UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ZOOTECNIA E ENGENHARIA DE ALIMENTOS

GABRIELA RIBEIRO

The search for genetic functional variants of feed efficiency in Nelore cattle

Pirassununga
2021

GABRIELA RIBEIRO

The search for genetic functional variants of feed efficiency in Nelore cattle

Dissertação apresentada à Faculdade de Zootecnia e Engenharia de Alimentos da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Mestre em Qualidade e Produtividade Animal do programa de Mestrado Acadêmico de Zootecnia.

Área de Concentração: Melhoramento genético e biologia molecular

Orientador: Prof. Dr. Heidge Fukumasu
Co-orientadora: Prof. Dra. Aline Silva Mello Cesar

Pirassununga

2021

Ficha catalográfica

Ficha catalográfica elaborada pelo
Serviço de Biblioteca e Informação, FZEA/USP,
com os dados fornecidos pelo(a) autor(a)

Ribeiro, Gabriela
R484t The search for functional genetic variants of feed efficiency in Nellore cattle / Gabriela Ribeiro ; orientador Heidge Fukumasu ; coorientadora Aline Silva Mello Cesar. -- Pirassununga, 2021.
72 f.

Dissertação (Mestrado - Programa de Pós-Graduação em Zootecnia) -- Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo.

1. RNA-seq. 2. eficiência alimentar. 3. variantes funcionais. 4. gado. 5. GWAS. I. Fukumasu, Heidge , orient. II. Silva Mello Cesar, Aline , coorient. III. Título.

FOLHA DE APROVAÇÃO

Nome: Gabriela Ribeiro

Título: A procura por variantes genéticas funcionais da eficiência alimentar em bovinos Nelore

Tese apresentada á Faculdade de Zootecnia e Engenharia de Alimentos da Universidade de São Paulo, como parte dos requisitos para a obtenção do título de Mestre em Qualidade e Produtividade Animal do programa de Pós-graduação em Zootecnia.

Área de Concentração: Melhoramento genético e biologia molecular

Data: _____ /_____/ _____

BANCA EXAMINADORA

Prof. Dr._____
Instituição:_____Julgamento:_____
Prof. Dr._____
Instituição:_____Julgamento:_____
Prof. Dr._____
Instituição:_____Julgamento:_____

DEDICATÓRIA

Dedico este trabalho a todos os que me ajudaram ao longo desta caminhada.

AGRADECIMENTOS

Ao meu orientador e amigo, Professor Heidge Fukumasu, que serei grata eternamente por todos os ensinamentos biológicos, conselhos pessoais e profissionais. Por sempre acreditar no meu potencial e estimular o meu processo científico. Um exemplo de liderança e profissionalismo. Por sempre estimular a minha curiosidade com a famosa frase “Eu já sei a resposta disso, agora é sua vez de saber”, isso me ensinou a ir atrás das respostas.

A Professora Dra. Aline Cesar, por me ajudar com a parte técnica do projeto, sempre com toda a paciência do mundo, para que eu entenda as escolhas dos softwares e dos códigos, o que engrandeceu o meu conhecimento técnico em bioinformática.

A Dra. Pâmela Alexandre por ser uma amiga e uma professora incrível, que me fez me ajudou a entender o que é ser um pesquisador e sempre acreditou no meu potencial, obrigada por toda paciência e compreensão, nunca vou esquecer do que me falou “Gabi tudo na vida passa, os momentos bons e ruins vão passar, saiba aproveitá-los”.

Ao grupo de pesquisa do Dr. Hans D. Daetwler, a Dr. Amanda Chamberlain, a Dr. Jennie Price e Dr. Ruidong Xiang, por serem um exemplo de pesquisadores e profissionais, que abriu as postas para a realização da minha BEPE. Serei eternamente grata pelos ensinamentos e ajuda. Obrigada por me acolherem e me ajudarem muito durantes os 5 meses que trabalhei por lá.

A FZEA, ao programa de Zootecnia e a todos os professores, pesquisadores e funcionários que dele fazem parte. Obrigado pela oportunidade!

Aos meus amigos de longa data Tamires Romão, Giuliana e José Vitor Pizol, que sempre me apoiaram e se mantiveram ao meu lado independente das dificuldades.

Ao meu namorado Ricardo e a minha família de consideração, por todo apoio, companheirismo e paciência.

Ao Professor José Bento e a técnica de bioinformática Elisangela por toda ajuda e pela disponibilização do servidor.

Aos amigos que fiz no LOCT nesses anos. Especialmente a Yô, Nina, Porco, Roberta Berezin, Roberta, Victor, Pedro, Pedrinho, Pâmela, Tais, Jéssika e Lídia. Muitos momentos juntos, ida ao bandeco, churrascos e comemorações. Foi muito bom partilhar esses momentos com vocês.

A todos os amigos que a FZEA e Pirassununga me deram.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES).

A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) por todo o suporte financeiro para a execução desses trabalhos e pelas bolsas (processos 2019/01234-1 e 2019/18647-7).

"Tudo o que temos de decidir
é o que fazer com o tempo
que nos é dado."
J. R. R. Tolkien

RESUMO

RIBEIRO, G. **A procura por variantes genéticas funcionais da eficiência alimentar em bovinos Nelore**. 72F. Dissertação (Mestrado). Faculdade de Zootecnia e Engenharia de Alimentos. Universidade de São Paulo. 2021.

A eficiência alimentar (EA) em sistemas de produção de proteína animal é uma característica de grande importância, já que animais mais eficientes refletem diretamente em melhores índices de produtividade e sustentabilidade das cadeias produtivas. No entanto, a avaliação desta característica é complexa, lenta e onerosa, o que justifica a busca de abordagens mais eficazes para identificar animais quanto à EA. Uma possível estratégia é a identificação de marcadores moleculares que permitam a seleção de animais com melhor ou pior EA. Abordagens baseadas em estudos de associação ampla do genoma (GWAS) tem gerado bastante informações, porém na maioria das vezes os marcadores encontrados não são os responsáveis pelos efeitos fenotípicos (causais), o que faz com que a análise dos resultados seja especulativa utilizando-se genes presentes em grandes janelas cromossomais. Outro problema destes estudos é que as análises são realizadas a partir do DNA de qualquer célula, se perdendo a possibilidade de compreender a importância funcional e tecido-específica das variantes causais em determinado fenótipo. Assim este trabalho se organiza em capítulos, onde no primeiro foi analisado dados gerados a partir de RNA-seq de amostras de biópsia de fígado de animais de alta e baixa EA, com o objetivo de identificar variantes funcionais que regulem genes e vias biológicas relacionadas com o fenótipo estudado. Foi possível identificar 256 variantes (247 SNPs e 9 INDELs (inserções e deleções)) distribuído em 190 genes que regulam vias importantes da imunidade inata, o que leva a concluir que o sistema imune é uma das principais vias que regulam a eficiência animal, seja direta ou indiretamente. Dentre os principais achados está 4 variantes homozigóticas para animais de alta eficiência alimentar (AEA) nos genes CFH e CFH5, que são responsáveis pela fagocitose de micróbios e células danificadas que induzem à infecção, alterações nessas variantes podem levar a alterações das vias biológicas relacionadas a imunidades, causando perdas econômicas. No segundo capítulo, realizou-se análises de dados a partir de RNA-seq de diversos orgãos (adrenal, hipotálamo, pituitária, fígado, musculo), para identificação de variantes

funcionais que podem ser reguladores centrais, suas interações biológicas com a característica de eficiência alimentar e por fim validar os resultados em uma população independente. Com esta pesquisa foi possível identificar 169 variantes expressos em comum nos cinco tecidos, sendo que essas variantes estão relacionadas principalmente com o MHC classe I. Também foram identificados um total de 144, 252, 413, 416 e 340 potenciais variantes funcionais (PVFs) apresentados no fígado, músculo, hipotálamo, hipófise e adrenal, que foram adjacentes a 223, 422, 694, 697 e 554 marcadores SNP apresentados no BovineHD (Illumina), respectivamente. Para realizar a validação e as previsões genômicas para o conjunto de validação, os marcadores SNP adjacentes ao PFV foram diferencialmente ponderados com base nos resultados obtidos com ssGWAS usando o conjunto de treinamento. Através dos dados obtidos e da metodologia aplicada, foi possível constatar que a precisão da predição aumentou de 31,03% para 40%, quando adicionados os dados das PVFs ponderados diferencialmente na matriz de predição. Através dos dados genômicos obtidos e utilizando a metodologia de identificação de PVFs e validação por ssGWAS, foi possível desenvolver um pipeline consistente que pode ser utilizada para a aprimorar os programas de melhoramento genético.

Palavras-Chave: RNA-seq, eficiência alimentar, variantes funcionais, multi-tecidos, sistema imunológico, GWAS, gado.

ABSTRACT

RIBEIRO, G. **The search for functional genetic variants of feed efficiency in Nellore cattle**. 72F. (Master) Dissertation. Faculty of Animal Science and Food Engineering. University of São Paulo. 2021.

Feed efficiency (FE) in animal protein production systems is a feature of great importance, since more efficient animals reflect directly on better productivity and sustainability indexes of the production chains. However, the evaluation of this characteristic is complex, slow and costly, which justifies the search for more effective approaches to identify animals regarding FE. A possible strategy is the identification of molecular markers that allow the selection of animals with better or worse FE. Approaches based on studies of broad genome association (GWAS) have generated a lot of information, however in most cases the markers found are not responsible for the phenotypic (causal) effects, which makes the analysis of the results speculative using genes present in large chromosomal windows. Another problem of these studies is that the analyzes are performed from the DNA of any cell, losing the possibility of understanding the functional and tissue-specific importance of the causal variants in a given phenotype. Thus, this work is organized in chapters, where the first one analyzed data generated from RNA-seq from liver biopsy samples from animals of high and low FE, with the objective of identifying functional variants that regulate genes and biological pathways related to the studied phenotype. It was possible to identify 256 variants (247 SNPs and 9 INDELs (insertions and deletions)) distributed in 190 genes that regulate important pathways of innate immunity, which leads to the conclusion that the immune system is one of the main pathways that regulate animal efficiency, be it directly or indirectly. Among the main findings are 4 homozygous variants for highly efficient animals (HFE) in the CFH and CFH5 genes, which are responsible for the phagocytosis of microbes and damaged cells that induce infection, changes in these variants can lead to changes in biological pathways related to immunities, causing economic losses. In the second chapter, data analysis was performed using RNA-seq from several organs (adrenal, hypothalamus, pituitary, liver, muscle), to identify functional variants that can be central regulators, their biological interactions with the efficiency characteristic and finally to validate the results in an independent population. With this research it was possible to identify 169 variants expressed in common in the

five tissues, and these variants are mainly related to MHC class I. A total of 144, 252, 413, 416 and 340 potential functional variants (PFVs) were also identified. in the liver, muscle, hypothalamus, pituitary and adrenal, which were adjacent to 223, 422, 694, 697 and 554 SNP markers presented in BovineHD (Illumina), respectively. To perform the validation and genomic predictions for the validation set, the SNP markers adjacent to the PFV were differentially weighted based on the results obtained with ssGWAS using the training set. Through the obtained data and the applied methodology, it was possible to verify that the precision of the prediction increased from 31.03% to 40%, when adding the data of the differently weighted PFVs in the prediction matrix. Through the genomic data obtained and using the methodology of identification of PFVs and validation by ssGWAS, it was possible to develop a consistent pipeline that can be used to improve the breeding programs.

Keywords: RNA-seq, feed efficiency, functional variants, multi-tissues, immune system, GWAS, cattle.

LIST OF FIGURES

Chapter 1

Chapter 2

LIST OF TABLES

Chapter 1

Chapter 2

LIST OF ABBREVIATIONS AND ACRONYMS

ADG	Average Daily Gain
ANPC	*Associação Nacional de Criadores e Pesquisadores*
APCs	Antigen-presenting cells
BIF	Beef Improvement Federation
BLUP	Best Linear Unbiased Prediction
CG	Contemporary Group
DMI	Dry Matter Intake
DP	Depth Plot
EBV	Estimated breeding values
eQTL	Expression Quantitative Trait Loci
FCR	Feed Conversion Ratio
FDR	False Discovery Rate
FE	Feed Efficiency
FindVar	Find Variables By Name
FS	Fisher Strand
GATK	Genome Analysis Toolkit
GBLUP	Genomic Best Linear Unbiased Prediction
GEBV	Vector of Genomic EBV
GWAS	Genome-Wide Association Studies
GWAS	Genome-Wide Association Studies
HFE	High Feed Efficiency
INDELs	Insertion and Deletions
LFE	Low Feed Efficiency
MAC	Complex Attack Complex
MAF	Minor Allele Frequency

MHC	Major Histocompatibility Complex
MQ	RMSMapping Quality
NCBI	Natural Center for Biotechnology Information
PEV	Position-effect variegation
PFVs	Potential Functional Variants
QD	Quality by Depth
QTL	Quantitative Trait Locus
QTN	Quantitative Trait Nucleotide
QUAL	Variant Quality Score
RCA	Regulation of the Complement Activation
rec	RFI records
RFI	Residual Feed Intake
RIG	Residual Intake and Weigh Gain
RIN	RNA Integrity Number
RNA-seq	RNA-sequencing
RWG	Residual Weigh Gain
SAMtools	Sequence Alignment/Map
SIFT	Scale-Invariant Feature Transform
SNP	Single Nucleotide Polymorphisms
ssBLUP	single step Best Linear Unbiased Prediction
ssGBLUP	single step Genomic Best Linear Unbiased Prediction
ssGWAS	single-step GWAS
STAR	Spliced Transcripts Alignment to a Reference
VEP	Variant Effect Predictor
wG	wighted G matrix

SUMMARY

1. Introduction

Ruminants can transform non-edible foods for humans, such as grasses, fodder and by-products rich in cellulose, into high-quality edible foods such as meat, milk, etc. This exclusive advantage of them results in low efficiency of feed conversion, being necessary to use approaches such as high-grain diets in feedlots to reach the appropriate weight and carcass scores for quality meat. However, to achieve the proper carcass, it is necessary to add grains to the animals' feed (concentrate), such as corn and soybeans, which are products that make up human food, generating, therefore, a competition for these food products. Another critical point is the selection of animals for feedlot, where individual feed efficiency is often not considered incurring additional expenses for the producer since animals with low feed conversion consume more dry matter (DM) to produce the same amount of products (meat or milk).

