagram shown in Figure 7A revealed 108 functions shared between the two environments of Bos. Furthermore, the statistical analysis conducted using the Wilcoxon test and log2 fold change revealed that nine functions were differentially abundant between the rumen and feces, as illustrated in Figure 7B. In this figure, a heatmap is presented, where the intensity of the colors indicates the relative abundance of differentially expressed functions in two environments: rumen and feces. Regarding unique functions, the rumen exhibited 60 unique functions, while feces contained only 2 unique functions. We selected the top 5 most abundant functions in rumen and the top 2 most abundant functions in feces (Figure 7C).

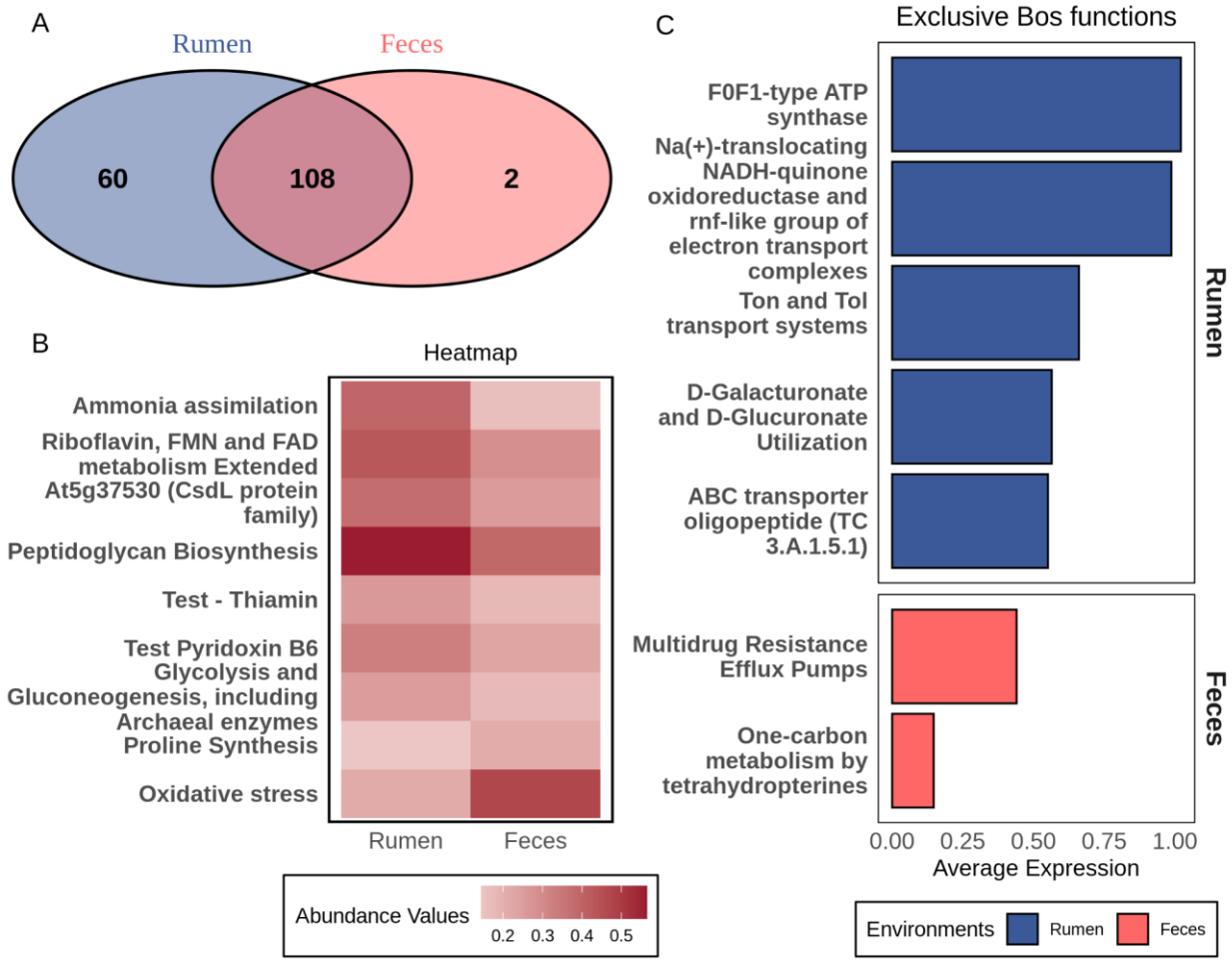


Figure 7: Functions grouped by SUPER-FOCUS level 3 of MAGs obtained from the rumen and feces of Bovine. Blue dark bars represent the functions of MAGs from the rumen, while red bars represent the functions of MAGs from the feces. (A) Venn diagram showing the functions unique and common to each environment. (B) Heatmap showing the significant functions between rumen and feces of Bovine' MAGs. (C) Barplot with unique functions to the rumen and feces.

