UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

CAUSAL LEARNING TECHNIQUES USING MULTI-OMICS DATA FOR CARCASS AND MEAT QUALITY TRAITS IN NELORE CATTLE

Tiago Bresolin

Zootecnista

2019

UNIVERSIDADE ESTADUAL PAULISTA - UNESP

CÂMPUS DE JABOTICABAL

CAUSAL LEARNING TECHNIQUES USING MULTI-OMICS DATA FOR CARCASS AND MEAT QUALITY TRAITS IN NELORE CATTLE

Tiago Bresolin

Orientadora: Profa. Dr. Lucia Galvão de Albuquerque
Coorientador: Dr. Roberto Carvalheiro

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias - Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Genética e Melhoramento Animal

2019

B842c

Bresolin, Tiago

Causal learning techniques using multi-omics data for carcass and meat quality traits in Nelore cattle / Tiago Bresolin. -- Jaboticabal, 2019

58 p. : il., tabs.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientadora: Lucia Galvão de Albuquerque

Coorientador: Roberto Carvalheiro

1. Beef cattle. 2. Bayesian networks. 3. Graphical models. 4. Quantitative trait loci. 5. Structural equation model. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

unesp

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: CAUSAL LEARNING TECHNIQUES USING MULTI-OMICS DATA FOR CARCASS AND MEAT QUALITY TRAITS IN NELORE CATTLE

AUTOR: TIAGO BRESOLIN
ORIENTADORA: LUCIA GALVÃO DE ALBUQUERQUE
COORIENTADOR: ROBERTO CARVALHEIRO

Aprovado como parte das exigências para obtenção do Título de Doutor em GENÉTICA E MELHORAMENTO ANIMAL, pela Comissão Examinadora:

Profa. Dra. LUCIA GALVÃO DE ALBUQUERQUE
Departamento de Zootecnia / FCAV / Unesp - Jaboticabal

Prof. Dr. FABYANO FONSECA E SILVA (Videoconferência)
Departamento de Zootecnia-UFV / Viçosa/MG

Prof. Dr. HENRIQUE NUNES DE OLIVEIRA
Departamento de Zootecnia / FCAV / Unesp - Jaboticabal

Maria Eugênia Z. Mercadante

Pesquisadora Dra. MARIA EUGÊNIA ZERLOTTI MERCADANTE
Instituto de Zootecnia / Sertãozinho/SP

Pós-doutoranda LARISSA FERNANDA SIMIELLI FONSECA
Departamento de Zootecnia / FCAV / UNESP - Jaboticabal

Jaboticabal, 18 de julho de 2019

CURRICULUM INFORMATION (RESUME)

Tiago Bresolin, was born in Caxambu do Sul - SC, on January 06th 1987, child of Divaldino João Bresolin and Rilde Pagliari Bresolin. He started his undergraduation course in Animal Science at Santa Catarina State University (UDESC), Chapecó - SC on February 2009 and he received a Bachelor's Degree in Animal Science on July 2013 at Santa Maria Federal University (UFSM), Santa Maria - RS. During the undergraduate studies he had involvement in scientific projects as internship supported by grants from CNPq, under supervision of Prof. Dr. Paulo Roberto Nogara Rorato. In August 2013 he started a Master within the postgraduate program on Genetics and Animal Breeding at School of Agricultural and Veterinarian Science/ São Paulo State University (FCAV/ Unesp), when he received financial support from FAPESP. From October 2014 to March 2015, he conducted part of his research project at University of Wisconsin (UW), Madison - WI, under supervision of Prof. Dr. Guilherme Jordão de Magalhães Rosa. His Master thesis was focused on genomic selection in Nelore cattle. In July 2015, he received the Master degree in Genetics and Animal Breeding, under advice of Prof. Dr. Lucia Galvão de Albuquerque and Dr. Roberto Carvalheiro. In the same year he started a Doctorate in Genetics and Animal Breeding at FCAV/Unesp, receiving financial support from FAPESP. During his Doctorate (from March 2018 to February 2019) he conducted part of his research project at University of Wisconsin (UW), Madison - WI, under supervision of Prof. Dr. Guilherme Jordão de Magalhães Rosa. In July 2019, he defended his Doctorate thesis, under advice of Prof. Dr. Lucia Galvão de Albuquerque and Dr. Roberto Carvalheiro.