Another relevant point about feed efficiency (FE) is the fact that efficient animals generally produce less waste and gases responsible for the greenhouse effect, such as methane. Studies indicate that methane emissions are linearly related to DM consumption, that is, selecting efficient animals can reduce methane production by up to 28%, without affecting the animal's performance (Difford et al., 2020; Hegarty et al., 2007). Since efficient animals consume less DM to produce similar amounts of product (meat) compared to animals with low efficiency, animals with high feed efficiency (HFE) are more sustainable (Difford et al., 2020; Hegarty et al., 2007).

Although the selection of FE has several advantages, identify animals for FE is a complex task, since it is a phenotype regulated by several variables. There are two main metrics that can be used to identify efficient animals. The first consists of the relationship between the amount of feed consumed and weight gain, which is called feed conversion rate (FCR). This metric, however, is less used, as it has a strong negative correlation with body weight gain, resulting in the selection of undesirable animals in terms of the animal's body size and feed consumption. The second metric, called Residual Feed Intake (RFI), consists of the difference between the actual feed consumption and the expected feed requirements for maintenance and weight gain. When an animal has low RFI, it is considered more efficient, as it tends to consume lower amounts of feed. This measure became common for the selection of FE due to the

lack of phenotypic correlation with the body weight gain and the size of the animal, being independent of the animal's performance.

Studies indicate that the variation of the RFI is caused by basic metabolic processes, which determine the efficiency in production, however, this measure is still not well accepted by the industry, as animals classified as superior in terms of RFI can develop slowly. In addition, performing RFI estimatives are expensive and not an easy task, since all feed consumption measures used in beef cattle have complex interactions with the environment and biological processes (digestion, metabolism, physical activity, thermoregulation and food intake). Additionally, FE is a polygenic characteristic that makes it even more difficult to select animals with HFE (KENNY et al., 2018).

Biological variations from animal to animal in feed efficiency are an important factor, but they are far from being understood and further studies are needed to identify the undesirable side effects of this selection. By better understanding the biological processes linked to FE, it makes it possible to plan more balanced breeding programs, reveal new management strategies and adapt the formulation of the feed to the individual needs of the animal, as well as contribute to the discovery of cheap and quick methods (markers) for classify early (young) animals with high feed efficiency, without the need to measure feed consumption.

Studies based on phenomics and systems biology have been intensified in recent years and most of them aimed to identify the complex interation and the molecular bases that explain why similar animals reared under the same conditions differ in feed efficiency. These studies have shown as a result regulatory genes, biological pathways and differentially expressed genes that can influence the phenotypic characteristic. However, they still did not identify the causal variants, responsible to control all these genes and their interactions.

Thus, the objective of this work to identify potential functional variants, genes and biological pathways that regulate feed efficiency through RNA-seq data from Nelore cattle.

2. **Chapter 1:** Potential Functional Variants in Innate Immune Response Genes Associated with Feed Efficiency in Beef Cattle

Submmited to Genetics Selection Evolution in June 2020.

Gabriela Ribeiro[1], Aline Silva Mello Cesar[2], Pâmela Almeida Alexandre[1,3], José Bento Sterman Ferraz[1], Heidge Fukumasu[2]

[a] Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, São Paulo, Brazil.

[b] Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, São Paulo, Brazil.

[c] Present address: CSIRO Agriculture & Food, 306 Carmody Rd., St. Lucia, Brisbane, QLD 4067, Australia

Abstract

Identifying and selecting animals for feed efficiency (FE) is extremely important for the beef production chain. Currently, the most common parameter to access the FE animals is residual feed intake (RFI), which is the residual of the linear regression that estimates DMI based on average daily gain and mid-test metabolic body weight. However, it relies on costly and time-consuming data collection, creating a growing demand for alternative approaches to identify genetically superior animals for FE. This study aimed to detect potential liver-specific functional variants from RNA-seq data of 16 Nellore bulls (*Bos indicus*) divergently selected for FE. The variant call analysis detected 247 missense SNPs and nine insertion-deletions (INDELs) that alter the protein functions. These variants were found within 190 genes differentially found ($P < 0.05$) in liver tissue between high FE (HFE) and low FE (LFE) animals. To better understand the role of these variants in biological pathways, we performed a functional enrichment analysis, which highlighted six genes involved in complement cascade and cascade complement regulation pathways, and 20 genes involved in the regulation of the innate immune system. They had four different significant variants in the complement factor H (*CFH*)

family genes, and all were homozygous in HFE animals rather than some degree of heterozygous in LFE animals. We developed a pipeline to detect potential liver-specific functional variants from RNA-seq data from animals divergently selected for feed efficiency. With this approach, we found potential functional variants in innate immune response genes associated with feed efficiency in beef cattle.

Keywords: SNP, INDEL, transcriptomic, variant calling, liver, bovine.

2.1. Introduction

Studies on feed efficiency (FE) in beef cattle (*Bos indicus*) are very important. The identification and selection of efficient animals can improve the productivity by reducing feed costs, which can reach 75% in feedlot systems (Paper, 2010). FE selection is also justified by the demand for less environmental impact of livestock. Feed efficient animals are recognized as more sustainable due to decreased production of waste and greenhouse gases.

One of the most common ways to evaluate FE is by residual feed intake (RFI), an independent measure of the level size and growth rate in beef cattle (Koch et al., 1963). At least five major physiological processes contribute to RFI as intake of feed, digestion of feed, metabolism, physical activity and thermoregulation (Arthur and Herd, 2008). Thus, the liver is considered a central organ for FE since it is responsible for the metabolism of nutrients as proteins, lipids and carbohydrates, along with other functions as metabolism of bilirubin, bile acids, xenobiotics, protein synthesis and immunity (Stalker and Hayes, 2007). Our group was the first to described a pathophysiological mechanism associated with FE in beef cattle: liver inflammation due to altered metabolism and/or bacterial translocation/infection (Alexandre et al., 2015), which was in part corroborated by others (Paradis et al., 2015; Tizioto et al., 2015; Weber et al., 2016). We also unravel the metabolic pathways related to FE in Nellore cattle showing an increased bacterial load in low feed efficient animals (LFE) which is in part responsible for the hepatic lesions and inflammation in these animals (Fonseca et al., 2019).

The RFI estimation in beef cattle is costly and time-consuming, therefore other approaches to identify genetically superior animals are needed. A usual strategy is the use of molecular markers for genomic selection based on SNPs (single nucleotide polymorphisms) genotyping, which allows the detection of DNA variants associated with FE. However, the majority of the SNPs previously identified in Genome-Wide Association studies (GWAs) and expression quantitative loci (eQTL) association study (Cesar et al., 2018) are non-causal variants (de Oliveira et al., 2014; Saatchi et al., 2014; Santana et al., 2014), which are located in intron or intergenic genomic regions. There is an urgent need to establish the functional (causal) variants of this important phenotype. One possibility is the analysis of whole genome sequencing, an approach which is still very costly. Another possibility is the whole exome sequencing that has a reduced cost, but still lack information regarding the importance of the polymorphism within cell and/or tissue architecture. Therefore, we propose an approach to overcome these limitations based on the identification of genetic variants from RNA-sequencing (RNA-seq) data from a physiologically phenotype-related organ, followed by a classification of the potential functional variants according to their effects on protein expression and function.

In this work, we developed a pipeline to detect potential liver-specific functional variants from RNA-seq data from animals divergently selected for feed efficiency. With this approach, we found potential functional variants in innate immune response genes associated with feed efficiency in beef cattle.

2.2. Methods

2.2.1. Phenotypic data and biological sample collection

All animal protocols were approved by the Institutional Animal Care and Use Committee of Faculty of Food Engineering and Animal Sciences, University of São Paulo (FZEA-USP – protocol number 14.1.636.74.1).

All procedures to collect the phenotypes and biological samples were carried out at FZEA-USP, Pirassununga, State of São Paulo, Brazil. Ninety-eight Nellore bulls (*Bos indicus*) (16 to 20 months old and 376 ± 29 kg BW) were evaluated in a feeding trial comprised of 21 days of adaptation to feedlot diet followed by a 70-day period of data collection. Total mixed ration was offered *ad libitum* and daily dry matter intake

(DMI) was individually measured. Animals were weighted at the beginning, at the end and every two weeks during the experimental period. Feed efficiency (FE) was estimated by residual feed intake (RFI) (Koch et al., 1963). At the end of the experimental period, liver biopsy samples from the left hepatic lobes were collected from each animal and quickly frozen in liquid nitrogen and stored at -80 °C. Further information about management and phenotypic measures of the animals used in this study can be found elsewhere (Alexandre et al., 2015).

2.2.2. RNA-seq data

Samples of 8 animals from each FE group (high and low) were selected for RNA-seq using RFI measure extremes. The total RNA from liver samples was extracted by using the RNeasy mini kit (QIAGEN, Crawley, West Sussex, UK) according to the instructions provided by the manufacturer. The total RNA quality and quantity were assessed using automated capillary gel electrophoresis on a Bioanalyzer 2100 with RNA 6000 Nano Labchips according to the manufacturer's instructions (Agilent Technologies Ireland, Dublin, Ireland). Samples that presented RNA integrity number (RIN) less than 8.0 were discarded. The mRNA libraries were constructed using the TruSeq™ Stranded mRNA LT Sample Prep Protocol and sequenced on Illumina HiSeq 2500 equipment in a HiSeq Flow Cell v4 using HiSeq SBS Kit v4 (2x100pb). FastQC software (Babraham Institute, Cambridge, UK, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to obtain the sequencing quality. Removal of PolyA / T tails and adapters was performed using Seqyclean software (University of Idaho: Institute of Bioinformatics and Evolutionary Studies, Moscow, USA, https://bitbucket.org/izhbannikov/seqyclean), so only bases with quality ≥ 20 and complete readings with at least 50 bp, were further studied for further analysis. The alignment of the reads was done using the STAR software version 2.7 (Dobin et al., 2013) with the reference genome *Bos taurus* UMD3.1 (Ensembl, ftp://ftp.ensembl.org/pub/release-98/fasta/bos_taurus/dna/), allowing two mismatches per read.

2.2.3. Call of functional genomic variations associated with FE

GATK software (McKenna et al., 2010) was used to perform the call the variants analysis. HaplotypeCaller command was applied to identify the variants (SNPs and insertions and deletions (INDELs). The variants underwent quality control on the GATK software as follows: (Variant Quality Score - QUAL) > 30, depth of sequencing (Deph Plot - DP) > 4, amount of available coverage (QualByDepth - QD) > 3, polarization trend (FisherStrand - FS) > 30 and the general mapping quality of readings that support a variant call (RMSMappingQuality- MQ) < 35. Statistical analysis was performed by Plink software (Purcell et al., 2007) considering MAF < 40% and call rate equal to 50%. For the allele frequency test between HFE and LFE groups, the Cochran-Armitage trend analysis was used considering significant differences when $P<0.05$.

2.2.4. Characterization of the effects of variants on protein sequence and function

The potential functional variants were analyzed in the Variant Effect Predictor (VEP) online tool (McLaren et al., 2016) which predicts the functional effects of the variants. Potential functional variants were analyzed by the Scale-Invariant Feature Transform (SIFT) score, a statistical tool of the VEP online software. SIFT is an algorithm that predicts whether an amino acid substitution affects the function of the protein. This analysis was performed on the basis of the gene sequence homology and the physical properties of the amino acids, where the like sequences are first sought, then strictly related sequences (which may share functions similar to the query sequence), with the alignment of the sequences chosen is possible to calculate the normalized probability for all possible substitutions of the alignment. Thus, generating a SIFT score ranging from 0 to 1, being classified as deleterious values below 0.05 and above 0.05 is considered as tolerated. The deleterious classification indicates that the amino acid change will have a great impact on the proteins, since the tolerated classification indicates that the impact will not be so great. With this information it was possible to compare the position of the variants in the protein with the protein database of the National Center for Biotechnology (NCBI), in order to find information about the name of the region and what function that region performs.

2.2.5. Functional enrichment analysis

The functional enrichment analysis was performed with Panther version 14.1 (Thomas et al., 2003) to identify biological pathways over-represented in the set of genes with potential functional variants. A multiple test correction was used and significant pathways were considered when $P<0.05$.

2.3. Results

2.3.1. Characterization of feed efficiency groups

Two groups of eight animals with extreme values of feed efficiency were selected and named High Feed Efficient (HFE, lowest RFI) and Low Feed Efficient (LFE, highest RFI). These groups were significantly different for feed efficiency traits RFI, feed conversion ratio (FCR), residual weight gain (RWG), residual intake and weight gain (RIG), for dry matter intake (DMI) and average daily gain (ADG) (Table 1). The LFE animals presented higher backfat thickness at the end of the experiment ($P<0.05$), supporting that HFE animals are more feed efficient because they eat less, have similar ADG and are leaner than LFE animals (Alexandre et al., 2015; Novais et al., 2019). A total of 11,361 genes were expressed in liver samples from HFE and LFE animals and eight genes were differentially expressed (DE) ($P \leq 0.1$): NR0B2, SOD3, RHOB, Bta-mir-2904-2, FTL, CYP2E1, GADD45G and FASN (Alexandre et al., 2015).

Table 1. Phenotypic traits. Mean of the groups of high feed efficiency ($n = 8$) and low feed efficiency ($n = 8$).

Trait	HFE mean	LFE mean	*P*-value
BWi (kg)	413.20 ± 47.20	407.80 ± 27.05	0.78
BWf (kg)	562.20 ± 49.50	530 ± 30.76	0.10
ADG(kg/d)	2.13 ± 0.52	1.75± 0.23	0.08
DMI (kg/d)	10.20 ± 1.26	12.32 ± 0.95	2.09x10-3*
FCR	4.96 ± 0.80	7.18 ± 0.62	2.82x10-5*
RFI (kg/d)	−1.43 ± 0.33	1.66 ± 0.41	9.39x10-4*
RWG (kg/d)	0.38 ± 0.30	−0.38 ± 0.14	7.98x10-5*
RIG	1.80 ± 0.31	−2.03 ± 0.37	9.39x10-4*
REAi	67.45±5.38	65.84 ±4.08	0.5115
REAf	83.38±7.44	82.79±4.54	0.8521
REAg	15.93±10.75	16.95±5.02	0.812
BFTi	1.087±1.19	1.97±1.32	0.1795
BFTf	3.00±1.88	5.96±1.52	0.004*
BFTg	1.91±2.06	3.99±1.15	0.03*
RFTi	2.49±1.40	5.12±1.26	0.001*
RFTf	5.69±2.58	9.50±2.05	0.006*
RFTg	3.20±1.95	4.37±1.70	0.2206

HFE = high feed efficiency; LFE = low feed efficiency; BWi = initial body weight; BWf = final body weight; DMI = dry matter intake; ADG = average daily gain; FCR = feed conversion ratio; RFI = residual feed intake; RWG = residual body weight gain; RIG = residual intake and body weight gain; REAi = initial rib eye area; REAf = final rib eye area; REAg = gain of rib eye area; BFTi = initial back fat thickness; BFTf = final back fat thickness; BFTg = gain of back fat thickness; RFTi = initial rump fat thickness; RFTf = final rump fat thickness; RFTg = gain of rump fat thickness

*$P \leq 0.05$

♦Student's t-test

°Mann–Whitney-Wilcoxon test

Adapted from Alexandre et al. 2015

2.3.2. Characterization of expressed SNPs and INDELs associated with feed efficiency

The variant calling analysis detected 268,393 and 37,587 SNPs and INDELs, respectively, from the 16 samples. These variants were tested for Minor Allele Frequency (MAF)<0.4 and call rate=50%, which left 68,852 SNPs and 5,148 INDELs, respectively, for statistical analysis. From these, 2,149 SNPs and 139 INDELs were significantly different between HFE and LFE animals (P<0.05). Part of these variants were found in exonic regions (51% for SNPs and 29% for INDELs) but a considerable fraction was found outside the gene (43% for SNPs and 66% for INDELs) or in introns (6% for SNPs and 5% for INDELs, Fig. 1).