Also, to facilitate the visualization of results when the pipeline is run individually (without informing the settings of the environments/phenotypes) or with different species, additional scripts are available to run histograms and Venn diagrams comparing more than one analysis (Figure 8).

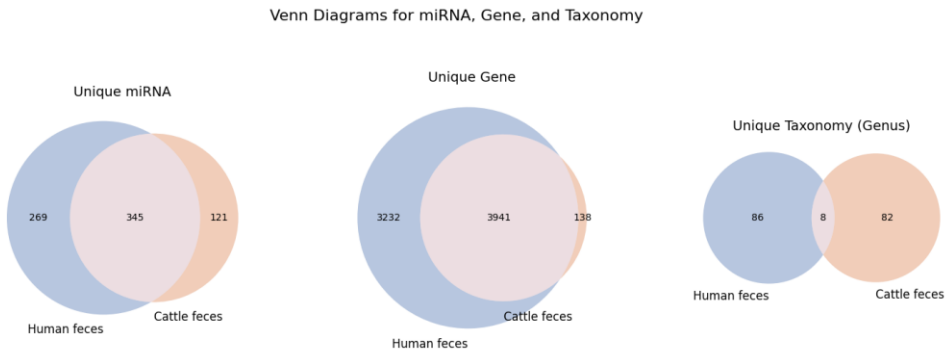


Figure 8. Histograms comparing unique and shared miRNAs, genes, and taxonomies between different sample species.

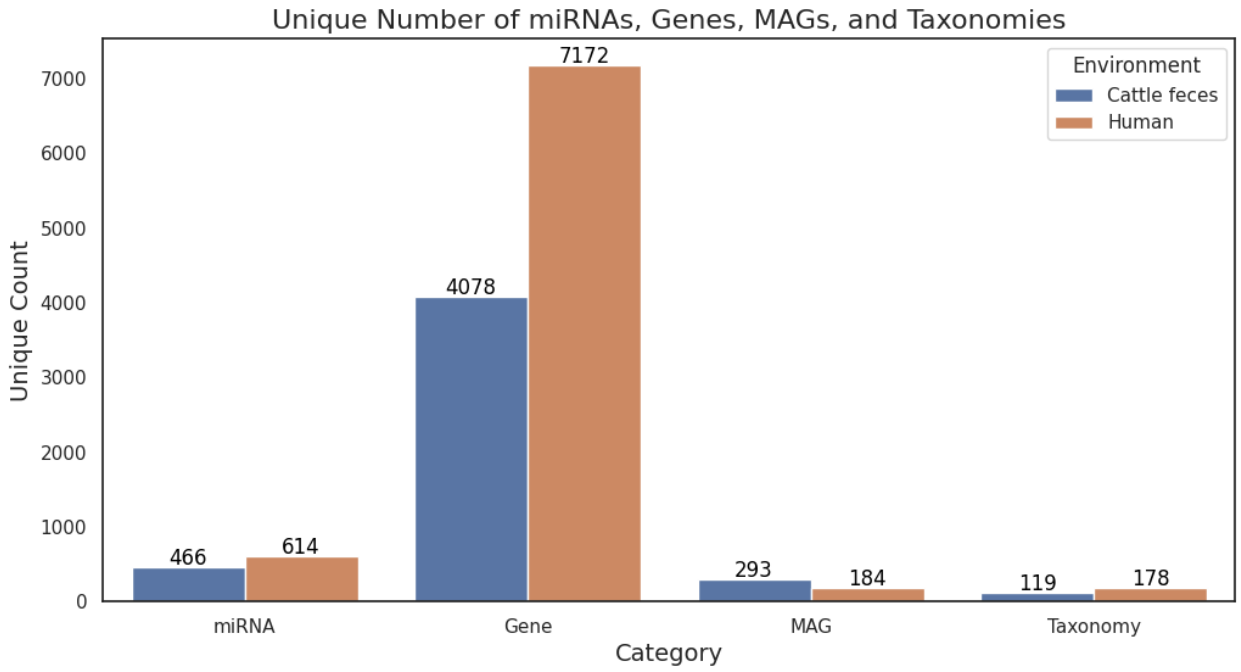


Figure 9. Histograms comparing unique miRNAs, genes, MAGs, and taxonomies between different sample species.