"Development of Western science is based on two great achievements: the invention of the formal logical system (in Euclidean geometry) by the Greek philosophers, and the discovery of the possibility to find out causal relationships by systematic experiment (during the Renaissance)."

By Albert Einstein (1953)

To my dear parents, Divaldino and Rilde.

To Camila, for the partnership and dedication.

To my siter Sinara.

To all those who helped me on the hike until here.

I dedicate and offer.

ACKNOWLEDGMENTS

I thank God for the miracle of life, and for the strength to carry on.

I would like to express my gratitude to Prof. Dr. Lucia Galvão de Albuquerque and to Dr. Roberto Carvalheiro, for their guidance, friendship, support, advice, encouragement and for every contribution to my personal and professional development.

I would like to express my gratitude to Prof. Dr. Guilherme Jordão de Magalhães Rosa for his support, advice and for every contribution to my personal and professional development.

I would like to thank all internal and external members of the examining committee: Dr. Danísio Prado Munari, Dr. Henrique Nunes de Oliveira, Dr. Fernando Sebastián Baldi Rey, Dr. Maria Eugênia Zerlotti Mercadante, Dr. Fabyano Fonseca e Silva, and Dr. Larissa Fernanda Simielli Fonseca for their kind attention and for their valuable suggestions and contributions to improving this thesis.

I thank my friends from Albuquerque‘s lab for many remarkable moments throughout these years. I especially thank Diogo, Gordo, Espigolan, Anderson, Andre, and Andres for amazing moments together.

I thank the friends I made in Madison at the University of Wisconsin in special Tiago, Renata, Arthur, Anna, João, Ligia, and Isabela.

I would like to thank the financial support from the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), grants #2016/02366-0 and #2017/18779-5 are gratefully acknowledged.

The School of Agricultural and Veterinarian Science/ São Paulo State University (FCAV/ Unesp) is acknowledged for all support provided to the development of this thesis as well as for every contribution to my professional development.

The support and incentive of the postgraduate program on Genetics and Animal Breeding at FCAV/ Unesp are gratefully acknowledged.

The National Council for Scientific and Technological Development (CNPq) for supporting the present study.

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

CONTENTS

TÉCNICAS DE APRENDIZAGEM CAUSAL UTILIZANDO DADOS MULTI-ÔTICOS PARA CARACTERÍSTICAS DE CARCAÇA E QUALIDADE DE CARNE EM BOVINOS DA RAÇA NELORE