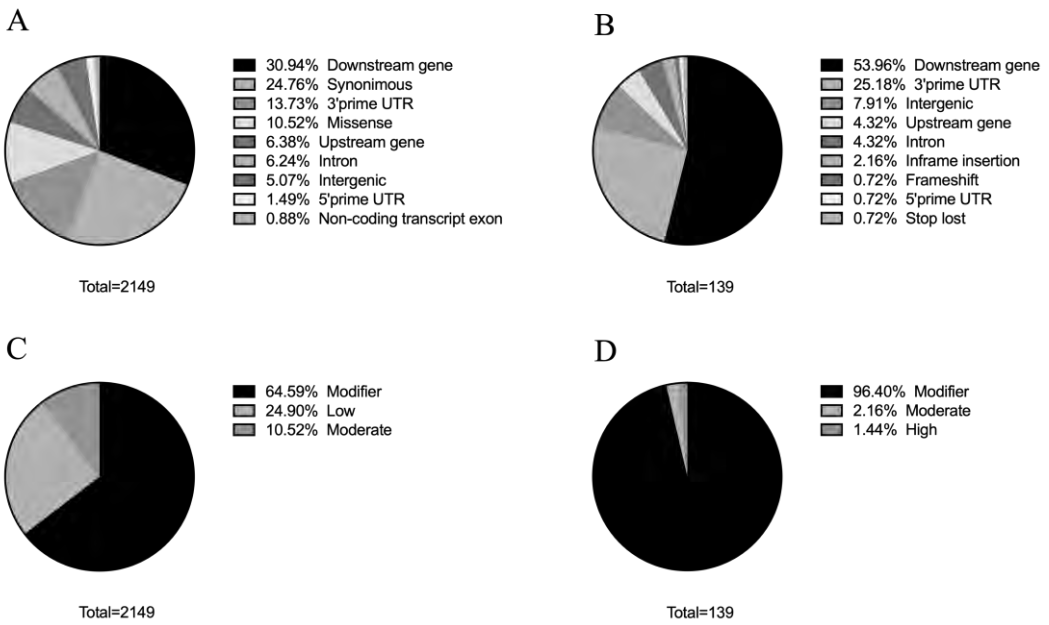


Figure 1. Relative frequency of significant SNPs and INDELs between HFE and LFE groups classified by consequence and impact on protein expression. SNPs and INDELs classified by the consequence on gene sequence (A and B) and by the impact on protein expression (C and D).

We considered the variants with moderate and high impact as potential functional variants totalizing 256 variants (247 SNPs and 9 INDELs) selected for further analysis. They generated effects on proteins as follows: shorter proteins due to frameshift INDELs, bigger proteins due to insertion of one or two AA by inframe insertion INDELs, changes in single amino acid (AA) on proteins due to missense SNPs and a stop retained INDEL (able to change at least one base of the terminator codon but maintaining the function of the terminator). These 256 potential functional variants were distributed across 190 genes, as some variants were found in spliced isoforms and there were genes with more than one variant (Supplementary table 1).

2.3.3. Depicting the biology of potential functional variants

In order to understand the importance of these 256 potential functional variants we performed the functional enrichment analysis for the selected genes, which detected three different significant pathways (False Discovery Rate - FDR<0.10): "Regulation of

Complement cascade", "Complement cascade" and "Innate Immune System" with fold enrichments of 14.79x, 12.19x and 2.57x respectively (Table 2).

Table 2. Significant pathways detected by functional enrichment analysis of genes with potential functional variants.

Reactome pathways	*Bos taurus* REFLIST	Input genes	Fold Enrichment	P-value	Q-value
Regulation of complement cascade	47	6	14.79	$5.97x10^{-06}$	$9.60x10^{-03}$
Complement cascade	57	6	12.19	$1.65x10^{-05}$	$1.32x10^{-02}$
Innate immune system	901	20	2.57	$1.31x10^{-04}$	$6.99x10^{-02}$

The three pathways associated with FE by our approach were all related to the innate immune system, in particular with the regulation of the complement cascade. The first two pathways enriched from the same set of six genes (Supplementary table 2) and the Innate Immune System enriched from twenty genes (Supplementary table 2), including the six genes from the complement cascade/regulation of complement cascade pathways. The list of these 20 genes, their potential functional variants, the impact on protein sequence and the frequency in each group can be found in the Supplementary table 3. Considering the “regulation of complement cascade” and “complement cascade” as one pathway since they were enriched for the same set of genes, it calls the attention the overrepresentation of Complement H factor genes (complement factor H-related 5, complement factor H and complement factor H precursor). There were four different significant variants in these genes and all were homozygous in HFE animals instead of some degree of heterozygosity in LFE animals (Table 4).

Table 4. Frequency and missense alteration of the potential functional variants enriched from the complement cascade pathway associated with feed efficiency.

Gene ID	Gene	SNP	Position protein	Protein alteration	LFE	HFE	P-value
ENSBTAG00000023177	*CFH*	T/A	445	I/K	2|6	0|16	0.028
ENSBTAG00000038171	*CFHR5*	T/C	558	P/S	3|11	0|16	0.038
ENSBTAG00000039995	*CFH*	G/C	15	P/R	4|6	0|8	0.016
ENSBTAG00000039995	*CFH*	C/T	30	Y/H	4|8	0|8	0.035

I = isoleucine; K = lysine; P = proline; S = serine; R = arginine; Y = tyrosine; H = histidine; LFE = Low Feed Efficiency; HFE = High Feed Efficiency

2.3.4. Effects of other important potential causal variants

At last, we analyzed the effects of high impact INDELs on protein function since they altered the protein sequence considerably but were not enriched in the processes already described. The frameshift INDEL on *GAS2L1* caused a deletion of 7 AA in the calponin homology domain and had an allele frequency of 35% (5/14) in the HFE animals whereas it was not found in LFE group (0/8, $P<0.05$). Another frameshift INDEL generated a deletion of 7 AAs in the ATP binding cassette (ABC) domain of the TAP2 and had an allele frequency of 50% (7/14) in the HFE group whereas only 25% in the LFE group (3/12, $P<0.05$). Interestingly, this INDEL was found significantly associated not only in one but also in two spliced isoforms of *TAP2*. An INDEL in the transcription factor *RREB1* caused an insertion of 2 AAs in one of the C2H2 Zinc finger domains of *Bos indicus RREB1* and was associated with LFE since the allele frequency in this group was 30% (3/10) but in the HFE no allele was found (0/14, $P<0.05$). An INDEL was found in *SEMA4F*, which generated a premature stop codon reducing the COOH-terminal of the SEMA4F in 5 AAs. This region contains the PDZ domain of the protein responsible for anchoring the SEMA4F in the membrane to cytoskeletal components. The allele with the deletion of 40 bp was found in 50% in LFE group (5/10) and 16% in HFE group (2/12, $P<0.05$). An inframe insertion INDEL in the

ECSIT insert a glutamic acid in the position 423 at the COOH terminal of ECSIT and this variant is associated with feed efficiency as the allele frequency was 29% (4/14) in the HFE and 0% (0/12) in the LFE group ($P<0.05$).

Another inframe insertion INDEL was found in the ENSBTAG00000011926 gene, which is probably the transcription factor *ZFN665-like* that inserted a leucine in one C2H2 Zn finger region of the protein. The variant with the inserted leucine is associated with lower feed efficiency with an allele frequency of 38% (3/8) instead of 6% (1/15) in the HFE group ($P>0.05$). Another potential functional variant is the insertion of one Alanine (252) just one position from the phosphorylation serine site of RPP30 (S251) with a higher frequency in the LFE group (LFE=50%, 5/10; HFE=14%, 2/14; $P<0.05$). The INDEL in the *MGAT2* although considered a frameshift INDEL, it did not change the function since it generated a stop retained codon.

2.4. Discussion

There is increasing evidence for the importance of the immune system on feed efficiency in a variety of domestic species. Our group previously showed that feed efficiency is associated with altered inflammatory response in beef cattle partially because of bacterial translocation from the digestive tract to the liver (Alexandre et al., 2015; Fonseca et al., 2019). Here we reported for the first time the existence of potential functional variants in immune system-related genes associated with feed efficiency, with the potential to alter the function of the innate immune system by regulation of the complement cascade. To achieve these results, we performed a screening for potential functional variants from transcriptomic data of animals evaluated for feed efficiency using RNA-seq data and bioinformatic tools such as GATK, VEP, functional enrichment using Panther and evaluation of effects from the potential variants.

Our pipeline found potential functional variants in genes that significantly enriched for three biological pathways related to the Immune System: “Regulation of Complement Cascade”, “Complement Cascade” and “Innate Immune System”. Similar genes enriched for Regulation of Complement Cascade and Complement Cascade (*C8G, CFHR5, CFH, IGHG, CPN2*) and it calls the attention the overrepresentation of potential functional variants in Complement H factor genes (*CFHR5*, *CFH* and *CFH*

precursor). The complement system is one of the major biological processes of the innate immune system related to the ability of antibodies and phagocytic cells to clear microbes and damage cells inducing inflammation. The system consists of a number of small proteins mainly synthesized by hepatocytes, but can also be secreted by endothelial cells, white blood cells and epithelial cells (Peng et al., 2008; Strainic et al., 2008). This enzymatic cascade helps in the defense of infections and is one of the main effectors of humoral immunity, regulating various biological processes such as phagocytosis, opsonization, leukocyte chemotaxis, release of mast cells, basophils and active oxygen species by leukocytes, vasoconstriction, smooth muscle contraction, increased vessel permeability, platelet aggregation and cytolysis (Frank and Fries, 1991; Haeney, 1998). Activation of this cascade can be performed by 3 pathways (classical, lectin and alternative) and the genes found in our study regulate the alternative pathway, which is triggered in the presence of an exogenous activator, such as the presence of fungi, bacteria, some types of viruses and parasites, which activate C3 molecules and trigger the cascade (Frank, 1989; Iturry-Yamamoto and Portinho, 2001; Ochs et al., 1983). As already mentioned, we showed in previous works that less feed efficient animals have increased inflammatory response in the liver (Alexandre et al., 2015) and there is a higher level of endotoxins in the blood of LFE animals (Fonseca et al., 2019) which corroborates with the activation of the complement cascade by the alternative pathway. In fact, another study comparing the hepatic transcriptome of high and low FE animals found the complement system as the most significant canonical pathway enriched from the differential expressed genes (Tizioto et al., 2015).

The proteins responsible for regulation of the complement activation (RCA) can be divided into two main groups: membrane-bound regulators and soluble regulators (Jiang et al., 2015). Factor H Complement belongs to the RCA family and we found different alleles predominantly in LFE animals. The *CFH* also has five additional members, represented by five separate CFH-related genes (*CFHR*), ranging from 1 to 5. The proteins produced by *CFHR* contain a set of domains that are homologous to those of *CFH* (Zipfel et al., 2002). These proteins act at the C3 level (Józsi and Zipfel, 2008) negatively regulating complement activation, acting as a cofactor for factor I-mediated C3b cleavage and facilitating C3 convertase acceleration of deterioration (Zipfel et al., 2006, 2002). In this work, we found heterozygosity in the LFE animals for the *CFH* genes whereas the HFE animals were homozygous. Another gene that influences the

complement system in all pathways (classical, lectin and alternative) is the C8 gene found in heterozygosity only in LFE animals. The C8 protein is part of the complex membrane attack complex (MAC) that assembles on bacterial membranes to form a pore, allowing the disruption of bacterial membrane organization, cell death and consequently inflammatory response (Morgan, 2016; Oikonomopoulou et al., 2012).

The genetic modulation of the innate immune system supports our previous studies that demonstrate the relationship between feed efficiency and hepatic inflammation (Alexandre et al., 2015; Fonseca et al., 2019). Along with the genes related to complement cascade regulation, other genes with potential functional variations enriched for the innate immune system as *BOLA*, *ECSIT*, *GLB1*, *ADA2*, *RIPK3*, *LILRA4*, *SIRPB1* and *YEAR6*. Olivieri and colleagues found candidate genes associated with feed efficiency traits in Nellore cattle involved in immune system as well as other processes (Olivieri et al., 2016). Specifically, they proposed the *NLRP14*, which was shown to regulate the innate immune signaling. In humans, a germline mutation in *NLRP14* impairs its function and affects the innate immune signaling (Abe et al., 2017) as well as other polymorphisms in complement genes are linked to human diseases as hemolytic uremia and age-dependent macular degeneration (Degn et al., 2011), supporting the idea that potential functional variations affect the innate immune system. Although we haven't found the same candidate genes from the GWAS study of Olivieri and colleagues (2016), our approach confirmed the innate immune system as an important pathway for feed efficiency in beef cattle.

Our initial idea was to identify causal genetic variants for feed efficiency and found these to be present exclusively in one group (LFE or HFE). However, in our study none of the 256 potential functional variants affecting protein sequences were found exclusively in one group and that is the reason why we choose not to call these as causal variants. It is shown for complex traits that even the most important loci in the genome have small effects sizes, and that, together, the significant hits only explain a modest fraction of the predicted genetic variance (Boyle et al., 2017). Indeed, in one of the largest performed GWAS study for feed efficiency in beef cattle, the authors found ten significant quantitative trait locus (QTL) for RFI but none explained more than 2.5% of the additive genetic variance in any population (Saatchi et al., 2014).