4. HolomiRA run times and memory use

Average run times and maximum memory used by each of the rules for 10 MAGs (mean 2.8M) and 100 miRNAs sequences.

	h:m:s	max_rss	io_in	io_out	mean_load	cpu_time
annotate_prokka	00:02:32	1317.198	26.94	99.196	34.775	367.743
cds	00:00:00	5.745	0	0.761	0	0.04
id	00:00:00	11.1	2.4	0.478	0	0.07
five_prime	00:00:02	30.717	3.968	3.661	34.439	1.268
find_targets	00:01:15	22.92	0.401	0.025	81.295	61.73
format_rnahybrid	00:00:02	82.305	5.062	0.149	48.125	1.386
aggregate_bsites	00:00:00	112.79	0.02	0.02	0	4.63
get_indiv_metrics	00:00:04	85.372	0.08	13.934	72.567	3.245
aggregate_finalresults	00:00:00	113.19	0.01	0.02	0	4.99
get_global_metrics	00:00:04	99.73	5.79	13.64	69.35	3.13
summary	00:00:05	155.04	70.45	0.32	66.41	3.95
plt_histogram	00:00:04	163.64	1.76	0.04	70.34	3.65
plt_venn	00:00:04	155.43	1.02	0.09	74.49	3.56
impacted	00:01:08	105.24	74.84	152.11	22.64	16.01
prep_superfocus	00:00:11	73.55	2.27	163.12	58.43	7.38
run_superfocus_MAG	00:28:56	5256.98	2134.67	1394.39	299.16	5421.93
run_superfocus_miRNA	04:31:41	5122.46	584.25	1426.47	89.83	30913.9

Conclusion

In summary, HolomiRA offers a robust and user-friendly framework for predicting host miRNA binding sites within prokaryotic genomes, enabling deeper insights into the regulatory interactions between hosts and their microbiomes. By integrating annotation, motif prediction, and functional analysis, this pipeline provides a comprehensive approach to uncovering the potential roles of miRNA in shaping microbiome function and dynamics. The application of HolomiRA to diverse datasets demonstrates its versatility and effectiveness, making it a valuable tool for advancing research in host-microbiome communication.

Bibliography

CHEN, Y.; WANG, X. miRDB: an online database for prediction of functional microRNA targets. **Nucleic Acids Research**, v. 48, n. D1, p. D127–D131, 8 jan. 2020.

DEAN M. **The Human ATP-Binding Cassette (ABC) Transporter Superfamily**. [s.l.] Bethesda (MD): National Center for Biotechnology Information (US), 2002.

ESPELT, M. V. et al. MicroRNAs contribute to ATP-binding cassette transporter- and autophagy-mediated chemoresistance in hepatocellular carcinoma. **World Journal of Hepatology**, v. 11, n. 4, p. 344–358, 27 abr. 2019.

FORSTER, S.C. et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. **Nature Biotechnology**, v. 37(2), p. 186–192, 4 feb. 2019

FROMM, B. et al. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. **Nucleic Acids Research**, v. 50, n. D1, p. D204–D210, 7 jan. 2022.

HUANG, H. et al. Identification of Six Thiolases and Their Effects on Fatty Acid and Ergosterol Biosynthesis in *Aspergillus oryzae*. **Applied and Environmental Microbiology**, v. 88, n. 6, p. e02372-21, 22 mar. 2022.

HYATT, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. **BMC Bioinformatics**, v. 11, n. 1, p. 119, dez. 2010.

KOZOMARA, A.; BIRGAOANU, M.; GRIFFITHS-JONES, S. miRBase: from microRNA sequences to function. **Nucleic Acids Research**, v. 47, n. D1, p. D155–D162, 8 jan. 2019.

KRUGER, J.; REHMSMEIER, M. RNAhybrid: microRNA target prediction easy, fast and flexible. **Nucleic Acids Research**, v. 34, n. Web Server, p. W451–W454, 1 jul. 2006.

LAI, L. et al. MicroRNA-33 Regulates the Innate Immune Response via ATP Binding Cassette Transporter-mediated Remodeling of Membrane Microdomains. **Journal of Biological Chemistry**, v. 291, n. 37, p. 19651–19660, set. 2016.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. **Bioinformatics**, v. 30, n. 14, p. 2068–2069, 15 jul. 2014.

SILVA, G. G. Z. et al. SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. **Bioinformatics**, v. 32, n. 3, p. 354–361, 1 fev. 2016.

TO, K. K. W. et al. Utilizing non-coding RNA-mediated regulation of ATP binding cassette (ABC) transporters to overcome multidrug resistance to cancer chemotherapy. **Drug Resistance Updates**, v. 73, p. 101058, mar. 2024.

WANG, Z. et al. Key enzymes involved in the utilization of fatty acids by Saccharomyces cerevisiae: a review. **Frontiers in Microbiology**, v. 14, p. 1294182, 11 jan. 2024.

XIE, F. et al. An integrated gene catalog and over 10,000 metagenome-assembled genomes from the gastrointestinal microbiome of ruminants. **Microbiome**, v. 9, n. 1, p. 137, 12 jun. 2021.