We are aware that one possible limitation of the present work is the sample size

but even in the present condition the results are well supported by the literature in transcriptomic (Alexandre et al., 2015; Paradis et al., 2015; Tizioto et al., 2015) and GWAs studies (Olivieri et al., 2016), including other species as pigs (Horodyska et al., 2019; Ramayo-Caldas et al., 2018) and poultry (Zhou et al., 2015). Therefore, the existence of potential functional variations in genes of the innate immune response affecting feed efficiency in beef cattle seems plausible and worth future studies.

2.5. Conclusion

In this research, we found possible hepatic causal variants that modulate the expression/function of genes involved in the pathways innate immune response. Having as one of the main findings, four applicable homozygous variants for HFE animals in the *CFH* and *CFH5* genes, which are responsible for phagocytosis of microbes and damaged cells that induce infection. Occasional changes in these variants regulate the alteration of immunity-related biological pathways, may alter or modify the model, which in our case is economical.

2.6. References

Abe, T., Lee, A., Sitharam, R., Kesner, J., Rabadan, R., Shapira, S.D., 2017. Germ-Cell-Specific Inflammasome Component NLRP14 Negatively Regulates Cytosolic Nucleic Acid Sensing to Promote Fertilization. Immunity 46, 621–634. https://doi.org/10.1016/j.immuni.2017.03.020

Alexandre, P.A., Kogelman, L.J.A., Santana, M.H.A., Passarelli, D., Pulz, L.H., Fantinato-Neto, P., Silva, P.L., Leme, P.R., Strefezzi, R.F., Coutinho, L.L., Ferraz, J.B.S., Eler, J.P., Kadarmideen, H.N., Fukumasu, H., 2015. Liver transcriptomic networks reveal main biological processes associated with feed efficiency in beef cattle. BMC Genomics 16. https://doi.org/10.1186/s12864-015-2292-8

Arthur, J.P.F., Herd, R.M., 2008. Residual feed intake in beef cattle. Rev. Bras. Zootec. 37, 269–279. https://doi.org/10.1590/S1516-35982008001300031

Boyle, E.A., Li, Y.I., Pritchard, J.K., 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. https://doi.org/10.1016/j.cell.2017.05.038

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., Fritz-Waters, E., Kramer, L., Garrick, D., Beiki, H., Geistlinger, L., Mourão, G.B., Zerlotini, A., Coutinho, L.L., 2018. Identification of putative regulatory regions and transcription factors associated with intramuscular fat content traits. BMC Genomics 19, 499. https://doi.org/10.1186/s12864-018-4871-y

de Oliveira, P.S.N., Cesar, A.S.M., do Nascimento, M.L., Chaves, A.S., Tizioto, P.C., Tullio, R.R., Lanna, D.P.D., Rosa, A.N., Sonstegard, T.S., Mourao, G.B., Reecy, J.M., Garrick, D.J., Mudadu, M.A., Coutinho, L.L., Regitano, L.C.A., 2014. Identification of genomic regions associated with feed efficiency in Nelore cattle. BMC Genet. https://doi.org/10.1186/s12863-014-0100-0

Degn, S.E., Jensenius, J.C., Thiel, S., 2011. Disease-causing mutations in genes of the complement system. Am. J. Hum. Genet. https://doi.org/10.1016/j.ajhg.2011.05.011

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Fonseca, L.D., Eler, J.P., Pereira, M.A., Rosa, A.F., Alexandre, P.A., Moncau, C.T., Salvato, F., Rosa-Fernandes, L., Palmisano, G., Ferraz, J.B.S., Fukumasu, H., 2019. Liver proteomics unravel the metabolic pathways related to Feed Efficiency in beef cattle. Sci. Rep. 9, 5364. https://doi.org/10.1038/s41598-019-41813-x

Frank, M.M., 1989. Complement: A brief review. J. Allergy Clin. Immunol. 84, 411–420. https://doi.org/https://doi.org/10.1016/0091-6749(89)90350-3

Frank, M.M., Fries, L.F., 1991. The role of complement in inflammation and phagocytosis. Immunol. Today 12, 322–326. https://doi.org/https://doi.org/10.1016/0167-5699(91)90009-I

Haeney, M.R., 1998. The role of the complement cascade in sepsis. J. Antimicrob. Chemother. 41, 41–46. https://doi.org/10.1093/jac/41.suppl_1.41

Horodyska, J., Hamill, R.M., Reyer, H., Trakooljul, N., Lawlor, P.G., McCormack, U.M., Wimmers, K., 2019. RNA-Seq of Liver From Pigs Divergent in Feed Efficiency Highlights Shifts in Macronutrient Metabolism, Hepatic Growth and Immune Response. Front. Genet. 10, 117. https://doi.org/10.3389/fgene.2019.00117

Iturry-Yamamoto, G.R., Portinho, C.P., 2001. Sistema complemento: ativação, regulação e deficiências congênitas e adquiridas. Rev. Assoc. Med. Bras. 47, 41–51. https://doi.org/10.1590/s0104-42302001000100029

Jiang, C., Zhang, J., Yao, J., Liu, S., Li, Y., Song, L., Li, C., Wang, X., Liu, Z., 2015. Complement regulatory protein genes in channel catfish and their involvement in

disease defense response. Dev. Comp. Immunol. 53, 33–41. https://doi.org/https://doi.org/10.1016/j.dci.2015.06.002

Józsi, M., Zipfel, P.F., 2008. Factor H family proteins and human diseases. Trends Immunol. 29, 380–387. https://doi.org/https://doi.org/10.1016/j.it.2008.04.008

Kenny, D. A., Fitzsimons, C., Waters, S. M., McGee, M. 2018. Invited review: Improving feed efficiency pf beef cattle - The current state of the art and future challenges. Animal. 12. 1825-1826. http://doi.org/10.1017/S1751731118000976

Koch, R.M., Swiger, L.A., Chambers, D., Gregory, K.E., 1963. Efficiency of Feed Use in Beef Cattle. J. Anim. Sci. 22, 486–494.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., DePristo, M.A., 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303. https://doi.org/10.1101/gr.107524.110

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., Cunningham, F., 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17, 122. https://doi.org/10.1186/s13059-016-0974-4

Morgan, B.P., 2016. The membrane attack complex as an inflammatory trigger. Immunobiology 221, 747–751. https://doi.org/https://doi.org/10.1016/j.imbio.2015.04.006

Novais, F.J., Pires, P.R.L., Alexandre, P.A., Dromms, R.A., Iglesias, A.H., Ferraz, J.B.S., Styczynski, M.P.-W., Fukumasu, H., 2019. Identification of a metabolomic signature associated with feed efficiency in beef cattle. BMC Genomics 20, 8. https://doi.org/10.1186/s12864-018-5406-2

Ochs, H.D., Wedgwood, R.J., Frank, M.M., Heller, S.R., Hosea, S.W., 1983. The role of complement in the induction of antibody responses. Clin. Exp. Immunol. 53, 208–216.

Oikonomopoulou, K., Ricklin, D., Ward, P.A., Lambris, J.D., 2012. Interactions between coagulation and complement--their role in inflammation. Semin. Immunopathol. 34, 151–165. https://doi.org/10.1007/s00281-011-0280-x

Olivieri, B.F., Mercadante, M.E.Z., Cyrillo, J.N.D.S.G., Branco, R.H., Bonilha, S.F.M., De Albuquerque, L.G., De Oliveira Silva, R.M., Baldi, F., 2016. Genomic regions associated with feed efficiency indicator traits in an experimental nellore cattle population. PLoS One. https://doi.org/10.1371/journal.pone.0164390

Paper, R., 2010. Simulation Modelling of the Cost of Producing and Utilising Feeds for Ruminants 14.

Paradis, F., Yue, S., Grant, J.R., Stothard, P., Basarab, J.A., Fitzsimmons, C., 2015. Transcriptomic analysis by RNA sequencing reveals that hepatic interferon-induced genes may be associated with feed efficiency in beef heifers1. J. Anim. Sci. 93, 3331–3341.

Peng, Q., Li, K., Anderson, K., Farrar, C.A., Lu, B., Smith, R.A.G., Sacks, S.H., Zhou, W., 2008. Local production and activation of complement up-regulates the allostimulatory function of dendritic cells through C3a–C3aR interaction. Blood 111, 2452 LP – 2461. https://doi.org/10.1182/blood-2007-06-095018

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., Sham, P.C., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575. https://doi.org/10.1086/519795

Ramayo-Caldas, Y., Ballester, M., Sánchez, J.P., González-Rodríguez, O., Revilla, M., Reyer, H., Wimmers, K., Torrallardona, D., Quintanilla, R., 2018. Integrative approach using liver and duodenum RNA-Seq data identifies candidate genes and pathways associated with feed efficiency in pigs. Sci. Rep. 8, 558. https://doi.org/10.1038/s41598-017-19072-5

Saatchi, M., Beever, J.E., Decker, J.E., Faulkner, D.B., Freetly, H.C., Hansen, S.L., Yampara-Iquise, H., Johnson, K.A., Kachman, S.D., Kerley, M.S., Kim, J.W., Loy, D.D., Marques, E., Neibergs, H.L., Pollak, E.J., Schnabel, R.D., Seabury, C.M., Shike, D.W., Snelling, W.M., Spangler, M.L., Weaber, R.L., Garrick, D.J., Taylor, J.F., 2014. QTLs associated with dry matter intake, metabolic mid-test weight, growth and feed efficiency have little overlap across 4 beef cattle studies. BMC Genomics 15. https://doi.org/10.1186/1471-2164-15-1004

Santana, M.H.A., Utsunomiya, Y.T., Neves, H.H.R., Gomes, R.C., Garcia, J.F., Fukumasu, H., Silva, S.L., Junior, G.A.O., Alexandre, P.A., Leme, P.R., Brassaloti, R.A., Coutinho, L.L., Lopes, T.G., Meirelles, F. V, Eler, J.P., Ferraz, J.B.S., Oliveira Junior, G.A., Alexandre, P.A., Leme, P.R., Brassaloti, R.A., Coutinho, L.L., Lopes, T.G., Meirelles, F. V, Eler, J.P., Ferraz, J.B.S., 2014. Genome-wide association analysis of feed intake and residual feed intake in Nellore cattle. BMC Genet. 15, 21. https://doi.org/10.1186/1471-2156-15-21

Stalker, M.J., Hayes, M.A., 2007. Liver and biliary system, in: Jubb, Kennedy and Palmer's Pathology of Domestic Animals. Elsevier Philadelphia, PA, pp. 297–388.

Strainic, M.G., Liu, J., Huang, D., An, F., Lalli, P.N., Muqim, N., Shapiro, V.S., Dubyak, G.R., Heeger, P.S., Medof, M.E., 2008. Locally produced complement fragments C5a and C3a provide both costimulatory and survival signals to naive CD4+ T cells. Immunity 28, 425–435. https://doi.org/10.1016/j.immuni.2008.02.001

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R.,

Diemer, K., Muruganujan, A., Narechania, A., 2003. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res. 13, 2129–2141. https://doi.org/10.1101/gr.772403

Tizioto, P.C., Coutinho, L.L., Decker, J.E., Schnabel, R.D., Rosa, K.O., Oliveira, P.S.N.N., Souza, M.M., Mourão, G.B., Tullio, R.R., Chaves, A.S., Lanna, D.P.D.D., Zerlotini-Neto, A., Mudadu, M.A., Taylor, J.F., Regitano, L.C.A.A., 2015. Global liver gene expression differences in Nelore steers with divergent residual feed intake phenotypes. BMC Genomics 16, 242. https://doi.org/10.1186/s12864-015-1464-x

Weber, K.L., Welly, B.T., Eenennaam, A.L. Van, Young, A.E., Reverter, A., Rincon, G., 2016. Identification of Gene Networks for Residual Feed Intake in Angus Cattle Using Genomic Prediction and RNA-seq 1–19. https://doi.org/10.1371/journal.pone.0152274

Zhou, N., Lee, W.R., Abasht, B., 2015. Messenger RNA sequencing and pathway analysis provide novel insights into the biological basis of chickens' feed efficiency. BMC Genomics 16, 195. https://doi.org/10.1186/s12864-015-1364-0

Zipfel, P.F., Heinen, S., Józsi, M., Skerka, C., 2006. Complement and diseases: Defective alternative pathway control results in kidney and eye diseases. Mol. Immunol. 43, 97–106. https://doi.org/https://doi.org/10.1016/j.molimm.2005.06.015

Zipfel, P.F., Skerka, C., Hellwage, J., Jokiranta, S.T., Meri, S., Brade, V., Kraiczy, P., Noris, M., Remuzzi, G., 2002. Factor H family proteins: on complement, microbes and human diseases. Biochem. Soc. Trans. 30, 971 LP – 978. https://doi.org/10.1042/bst0300971

3. **Chapter 2:** A pipeline for detection of potential genomic functional variants based on multi-tissue RNA-seq data: The case of feed efficiency in beef cattle

Gabriela Ribeiro[1], Fernando Baldi [2], Aline S.M. Cesar [3], Pâmela A. Alexandre [4], José B. S. Ferraz [5] and Heidge Fukumasu[6]

Submmited to Genetics Selection Evolution on 25 November of 2020.

[1] Department of Veterinary Medicine, Faculty of Animal Science and Food Engineering, University of Sao Paulo, Pirassununga, Sao Paulo 13635-900, Brazil;

[2] Department of Animal Science, São Paulo State University (UNESP), Jaboticabal, São Paulo, Brazil;

[3] Escola Superior de Agricultura "Luiz de Queiroz", University of Sao Paulo, Piracicaba, São Paulo, Brazil;

[4] CSIRO Agriculture & Food, 306 Carmody Rd., St. Lucia, Brisbane, QLD 4067, Australia;

[5] Department of Veterinary Medicine, Faculty of Animal Science and Food Engineering, University of Sao Paulo, Brazil; jbferraz@usp.br

[6] Department of Veterinary Medicine, Faculty of Animal Science and Food Engineering, University of Sao Paulo, Brazil; fukumasu@usp.br

Abstract

Background: The identification and selection of animals for feed efficiency (FE) is extremely important for sustainable and productive livestock and the best approach so far is the use of genomic selection based on DNA markers. However, FE is a complex phenotype, and the identification of potential functional variants is difficult. Here we

propose a pipeline for the identification of potential functional variants (PFV) for FE based on multi-tissue RNA-seq of relevant organs in beef cattle.

Results: The pipeline, was based on RNA-seq data from 5 different tissues from animals divergent for FE evaluated by residual feed intake, followed by the call of variants with the GATK tool, statistical analysis by Plink, identification of the consequences and impact of the variants by the Ensembl VEP, selection of relevant PFV and validation by weighted single-step GWAS (ssGWAS) and genomic selection, in which the linear model includes the fixed effects of the contemporary group, age as a covariate, and the direct additive genetic effect. At last, functional enrichment analysis with genes enriched for PFVs was performed using Panther tool. With the analysis of the call for variants, 169 significant variants were detected in all tissues, when a deeper analysis was carried out to know the impact and consequence of the variants, on the function of proteins, it was demonstrated that 20.4% have moderate or high impact. Adding information from PFV improved the prediction ability for RFI and less inflated prediction were obtained using PFV from liver and muscle, mainly those involved with the biological pathway of MHC, which is the main biological system of immunity.

Conclusions: Here we propose a consistent pipeline that identifies PFV and its biological pathways, and can be used in genetic prediction programs, helping to identify young animals without records.

3.1. Introduction

Genome wide association studies (GWAS) are often used to determine DNA variants related to a given phenotype. In livestock science they establish quantitative trait loci (QTL), identify candidate genes and are the basis for genome selection [1]. However, although there are some attempts to find functional (causal) variants of quantitative and complex traits based on GWAS results, it is a task comparable to looking for a needle in the haystack. That is because DNA genotyping chips have DNA markers interspaced in all chromosomes and the QTLs are often quite large [2]. On the other hand, there is an urgent need to establish the functional variants of complex phenotypes in livestock, because when the variant has a causal relationship with the phenotype, it can be used in different populations and even in different breed, from

which the discovery of the polymorphism was generated. With these specific DNA markers in hands, animal breeding will advance even further and faster.

One possibility to detect functional variants is the analysis of whole-genome sequencing, an approach that is still very costly nowadays. Another possibility is the whole-exome sequencing that has a reduced cost, but still lacks information regarding the importance of the polymorphism within cell and/or tissue architecture [3]. One should have in mind that a complex phenotype is made by contributions of several cell types, organs and their interactions. Therefore, a tissue-level systems biology approach should be considered, since it might point specific DNA variants associated with relevant biological processes for specific phenotypes.

Feed efficiency (FE) in beef cattle is one of the most important traits of livestock. While beef cattle produce high-quality meat from low-quality forage, they are one of the least efficient animals to convert feed into protein [4]. As a consequence, pasture grazing beef cattle are recognized as one of the biggest contributors to green-house emissions [5]. Therefore, more efficient animals are highly needed worldwide. Their improved productivity and sustainability are reflected in a reduced production cost, that can reach 75% in feedlot systems [6], and a decreased methane production, one of the green-house gases reducing the impact on the environment [7,8]. However, the identification of high FE animals is not an easy task. It is a complex phenotype, being controlled by several interconnected mechanisms [9,10]. Thus, it is necessary to understand the biological basis of FE to define future animal breeding programs [11].

Our group was the first to described a pathophysiological mechanism associated with FE in beef cattle: liver inflammation due to altered metabolism and/or bacterial translocation/infection [12], which was partially corroborated by others [13–15]. We also unravel the metabolic pathways related to FE in Nellore cattle showing an increased bacterial load in low feed efficient animals which is in part responsible for the hepatic lesions and inflammation in these animals [16]. Previously, some QTL's for feed efficiency in Nellore beef cattle were found by *conventional* GWAS [17–21], however only attempts to find causal variants were done.

Therefore, here we propose a *in silico* pipeline to overcome these limitations based on the identification of genetic variants from RNA-sequencing (RNA-seq) data from physiologically phenotype-related organs, followed by a classification of the potential

functional variants according to their effects on protein expression and function. We also validated the potential functional variants by a *weighted single-step* GWAS (ssGWAS) and genomic prediction in a different non-related population.

3.2. Methods

3.2.1. Phenotypic data and biological sample collection

Ninety-eight Nellore bulls (*Bos indicus*) (16 to 20 months old and 376 ± 29 kg BW) were evaluated in a feeding trial comprised of 21 days of adaptation to feedlot diet followed by a 70-day period of data collection. Total mixed ration was offered *ad libitum* and daily dry matter intake (DMI) was individually measured. Animals were weighted at the beginning, at the end and every two weeks during the experimental period. Feed efficiency was estimated by residual feed intake (RFI) [22]. Forty animals selected either as high feed efficiency (HFE) or low feed efficiency (LFE) groups were slaughtered on two days with a 6-day interval. Liver biopsy samples were collected from each animal and quickly frozen in liquid nitrogen and stored at -80 °C. Further information about management and phenotypic measures of the animals used in this study can be found elsewhere [12].

3.2.2. RNA-seq data

Samples of nine animals from each FE group (high and low) were selected for RNA-seq using RFI. The total RNA from liver, muscle, adrenal, pituitary and hypothalamus samples was extracted by using the RNeasy mini kit (QIAGEN, Crawley, West Sussex, UK) according to the instructions provided by the manufacturer. The total RNA quality and quantity were assessed using automated capillary gel electrophoresis on a Bioanalyzer 2100 with RNA 6000 Nano Labchips according to the manufacturer's instructions (Agilent Technologies Ireland, Dublin, Ireland). Samples that presented RNA integrity number (RIN) less than 8.0 were discarded. The mRNA libraries were constructed using the TruSeq™ Stranded mRNA LT Sample Prep Protocol and sequenced on Illumina HiSeq 2500 equipment in a HiSeq Flow Cell v4 using HiSeq SBS Kit v4 (2x100pb). FastQC software (Babraham Institute, Cambridge, UK, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to obtain the

sequencing quality. Removal of PolyA / T tails and adapters was performed using Seqyclean software (University of Idaho: Institute of Bioinformatics and Evolutionary Studies, Moscow, USA, https://bitbucket.org/izhbannikov/seqyclean), so only bases with quality ≥ 20 and complete readings with at least 50 bp, were further studied for further analysis. The alignment of the reads was done using the STAR software version 2.7 [23] with the reference genome *Bos taurus* ARS-UCD1.2 (Ensembl, ftp://ftp.ensembl.org/pub/release-98/fasta/bos_taurus/dna/), allowing two mismatches per read.

3.2.3. Protocol to call the potential functional variants associated with FE

Genome Analysis Toolkit (GATK) software version 4.0.11.0 [24] was used to call the variants. HaplotypeCaller command was applied to identify the variants (Single Nucleotide Polymorphism - SNPs) and insertions and deletions (INDELs). The variants underwent quality control on the GATK software as follows: (Variant Quality Score - QUAL) > 30, depth of sequencing (Deph Plot - DP) > 4, amount of available coverage (QualByDepth - QD) > 3, polarization trend (FisherStrand - FS) > 30 and the general mapping quality of readings that support a variant call (RMSMappingQuality- MQ) < 35. Statistical analysis was performed by Plink software [25] considering MAF < 40% and call rate equal to 50%. For the allele frequency test between HFE and LFE groups, the Cochran-Armitage trend analysis was used considering significant differences when $P<0.05$.

3.2.4. Characterization of the effects of variants on protein sequence and function

The potential functional variants were analyzed in the Variant Effect Predictor (VEP) online tool release 98 [26] which predicts the functional effects of the variants. Potential functional variants were analyzed by the Scale-Invariant Feature Transform (SIFT) score, a statistical tool of the VEP online software. SIFT is an algorithm that predicts whether an amino acid substitution affects the function of the protein. This analysis was performed on the basis of the gene sequence homology and the physical properties of the amino acids, where the like sequences are first sought, then strictly related sequences (which may share functions similar to the query sequence), with the

alignment of the sequences chosen is possible to calculate the normalized probability for all possible substitutions of the alignment. Thus, generating a SIFT score ranging from 0 to 1, being classified as deleterious values below 0.05 and above 0.05 is considered as tolerated. The deleterious classification indicates that the amino acid change will have a great impact on the proteins, since the tolerated classification indicates that the impact will not be so great. The tool also indicates the type of impact, which can be of the high type, causing protein truncation and loss of function; the moderate type, a non-disruptive variant that can alter the protein's effectiveness; low type, unlikely to alter the behavior of proteins; type modifier, non-coding variants or variants that affect non-coding genes, where predictions are difficult or there is no evidence of impact. In this research, only high and moderate impacts were considered.

3.2.5. Functional enrichment analysis

The functional enrichment analysis was performed with PANTHER version 15.0 [27] to identify biological pathways over-represented in the set of genes with potential functional variants. The analysis type was “PANTHER Overrepresentation test” using the gene list with potential functional variants against the reference list of *Bos taurus* (all genes in the database). The annotation data set used was “Reactome Pathways” and statistical analysis was performed with Fisher’s Exact test and correction by False Discovery Rate. Significant pathways were considered when FDR<0.05.

3.2.6. Validation of the potential function variants by genomic prediction

3.2.6.1. General information about the data

Records for RFI were obtained from feed efficiency tests carried out between 2011 and 2018, from an independent population of this study, with phenotypic and genotypic information of 4,653 and 5,117 animals, respectively, were considered [28]. The relationship matrix used in the analyses was calculated based on pedigree information from 19,507 animals, provided by the Nelore Brazil Breeding Program, coordinated by the National Association of Breeders and Researchers. More information regarding the set of animals used in this study can be found elsewhere [18].

3.2.6.2. Genomic information

Phenotypic and genotypic records from the Nelore Brazil breeding program coordinated by the National Association of Farmers and Researchers (*Associação Nacional de Criadores e Pesquisadores* – ANCP, Ribeirão Preto-SP, Brazil) were used for this study. A total of 963 animals were genotyped using the Illumina BovineHD BeadChip (Illumina Inc., San Diego, CA, USA), which contains 777,962 SNP markers of an independent population. These animals were used as a reference population to impute genotypes of 5,117 animals, previously genotyped with a low-density panel (CLARIFIDE® Nelore 3.1) encompassing over 27,000 SNP markers. Genotype imputation was performed using the FImpute 2.2 software [29]. The quality control criteria were performed by the PREGSF90 package [30], removing animals and markers with call rate < 0.90 and minor allele frequency (MAF) < 0.05. Monomorphic SNPs with redundant position and those located in non-autosome chromosomes were removed. Additionally, animals and SNPs with Mendelian conflicts were excluded.

3.2.6.3. Weighted single step Genome-Wide Association Study

The GWAS analysis was performed using the single-step GWAS (ssGWAS) methodology [31]. The ssGWAS was performed in order to estimate the weights for SNPs markers iteratively. The linear model included the fixed effects of contemporary group (CG) and the animal age as covariable, and the random direct additive genetic effect. Farm, management group, sex, feed efficiency test, year and birth season, composed the CG. Records within ±3.5 standard deviations from the CG mean were considered in the analysis, and CG that had at least four animals were considered in the analysis.

3.2.6.4. Prediction models

For the genetic evaluation of RFI the same model applied for ssGWAS analyses was applied. To calculate the prediction ability, the dataset was split into training (3,253 animals) and validation (1,864 animals) subsets. The validation subset consisted of genotyped young animals without progeny records and the training was composed by

genotyped sires and dams with progenies. Phenotypic and genotypic information from validation animals were omitted in the training set. To evaluate the prediction ability of GEBV in the validation subset, the prediction accuracies were calculated according to the Beef Improvement Federation (BIF)[32] as follows:

$$\mathrm{Acc_{BIF}} = 1 - \sqrt{\frac{\mathrm{PEV}}{(1+\mathrm{F_i}) \times \sigma^2_a}}$$

where PEV is the prediction error variance, σ^2_a the additive genetic variance, and $\mathrm{F_i}$ the inbreeding coefficient. The regressions coefficient between the GEBV obtained using the complete data set using the unweighted **G** and the GEBV estimated for different weighted G scenarios with or without potential functional variants (PFV) was used evaluate the prediction inflation.

The **G** matrix was obtained following [33]:

$$\mathbf{G} = \frac{(\mathbf{M} - \mathbf{P})(\mathbf{M} - \mathbf{P})'}{2\sum_{j=1}^{m} p_j (1 - p_j)}$$

where **M** is an allele-sharing matrix with *m* columns (*m* total number of markers) and *n* rows (*n* = total number of genotyped individuals), and **P** is a matrix containing the frequency of the second allele (*pj*), expressed as 2*pj*. M*ij* was 0 if the genotype of individual *i* for SNP *j* was homozygous AA, was 1 if heterozygous, or 2 if the genotype was homozygous BB. To account for heterogeneous SNP weights, a matrix of weights should be included in the formula for constructing **G**, where *var(s)* is the vector containing the variance of individual SNP effects, and d_i is the $_i$*th* diagonal element of **D**, accounting for the $_i$*th* SNP weight:

$$var(s) = \mathbf{D} = \begin{vmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_n \end{vmatrix}$$

Based on that, a weighted relationship matrix can be defined as:

$$\mathbf{G_w} = \frac{\mathrm{MDM'}}{2\sum_{\mathrm{j=1}}^{\mathrm{m}} \mathrm{p_j}\,(1-\mathrm{p_j})}$$

where **D** is a matrix of weights and each diagonal element of this matrix is defined as

$$d_i = \sigma_{u,i}^2 \frac{\sum_{j=1}^{m} 2 \;\; p_j \;\; q_j}{\sigma_a^2}$$

where $\sigma_{u,i}^2$can be understood as SNP "prior variances" [34,35]. In practice, $\sigma_{u,i}^2$ are not known (or even well defined; [35]).

Several approaches exist to estimate individual SNP variances, but $\sigma_{u,i}^2$ can be approximated from estimates of SNP effects as follows. Genomic predictions were obtained by ssGBLUP, SNP effects can be calculated using a backsolving process [33,36,37]:

$$\hat{u} = \frac{1}{\sum_{j=1}^{m} 2\, p_j \;\; q_j} \mathrm{DM'G_w^{-1}}\hat{a}$$

where $\hat{u}$ is the vector of estimated SNP effects, and $\hat{a}$ is a vector of genomic EBV (GEBV). After the SNP effect was calculated, the $\sigma_{u,i}^2$ was obtained as proposed by [37].

The G matrix was constructed using different combinations of SNPs and weights: (a) Unweighted G matrix with 460,992 SNPs; (b) weights in D calculated based on genome-wide association studies (ssGWAS) using iterative ssGBLUP as in [37], updating GEBV and SNP weights for 2 iterations; c) weighted SNPs as b) and also including differential weights for SNPs neighboring the causative PFV for liver, adrenal, pituitary, hypothalamus and muscle tissue, respectively. The SNP markers that are causal or in linkage disequilibrium with potential functional variants should be

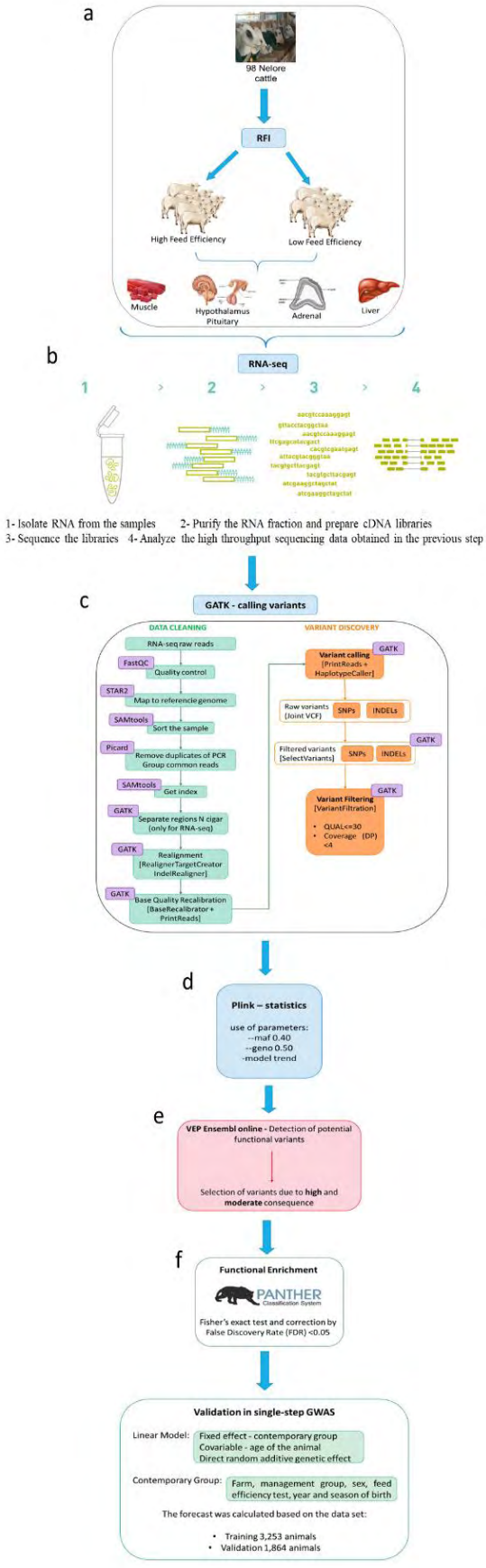
given higher weights. In these sense, three levels of weight for SNPs neighboring the PFV were tested, 1-fold, 2-fold and 3-fold the maximum weighted obtained in the ssGWAS after 2 iteration.

3.3. Results

3.3.1. Detection and characterization of potential functional variants (PFVs) associated with feed efficiency

The pipeline we proposed here (Fig. 1) is for the detection of PFVs based on RNA-seq data of relevant organs for a given phenotype, in this case feed efficiency in beef cattle. For this experiment, we used samples from 9 animals of each group (HFE and LFE) analyzing 18 samples of liver, hypothalamus and pituitary; 17 of muscle and 15 of adrenal gland, yielding 13,3 million reads per sample on average (Table 1 and Supplementary table 1). Initially, variants were called from the 5 different organs and the number of uniquely variants was 2,000,936 dues to the overlap of variant calling in different organs (Table 1). After filtering the variants by MAF and call rate, a total of 11,35% (227,225 uniquely variants) was used for statistical analysis where 4,39% (9,986 variants) were significantly associated with FE. Next, we classified the PFVs according to the impact on protein function where 20,0% (1,995) were classified with moderate impact and only 0.78% (78) were classified with high impact on protein function (Table 1 and Fig. 2).

Figure 1. The pipeline for PFV detection.



Step-by-step details of the material and methods for obtaining potential causal variants. a - selection of animals for feed efficiency and sample preparation; b - RNA-seq analysis (paired end); c - data treatment and call for variants by the GATK tool; d - statistical analysis and genetic association by Plink; e - identification of the impact and consequence of variants by the Ensembl VEP online; f - functional enrichment by Panther (GO); g - validation of results by the GWAS of an independent population.

Table 1. Total call for variants, filtering, significant variants, impact levels, distributed by tissue.

Organ	Total variant calling	After filtering (MAF and call rate)	Significant variants	Classification of the impact	
				Moderate	High
Liver	484,589	46,268	1,110	263	6
Muscle	459,057	80,818	2,540	501	25
Hypothalamus	1,037,253	143,540	4,586	834	24
Pituitary	846,809	125,141	3,782	847	21
Adrenal	745,878	130,738	3,575	657	10
Total	**3,573,586**	**526,505**	**15,593**	**3,102**	**86**
Uniquely variants*	**2,000,936**	**227,225**	**9,986**	**1,995**	**78**

The results of the Total variant number, Filter and Significant variants are in quantities, that is, the number of variants that can be SNPs or/and INDELs.

* this data means the number of uniquely variants since there are variants detected in more than one tissue.

The majority of PFVs with moderate or high impact are missense SNPs (Table 2 and Fig. 2), but there are other important protein consequences as frameshift INDELs, stop gained INDELs and inframe insertion INDELs all altering protein sequences and function.

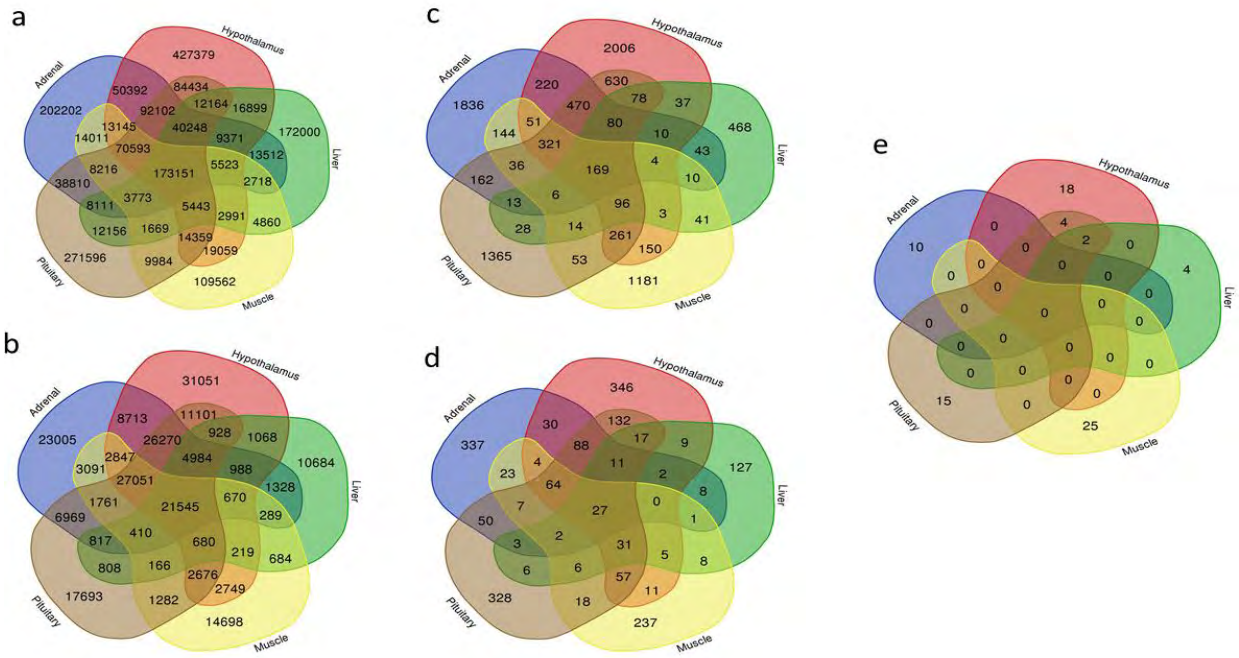
Table 2. Variants distributed according to the consequences on protein sequence.

	Consequence *												
Organ	Missense	Splice region	Splice donor	Frameshift	Stop gained	Inframe insertion	Inframe deletion	Splice acceptor	Start lost	Stop lost	Intron variant	Start retained	Affected genes
Liver	260	2	3	3	1	3	-	-	-	-	-	-	120
Muscle	494	24	8	14	-	3	4	1	1	1	-	-	211
Hypothalamus	818	11	11	10	3	10	9	-	-	-	-	-	352
Pituitary	823	9	5	12	3	10	17	-	2	1	-	2	343
Adrenal	635	6	2	3	2	9	15	3	-	-	2	-	286
Total	3.030	52	29	42	9	35	45	4	3	2	2	2	1,312

* The number of consequences of the variants can vary, as a variant can have more than one consequence.

Interestingly, we found 169 significant variants expressed in all the five organs (Fig. 2.c), however only 27 variants have moderate impact, and none had high impact providing evidence of tissue specific effects of high impact functional variants. In addition, the pituitary and the hypothalamus ranked as the first and second organs with the most significant variants in our study.

Figure 2. Variants overlay (SNPs and INDELs)



a – Number of SNPs and INDELs variants (insertions and deletions), b - variants filtered according to maf <0.40 and call rate 0.50 and associated with the genotype; c - significant variations associated with feed efficiency; d - potential functional variants with moderate impact, that is, non-disruptive variant that can alter the effectiveness of the protein; e - potential functional variants with high impact, causing protein truncation and loss of function.

3.3.2. Functional analysis of PFVs

Functional enrichment analysis was performed first with all genes carrying the PFVs and also for each tissue alone. In the first scenario, all enriched terms are related to the Class I MHC (major histocompatibility complex) mediated antigen processing and presentation, an important part of the adaptative immune response (Supplementary file 2). The analysis of each tissue alone enriched for

immune response terms in all tissues except for hypothalamus. In this specific tissue, although being the tissue with the higher number of significant PFVs detected, only one biological pathway was significantly enriched (DNA double-strand break response). The list of these 28 genes, their potential functional variants, the impact on the protein sequence and the frequency in each group can be found in the Supplementary file 2.

3.3.3. Validation of the findings by genomic prediction

A total of 144, 252, 413, 416 and 340 PFVs presented in the liver, muscle, hypothalamus, pituitary and adrenal were adjacent to 223, 422, 694, 697 and 554 SNP markers presented in the BovineHD (Illumina), respectively. To perform the validation and genomic predictions for validation set, the adjacent SNP markers to PFV were differentially weighted based on the results obtained with ssGWAS using the training set. The genomic prediction ability for RFI in the validation set when the PFV were not included in the analyses is presented in Table 3. The prediction accuracy for RFI using the weighted G matrix (ssGBLUP+wG) obtained in the ssGWAS of training set was higher than unweighted G matrix (ssGBLUP). However, the prediction accuracy improvement was higher when RFI records were added (ssGBLUP+rec) in the validation subset compared to apply a weighted G matrix (Table 3). As expected, the highest prediction accuracy in the validation set was obtained when all available information was considered and the G matrix was weighted (ssGBLUP+wG+rec), however, more inflated predictions for RFI were obtained. Applying also the ssGBLUP method, [38] (0.45) and [39] (0.22) reported higher prediction ability for RFI also in Nellore cattle. The less inflated predictions for RFI were obtained with the model that includes unweighted genomic information and records of validation set, however, in this scenario phenotypic information is necessary. It is important to highlight that the more realist scenario is only use genomic information to predict the GEBV of young animals without RFI records at early ages, in order to maximize the genetic progress for RFI and took the advantage of genomic selection. The availability of phenotypic records for RFI is not common in beef cattle breeding programs because is expensive to assess records for this important trait.

Table 3. Prediction ability (Acc) and regression coefficient (b) for RFI in the validation set.

Validation without PFVs[1]	Acc	b (SE)
SsGBLUP	0.10	0.48 (0.02)
ssGBLUP+Wg	0.15	0.85 (0.03)
ssGBLUP+records	0.23	1.00 (0.03)
ssGBLUP+wG+records	0.29	1.10 (0.02)

[1]ssGBLUP: ssGBLUP using unweighted G matrix; ssGBLUP+wG: ssGBLUP using weighted G matrix; ssGBLUP+records: ssGBLUP using unweighted G matrix and records; ssGBLUP+wG+records: ssGBLUP using weighted G matrix and records

Adding the information from PFVs for the five organs with a differentially weighted for adjacent SNPs, the prediction ability improved in comparison to ssGBLUP+wG (Table 4). The prediction ability for RFI using 1-fold, 2-fold and 3-fold were almost the same for the different tissues, however the highest prediction accuracy was obtained in the 3-fold scenario, where the prediction accuracy increased from 31.03% to 40% compared to weighted G matrix without consider the PFVs. Despite the higher prediction accuracy for RFI when SNP markers adjacent to PFV were differentially weighted, more inflated predictions were obtained for RFI as the weighted for PFV increased. However, is important to highlight that the increase of prediction inflation was lower in the liver and muscle tissue compared to adrenal, pituitary or hypothalamus.

Table 4. Prediction ability (Acc) and regression coefficient (b) for RFI differentially (SNPs and PFV).

Validation for functional mutations	Adrenal		Pituitary		Hypothalamus		Muscle		Liver	
	Acc	**b (SE)**	**Acc**	**b (SE)**	**Acc**	**b (SE)**	**Acc**	**b (SE)**	**Acc**	**b (SE)**
ssGBLUP+wG +QTN:1-fold	0.16	0.74 (0.03)	0.16	0.723 (0.03)	0.15	0.73 (0.03)	0.15	0.85 (0.03)	0.15	0.82 (0.03)
ssGBLUP+wG+QTN:2-fold	0.18	0.67 (0.03)	0.18	0.65 (0.03)	0.18	0.66 (0.03)	0.18	0.75 (0.03)	0.18	0.79 (0.03)
ssGBLUP+wG+QTN:3-fold	0.20	0.62 (0.03)	0.19	0.60 (0.03)	0.19	0.61 (0.03)	0.20	0.71 (0.03)	0.20	0.77 (0.03)
ssGBLUPrecords+wG+QTN:1-fold	0.31	1.02 (0.02)	0.31	1.00 (0.02)	0.31	1.02 (0.02)	0.29	1.11 (0.2)	0.30	1.08 (0.02)

Prediction ability (Acc) and regression coefficient (b) for weighted single-step GBLUP (ssGBLUP+wG) including selected variants (PFV) in the model and applying different weighting approaches for PFV (1-fold, 2-fold and 3-fold the maximum weighted obtained in the ssGWAS)

3.4. Discussion

It is imperative the determination of potential functional variants of complex, quantitative phenotypes of livestock species. Some works have been trying to find the best way for the identification of PFVs, but weren't able to state whether these variants were in fact causal and what are the consequences for production characteristics [40–43]. Here we proposed a new methodology for identifying PFVs, together with a powerful validation in GWAS in an independent population, and we showed that in fact the PFVs increased the accuracy for genome selection. Our approach uses a systems biology approach coupled with multi-tissue phenotyping and the effects on protein consequences of the PFVs for a given phenotype, in this case the feed efficiency in beef cattle.

The methodology used in this work consists of two major stages, the first being the identification of PFVs and the second a validation by weighted GWAS. In the first stage, GATK was used to call the variants: it compares the case RNA sequencing data with the control (bovine reference genome) using powerful filtering and statistical tools. This tool is widely used in several works to identify variants [44–47] and it is generally used with the HaplotypeCaller algorithm, which improves performance by making the tool more accurate [48]. To select a variant calling tool, one needs to have a good combination of processing time, precision and sensitivity (call quality) of the genotyping. When analyzing GATK with other tools (Findvar, SAMtools, Graphtyper) that have the same functions and considering the processing time, GATK is at a disadvantage compared to the other tools [48,49]. However, when comparing the number of polymorphic sites found by the tools (homozygous and heterozygous), GATK is of great advantage [49,50]. Regarding the call for false positives, the tool with the lowest percentages was Findvar, followed by GATK and later by SAMtools [49]. Although GATK has a long processing time, it can

be compensated by the number of polymorphic sites and the median filtering of false positives, making it a balanced tool for the identification of PFVs.

The second stage, on the other hand, involves validating the results obtained in the first stage using the ssGWAS and perform genomic prediction in an independent population. The ssGWAS is widely used to find DNA variants associated with a characteristic. Nevertheless, it does not allow the identification whether the variant is causal or not. Some studies [51–54], point out that the use of *expression quantitative trait loci (*eQTL) mapping can help in the identification of causal variants and also in the distinction between pleotropic and binding effects [55]. However, a denser SNP panel is needed to accurately locate mutations and the genes involved, making eQTL studies very expensive. A viable alternative is the methodology we have developed, a systems biology-based characterization of the phenotype to detect PFVs and further use as additional information for genomic prediction. In this study we used data from PFVs, coming from an analysis of RNA-seq, which can serve as a tool to improve genetic predictions. Thus, the two factors (GWAS + PFV) differentially weighted and added together increase the ability of genomic prediction. Adding external information from PFV identified by analysis of RNA-seq and also including information from ssGWAS both contributes to reduce selection risk by improving GEBV accuracies, however, more inflated predictions were obtained as the weight for genomic information and PFV increased. The prediction ability for RFI was close to those obtained in previous studies using taurine breeds [56,57] and indicine breeds [38,39]. Comparing the prediction inflation from different tissue, it was possible to see that the tissues with the less inflated predictions were liver and muscle compared to adrenal, pituitary and hypothalamus, indicating that the PFV present in the liver and muscle contributes to more informative SNPs compared to other tissues.

The gains in prediction accuracy are expected when widely known candidate regions identified by GWAS were included and weighted in the prediction models [58]. As examples: (1) when 1,623

variants from different breeds where added to a custom SNP chip, an accuracy gain of 2 % was found when more weight was assigned to the QTN (Quantitative Trait Nucleotide) [59]; (2) the addition of selected sequence variants from a multiracial GWAS generates an increase of up to 10% in accuracy [60] and; (3) the incorporation of potential causative SNPs and removal of adjacent SNPs increase the accuracy by 2.5% [61]. Thus, the results obtained in our validation study pointed out that the incorporation of information derived from PVFs improved the genomic prediction for RFI though increasing the chance or probability to pick-up genetic marker in strong linkage disequilibrium with causal mutations and providing higher contribution of these SNP markers to the additive genetic variance for RFI.

The proposed pipeline also allowed the functional enrichment analysis of the genes with PFVs and in this case, a significantly enrichment for Class I MHC mediated antigen processing and presentation was found for all organs, along with the DNA double-strand break response specifically in the hypothalamus. An overrepresentation of the BOLA gene and its polymorphisms (BOLA-DQA5, BLA-DQB, BOLA-DQA1, BoLA (ENSBTAG00000005182), JPS.1 (also known as BoLa) and BOLA-NC1) was found. These genes are part of the main bovine histocompatibility complex (MHC) class I (JSP.1, BOLA, BoLA (ENSBTAG00000005182), BOLA-NC1) and class II (BLA-DQB, BOLA-DQA5). MHC is a fundamental part of the immune system, which has several functions, being the most important the response against infectious diseases [62,63]. The MHC is divided into 3 groups, class I, class II and class III. Class I molecules have as their main function to present peptides to CD8 + T lymphocytes, which in turn kill cells infected by viruses and neoplasms [64]. Class II molecules have direct and indirect functions, which include participation in antigen-presenting cells (APCs), such as dendritic and macrophage cells. These APCs have antigens derived from extracellular CD4 + T cell pathogens, which in turn activate macrophages and B cells to generate inflammatory and antibody responses, respectively. Indirectly, class II molecules

participate in the immune process through steroid 21-hydroxylase enzymes and tumor necrosis factors [64]. In MHC class I there are two subclasses, classic MHC-I (MHC-Ia) and non-classic (MHC-Ib). The BOLA-NC1 gene participates in the non-classical pathway, responsible for generating membrane isoforms from alternative splicing (differential splicing) [65–67]. In humans, the BOLA-NC1 gene is responsible for secreting molecules that interact with inhibitory receptors expressed by natural killer cells (NK), T lymphocytes and APC, in order to inhibit cells [68–76]. Class II is classified in two groups: DR and DQ [77]. BLA-DQB and BOLA-DQA5 participate in the DQ group, which is highly polymorphic and has different effects, but mainly acts in decreasing the response in CD4 helper T cells [78,79]. These genes have already been reported in other studies such as bovine leukemia virus [77], comparison of MHC class II diversity between different breeds of cattle [80], how bovine MHC influences disease function and susceptibility [81] and bacterial infection and inflammation in dairy cattle [82,83].

An increasing number of studies have demonstrated the link between the immune system and feed efficiency in different livestock species. In one of our previous works [12] the transcriptomic analysis indicated that Low Feed Efficiency (LFE) animals had more periportal liver lesions and pronounced inflammatory response, which is mediated by the immune system. We also demonstrated that LFE animals have increased bacterial load which is at least in part, responsible for the hepatic lesions and inflammation in these animals [16]. Therefore, the identification of PFV for feed efficiency in beef cattle in genes related to immune response is very plausible and open the possibility for fast improvement, by genetic selection, of this important phenotype in this species. It doesn’t escape our attention that as the biological pathways of Feed Efficiency are similar for other species as pigs, poultry and dairy cows, the detection of PFVs for genetic selection in these species will be very important.

3.5. Conclusion

Here, a pipeline to identify potential functional variants for a complex polygenic phenotype was proposed. The strategy added accuracy to genomic prediction model and increased the prediction accuracy for young animals without phenotypic records. The study also found that liver and muscle variants provided better genomic predictions for young animals, highlighting the main importance of these two organs for the phenotypic characteristic of feed efficiency (FE). At last, in this work, we showed that one of the most important biological pathways related to FE is the Class I MHC mediated antigen processing and presentation, an important part of the adaptative immune response.

3.6. References

1. Wang M, Wang Q, Pan Y. From QTL to QTN: candidate gene set approach and a case study in porcine IGF1-FoxO pathway. PLoS One. 2013/01/14. Public Library of Science; 2013;8:e53452–e53452.

2. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19:491–504.

3. Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proc Natl Acad Sci U S A. 2015/03/31. National Academy of Sciences; 2015;112:5473–8.

4. Shepon A, Eshel G, Noor E, Milo R. Energy and protein feed-to-food conversion efficiencies in the US and potential food security gains from dietary changes. Environ Res Lett. IOP Publishing; 2016;11.

5. Zhuang M, Lu X, Caro D, Gao J, Zhang J, Cullen B, et al. Emissions of non-CO2 greenhouse gases from livestock in China during 2000–2015: Magnitude, trends and spatiotemporal patterns. J Environ Manage. 2019;242:40–5.

6. Paper R. Simulation Modelling of the Cost of Producing and Utilising Feeds for Ruminants. 2010;14.

7. Difford GF, Løvendahl P, Veerkamp RF, Bovenhuis H, Visker MHPW, Lassen J, et al. Can greenhouse gases in breath be used to genetically improve feed efficiency of dairy cows? J Dairy Sci. 2020;103:2442–59.
8. Hegarty RS, Goopy JP, Herd RM, McCorkell B. Cattle selected for lower residual feed intake have reduced daily methane production1,2. J Anim Sci. 2007;85:1479–86.
9. de Haas Y, Calus MPL, Veerkamp RF, Wall E, Coffey MP, Daetwyler HD, et al. Improved accuracy of genomic prediction for dry matter intake of dairy cattle from combined European and Australian data sets. J Dairy Sci. Elsevier; 2012;95:6103–12.
10. Veerkamp RF, Pryce JE, Spurlock D, Berry D, Coffey M. Selection on feed intake or feed efficiency: A position paper from gDMI breeding goal discussions. Interbull Bull. 2013;0.
11. Cantalapiedra-Hijar G, Abo-Ismail M, Carstens GE, Guan LL, Hegarty R, Kenny DA, et al. Review: Biological determinants of between-animal variation in feed efficiency of growing beef cattle. Animal. Cambridge University Press; 2018. p. S321–35.
12. Alexandre PA, Kogelman LJA, Santana MHA, Passarelli D, Pulz LH, Fantinato-Neto P, et al. Liver transcriptomic networks reveal main biological processes associated with feed efficiency in beef cattle. BMC Genomics. 2015;16.
13. Paradis F, Yue S, Grant JR, Stothard P, Basarab JA, Fitzsimmons C. Transcriptomic analysis by RNA sequencing reveals that hepatic interferon-induced genes may be associated with feed efficiency in beef heifers. J Anim Sci. American Society of Animal Science; 2015;93:3331–41.
14. Tizioto PC, Coutinho LL, Decker JE, Schnabel RD, Rosa KO, Oliveira PSN, et al. Global liver gene expression differences in Nelore steers with divergent residual feed intake phenotypes. BMC Genomics. 2015;16:1–14.
15. Weber KL, Welly BT, Van Eenennaam AL, Young AE, Port-Neto LR, Reverter A, et al. Identification of Gene networks for residual feed intake in Angus cattle using genomic prediction and RNA-seq. PLoS One. 2016;11:1–19.
16. Fonseca LD, Eler JP, Pereira MA, Rosa AF, Alexandre PA, Moncau CT, et al. Liver proteomics unravel the metabolic pathways related to Feed Efficiency in beef cattle. Sci Rep [Internet]. Nature Publishing Group; 2019 [cited 2019 Apr 29];9:5364. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30926873
17. Olivieri BF, Mercadante MEZ, Cyrillo JNDSG, Branco RH, Bonilha SFM, De Albuquerque LG, et al. Genomic regions associated with feed efficiency indicator traits in an experimental nellore cattle population. PLoS One. Public Library of Science; 2016;11.

18. Brunes LC, Baldi F, Lopes FB, Lôbo RB, Espigolan R, Costa MFO, et al. Weighted single-step genome-wide association study and pathway analyses for feed efficiency traits in Nellore cattle. J Anim Breed Genet [Internet]. Blackwell Publishing Ltd; 2020 [cited 2020 Aug 18]; Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/jbg.12496
19. Higgins MG, Fitzsimons C, McClure MC, McKenna C, Conroy S, Kenny DA, et al. GWAS and eQTL analysis identifies a SNP associated with both residual feed intake and GFRA2 expression in beef cattle. Sci Rep. Nature Publishing Group UK; 2018;8:14301.
20. Santana MHAMHA, Utsunomiya YTYT, Neves HHRHHR, Gomes RCRC, Garcia JFJF, Fukumasu H, et al. Genome-wide association analysis of feed intake and residual feed intake in Nellore cattle. BMC Genet. 2014;15:1–8.
21. Santana MH de A, Oliveira Junior GA, Cesar ASMASM, Freua MCMC, Gomes R da C, Silva S da L e, et al. Copy number variations and genome-wide associations reveal putative genes and metabolic pathways involved with the feed conversion ratio in beef cattle. J Appl Genet. Journal of Applied Genetics; 2016;57:495–504.
22. Koch RM, Swiger LA, Chambers D, Gregory KE. Efficiency of feed use in beef cattle. J Anim Sci. 1963;22:486–94.
23. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
24. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. Cold Spring Harbor Laboratory Press; 2010;20:1297–303.
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. Cell Press; 2007;81:559–75.
26. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. 2016 [cited 2019 Nov 18];17:122. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4
27. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: A library of protein families and subfamilies indexed by function. Genome Res. 2003;13:2129–41.
28. Lewis ZT, Mills DA. Differential Establishment of Bifidobacteria in the Breastfed Infant Gut. Nestle Nutr Inst Workshop Ser. 2017/03/27. 2017;88:149–59.
29. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation

using information from relatives. BMC Genomics. 2014;15.
30. Aguilar I, Misztal I, Tsuruta S, Legarra A. PREGSF90 – POSTGSF90 : Computational Tools for the Implementation of Single-step Genomic Selection and Genome-wide Association with Ungenotyped Individuals in BLUPF90 Programs. 10th World Congr Genet Appl to Livest Prod. 2014.
31. Wang H, Misztal I, Aguilar I, Legarra A, Muir WM. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet Res (Camb). 2012;94:73–83.
32. BIF. For Uniform Beef Improvement Programs. Beef. 2002;
33. VanRaden PM. Efficient methods to compute genomic predictions. J Dairy Sci. Elsevier; 2008;91:4414–23.
34. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
35. Gianola D. Priors in whole-genome regression: The Bayesian alphabet returns. Genetics. 2013;194:573–96.
36. Strandén I, Garrick DJ. Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci. Elsevier; 2009;92:2971–5.
37. Wang H, Misztal I, Aguilar I, Legarra A, Fernando RL, Vitezica Z, et al. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. Front Genet. Frontiers Media S.A.; 2014;5:134.
38. Silva RMO, Fragomeni BO, Lourenco DAL, Magalhães AFB, Irano N, Carvalheiro R, et al. Accuracies of genomic prediction of feed efficiency traits using different prediction and validation methods in an experimental Nelore cattle population. J Anim Sci. 2016;94:3613–23.
39. Brunes, L. C.; Baldi, F.; Lopes, F. B.; Narciso, M. G.; Lobo, R. B.; Espigolan, R.; Costa MFO. M. Genomic prediction ability for feed efficiency traits using different models and pseudo-phenotypes under several validation strategies in Nelore cattle. Animal (in press); 2020.
40. Kumaran M, Subramanian U, Devarajan B. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinformatics. BioMed Central; 2019;20:342.
41. Cai Z, Guldbrandtsen B, Lund MS, Sahana G. Weighting sequence variants based on their annotation increases the power of genome-wide association studies in dairy cattle. Genet Sel Evol. BioMed Central; 2019;51:20.

42. Schnepp PM, Chen M, Keller ET, Zhou X. SNV identification from single-cell RNA sequencing data. Hum Mol Genet. 2019;28:3569–83.
43. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;43:11.10.1-11.10.33.
44. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011/04/10. 2011;43:491–8.
45. Tessier L, Côté O, Bienzle D. Sequence variant analysis of RNA sequences in severe equine asthma. PeerJ. PeerJ Inc.; 2018;6:e5759–e5759.
46. Patel SM, Koringa PG, Nathani NM, Patel N V, Shah TM, Joshi CG. Exploring genetic polymorphism in innate immune genes in Indian cattle (Bos indicus) and buffalo (Bubalus bubalis) using next generation sequencing technology. Meta gene. Elsevier; 2015;3:50–8.
47. Zwane AA, Schnabel RD, Hoff J, Choudhury A, Makgahlela ML, Maiwashe A, et al. Genome-Wide SNP Discovery in Indigenous Cattle Breeds of South Africa. Front Genet. Frontiers Media S.A.; 2019;10:273.
48. Ren S, Bertels K, Al-Ars Z. Efficient Acceleration of the Pair-HMMs Forward Algorithm for GATK HaplotypeCaller on Graphics Processing Units. Evol Bioinform Online. SAGE Publications; 2018;14:1176934318760543–1176934318760543.
49. VanRaden PM, Bickhart DM, O'Connell JR. Calling known variants and identifying new variants while rapidly aligning sequence data. J Dairy Sci. 2019;102:3216–29.
50. Crysnanto D, Wurmser C, Pausch H. Accurate sequence variant genotyping in cattle using variation-aware genome graphs. Genet Sel Evol. BioMed Central; 2019;51:21.
51. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. 2008/07/01. 2008;24:408–15.
52. Pausch H, Emmerling R, Schwarzenbacher H, Fries R. A multi-trait meta-analysis with imputed sequence variants reveals twelve QTL for mammary gland morphology in Fleckvieh cattle. Genet Sel Evol. BioMed Central; 2016;48:14.
53. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48:481–7.
54. van den Berg I, Hayes BJ, Chamberlain AJ, Goddard ME. Overlap between eQTL and QTL associated with production traits and fertility in dairy cattle. BMC Genomics. BioMed Central;

2019;20:291.
55. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet. 2016;48:481–7.
56. Lu D, Akanno EC, Crowley JJ, Schenkel F, Li H, Pauw M De, et al. Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes 1. 2018;1342–53.
57. Pryce JE, Arias J, Bowman PJ, Davis SR, Macdonald KA, Waghorn GC, et al. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. J Dairy Sci. 2012;95:2108–19.
58. Fragomeni BO, Lourenco DAL, Masuda Y, Legarra A, Misztal I. Incorporation of causative quantitative trait nucleotides in single-step GBLUP. Genet Sel Evol. 2017;49:59.
59. Brøndum RF, Su G, Janss L, Sahana G, Guldbrandtsen B, Boichard D, et al. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci. 2015;98:4107–16.
60. van den Berg I, Boichard D, Lund MS. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genet Sel Evol. 2016;48:83.
61. VanRaden PM, Tooker ME, O'Connell JR, Cole JB, Bickhart DM. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol. 2017;49:32.
62. Hill AVS. THE IMMUNOGENETICS OF HUMAN INFECTIOUS DISEASES. Annu Rev Immunol. Annual Reviews; 1998;16:593–617.
63. Hedrick PW, Parker KM, Gutiérrez-Espeleta GA, Rattink A, Lievers K. MAJOR HISTOCOMPATIBILITY COMPLEX VARIATION IN THE ARABIAN ORYX. Evolution (N Y). John Wiley & Sons, Ltd; 2000;54:2145–51.
64. Behl JD, Verma NK, Tyagi N, Mishra P, Behl R, Joshi BK. The Major Histocompatibility Complex in Bovines: A Review. Nsahlai I, Benko M, Albenzio M, editors. ISRN Vet Sci. International Scholarly Research Network; 2012;2012:872710.
65. Araibi EH, Marchetti B, Dornan ES, Ashrafi GH, Dobromylskyj M, Ellis SA, et al. The E5 oncoprotein of BPV-4 does not interfere with the biosynthetic pathway of non-classical MHC class I. Virology. 2006;353:174–83.
66. Birch J, Codner G, Guzman E, Ellis SA. Genomic location and characterisation of nonclassical MHC class I genes in cattle. Immunogenetics. 2008;60:267–73.
67. Davies CJ, Eldridge JA, Fisher PJ, Schlafer DH. Evidence for Expression of Both Classical and

Non-Classical Major Histocompatibility Complex Class I Genes in Bovine Trophoblast Cells. Am J Reprod Immunol. John Wiley & Sons, Ltd; 2006;55:188–200.
68. Bainbridge DRJ, Ellis SA, Sargent IL. HLA-G suppresses proliferation of CD4+ T-lymphocytes. J Reprod Immunol. 2000;48:17–26.
69. Braud VM, Allan DSJ, O'Callaghan CA, Söderström K, D'Andrea A, Ogg GS, et al. HLA-E binds to natural killer cell receptors CD94/NKG2A, B and C. Nature. 1998;391:795–9.
70. Ellis SA, Palmer MS, McMichael AJ. Human trophoblast and the choriocarcinoma cell line BeWo express a truncated HLA Class I molecule. J Immunol. 1990;144:731 LP – 735.
71. Ellis SA, Sargent IL, Redman CW, McMichael AJ. Evidence for a novel HLA antigen found on human extravillous trophoblast and a choriocarcinoma cell line. Immunology. 1986;59:595–601.
72. Hunt JS, Langat DK, McIntire RH, Morales PJ. The role of HLA-G in human pregnancy. Reprod Biol Endocrinol. 2006;4:S10.
73. Hunt JS, Langat DL. HLA-G: a human pregnancy-related immunomodulator. Curr Opin Pharmacol. 2009;9:462–9.
74. Hunt JS, Petroff MG, McIntire RH, Ober C. HLA-G and immune tolerance in pregnancy. FASEB J. John Wiley & Sons, Ltd; 2005;19:681–93.
75. Bouteiller P Le. HLA-G in the human placenta: expression and potential functions. Biochem Soc Trans. 2000;28:208–12.
76. Park GM, Lee S, Park B, Kim E, Shin J, Cho K, et al. Soluble HLA-G generated by proteolytic shedding inhibits NK-mediated cell lysis. Biochem Biophys Res Commun. 2004;313:606–11.
77. Takeshima S, Ohno A, Aida Y. Bovine leukemia virus proviral load is more strongly associated with bovine major histocompatibility complex class II DRB3 polymorphism than with DQA1 polymorphism in Holstein cow in Japan. Retrovirology. 2019;16:14.
78. Glass EJ, Oliver RA, Russell GC. Duplicated DQ Haplotypes Increase the Complexity of Restriction Element Usage in Cattle. J Immunol. 2000;165:134 LP – 138.
79. Bai L, Takeshima S-N, Sato M, Davis WC, Wada S, Kohara J, et al. Mapping of CD4(+) T-cell epitopes in bovine leukemia virus from five cattle with differential susceptibilities to bovine leukemia virus disease progression. Virol J. BioMed Central; 2019;16:157.
80. Miyasaka T, Takeshima S, Matsumoto Y, Kobayashi N, Matsuhashi T, Miyazaki Y, et al. The diversity of bovine MHC class II DRB3 and DQA1 alleles in different herds of Japanese Black and Holstein cattle in Japan. Gene. 2011;472:42–9.
81. TAKESHIMA S-N, AIDA Y. Structure, function and disease susceptibility of the bovine major

histocompatibility complex. Anim Sci J. John Wiley & Sons, Ltd; 2006;77:138–50.
82. Kosciuczuk EM, Lisowski P, Jarczak J, Majewska A, Rzewuska M, Zwierzchowski L, et al. Transcriptome profiling of Staphylococci-infected cow mammary gland parenchyma. BMC Vet Res. BioMed Central; 2017;13:161.
83. Hou Q, Huang J, Ju Z, Li Q, Li L, Wang C, et al. Identification of Splice Variants, Targeted MicroRNAs and Functional Single Nucleotide Polymorphisms of the BOLA-DQA2 Gene in Dairy Cattle. DNA Cell Biol. Mary Ann Liebert, Inc., publishers; 2011;31:739–44.

4. Conclusion and Perspectives

Through the methodology used in this work, it was possible to identify potential functional variants, genes and biological pathways that regulate feed efficiency (FE) using RNA-seq data from multiple central tissues, such as adrenal, muscle, liver, hypothalamus and pituitary of Nellore cattle phenotyped for FE. Although the methodology has proved to be effective, there is still a long way to go for genetic evaluation of the variants / genes / biological pathways and then they will be added to genetic improvement programs.

As future work, it is also intended to complete the differential splicing analyzes associated with feed efficiency and identification of variant sequences associated with splicing events (SVASE) in cattle, with the aim of exploring alternative splicing events, which generate protein diversity.

APPENDIX A – SUPPLEMENTARY MATERIAL OF CHAPTER 1

Supplementary Table 1: Information on potential functional variants associated with feed efficiency

Description: Information on the high and moderate impact variants associated with the feed efficiency phenotype in Nellore cattle. This table also contains information on genomic positions, genes, transcripts, consequences of variants and produced amino acids.

Supplementary Table 2: List of genes enriched for the significant pathways found from potential functional variants associated with Feed Efficiency.

Description: Nesta tabela contem as vias biológicas

Pathways	Ensembl code	Gene symbol	Gene name
Regulation of	ENSBTAG00000010520	*C8G*	Complement C8 gamma chain

Complement Cascade	ENSBTAG00000038171	*CFHR5*	Complement factor H-related 5
And	ENSBTAG00000023177	*CFH*	Complement factor H
Complement Cascade pathway	ENSBTAG00000048135	*IGHG*	A member of immunoglobulin heavy constant gamma
	ENSBTAG00000039995	*CFH*	Complement factor H
	ENSBTAG00000032656	*CPN2*	Carboxypeptidase N subunit 2
	ENSBTAG00000018422	*RIPK3*	Receptor interacting serine/threonine kinase 3
	ENSBTAG00000002472	*CAS9*	Caspase-9
	ENSBTAG00000008959	*BOLA*	Class I histocompatibility antigen alpha chain precursor MHC class I antigen
	ENSBTAG00000010520	*C8G*	Complement C8 gamma chain
	ENSBTAG00000004043	*LOC407171*	Fc gamma 2 receptor
	ENSBTAG00000017313	*ADA2*	Adenosine deaminase 2
	ENSBTAG00000038171	*CFH5*	Complement factor H-related 5
	ENSBTAG00000007213	*SIRPA*	Tyrosine-protein phosphatase non-receptor type substrate 1
	ENSBTAG00000006378	*RIPK1*	Receptor (TNFRSF)-interacting serine-threonine kinase 1
Innate Immune System	ENSBTAG00000023177	*CFH*	Complement factor H
	ENSBTAG00000048135	*IGHG*	Member of immunoglobulin heavy constant gamma

ENSBTAG00000006859	*CEACAM*	Carcinoembryonic antigen-related cell adhesion molecule
ENSBTAG00000039520	*SIRPB1*	Signal-regulatory protein beta 1
ENSBTAG00000026944	*LILRA4*	Immunoglobulin receptor precursor receptor
ENSBTAG00000039995	*CFH*	Complement factor H
ENSBTAG00000002902	*ANO6*	Anoctamin
ENSBTAG00000015254	*GLB1*	Beta-galactosidase
ENSBTAG00000006556	*COPB1*	Coatomer subunit beta
ENSBTAG00000032656	*CPN2*	Carboxypeptidase N subunit 2
ENSBTAG00000015049	*ECSIT*	Evolutionarily conserved signaling intermediate in Toll pathway

Supplementary Table 3: Potential functional variants found in genes enriched for the significant pathways related to immune system.

Description: Enriched functional variants for significant pathways related to the immune system. In this table contains the gene, the name of the variant, the SNP, protein position, altered protein, allele frequency in animals of low and high feed efficiency, p value and the function of the region.

APPENDIX B – SUPPLEMENTARY MATERIAL OF CHAPTER 2

Supplementary Table 1: Mean and standard deviation (SD) of tissue alignment results

Description: Quantities and standard deviation of reads by fabrics, obtained through alignment. In the table we have the initial number, the percentage of reads mapped only, reads mapped to multiple loci, reads not mapped and finally the percentage of alignment coverage.

Tissue	Number of initial reads*	% Reads mapped only	% Reads mapped to multiple loci	% Reads unmapped	% Reads unmapped: other	% Coverage
Adrenal	14,133,167 ± 1,457,849	82 ± 4	4 ± 0,3	14 ± 4	0,03 ± 0,01	86 ± 4
Hypothalamus	13,309,273 ± 1,905,237	81 ± 5	2 ± 0,1	16 ± 5	0,07 ± 0,02	84 ± 5
Muscle	13,205,295 ± 1,045,274	87 ± 3	2 ± 0,2	11 ± 3	0,08 ± 0,01	89 ± 3
Liver	13,332,857 ± 1,149,619	69 ± 16	6 ± 2	25 ± 15	0,04 ± 0,02	75 ± 15
Pituitary	12,512,647 ± 1,440,560	91 ± 1	3 ± 0,1	6 ± 1	0,06 ± 0,01	94 ± 1

* these values are not in percentage; they are in number of reads

Supplementary Table S2: Multi table

Description: The file has several tables, in which it contains information by tissue (adrenal, hypothalamus, muscle, pituitary, liver), about the genes that were included in the functional enrichment and the tabulated result of this enrichment. It also has a table with the significant genes and finally a table with information on SNP, protein position, allelic frequency of the groups, among other information of the variants.