RESUMO - Registros de características quantitativas e informações genotípicas coletadas para cada animal são utilizados para identificar regiões do genoma associadas à variação fenotípica. No entanto, essas investigações são, geralmente, realizadas com base em testes estatísticos de correlação ou associação, que não implicam em causalidade. A fim de explorar amplamente essas informações, métodos poderosos de inferência causal foram desenvolvidos para estimar os efeitos causais entre as variáveis estudadas. Apesar do progresso significativo neste campo, inferir os efeitos causais entre variáveis aleatórias contínuas ainda é um desafio e poucos estudos têm explorado as relações causais em genética quantitativa e no melhoramento animal. Neste contexto, dois estudos foram realizados com os seguintes objetivos: 1) Buscar as relações causais entre as características de carcaça e qualidade de carne usando um modelo de equação estrutural (MEE), sob modelo linear misto em bovinos da raça Nelore, e 2) Reconstruir redes de genes-fenótipos e realizar análise de rede causal por meio da integração de dados fenotípicos, genotípicos e transcriptômicos em bovinos da raça Nelore. Para o primeiro estudo, um total de 4.479 animais com informação fenotípica para o peso da carcaça quente (PCQ), área de olho lombo (AOL), espessura de gordura subcutânea (EGS), força de cisalhamento (FC) e marmoreio (MAR) foram usados. Os animais foram genotipados usando os painéis BovineHD BeadChip e GeneSeek Genomic Profiler Indicus HD - GGP75Ki. Para inferência causal usando MEE, uma metodologia de múltiplos passos foi utilizada: a) um modelo multicaracteristica padrão, considerando as características estudadas, foi ajustado e as (co)variâncias residuais *a posteriori* foram estimadas, b) o algoritmo "Inductive Causation" (IC) foi utilizado para inferir as estruturas causais entre as caracteríticas usando as (co)variância residuais *a posterior*, e c) a partir da estrutura causal recuperada pelo algoritmo IC, o MEE foi ajustado. Aplicando intervalo de maior densidade a posteriori (HPD) de 95 %, 90 % e 85 %, as mesmas estruturas causais entre as características foram detectados pelo algoritmo IC, com links não direcionados entre EGS com PCQ e MAR. Ligação extra entre FC e PCQ e a direção entre EGS e PCQ foram identificados usando intervalo de HPD menor (80 %), enquanto que o link entre EGS e MAR permaneceram estatisticamente sem direção. Dois MEE diferentes foram ajustados com base na rede causal recuperada pelo algoritmo IC, com a seta EGS → MAR ou com a seta EGS ← MAR. O MEE que melhor se ajustou compreende as seguintes ligações entre características: FC → AOL, FC → PCQ, PCQ → AOL, EGS → PCQ e EGS → MAR com coeficientes estruturais *a posteriori* igual a -0,29, 0,43, 0,10, 1,92 e 0,03, respectivamente. O MEE final revelou relações causais entre as características, e os efeitos causais sugerem que intervenções em FC e no EGS afetariam diretamente o PCQ e a MAR. Para o segundo estudo, um total de 4.599 animais com informações fenotípicas (AOL, EGS e FC) e genotípicas (como descrito anteriormente) foi utilizados. O sequenciamento do RNA (RNA-Seq) para 80 amostras de tecido muscular de animais da raça Nelore foi reali-zado pelo sistema Illumina HiSeq

2500 produzindo leituras pared-end de 2x100 pares de bases usando amostra de tecido muscular. Redes de gene-fenótipo e análise de rede causal foram realizadas usando uma abordagem de três passos: a) análises de varredura do genôma para identificar a associação entre dados genotípicos e fenotípicos (pQTL - mapeamento de locos de características quantitativas fenotípicas) e entre dados genotípicos e de expressão gênica (eQTL - mapeamento de locos de características quantitativas de expressão). Os efeitos dos marcadores estimados em cada mapeamento de pQTL para os fenótipos estudados (AOL, EGS e FC) foram usados para realizar uma análise multicaracteristica. b) regiões significativas para os dois mapeamentos de QTL (multicaracteristica e eQTL) foram co-localizadas, e c) a reconstrução da rede usando um algoritmo de aprendizado estrutural causal considerando AOL, EGS, FC, eQTL e características de expressão gênica foi realizada. A partir da análise multi-característica, 14 regiões do genoma foram associadas significativamente com AOL, EGS e FC e 19 *cis*-eQTL estavam sobrepondo cinco das regiões do genoma. Com base na posição *cis*-eQTL (a mais significativa em cada região do genoma), trinta e dois genes próximos foram identificados. Integrando dados fenotípicos, genotípicos e de expressão gênica a rede inferida indicou que o *rs137704711*, localizado no cromossomo 20, afetou os três fenótipos (AOL, EGS e FC), e o *rs133894950*, localizado no cromossomo 16, afetou o EGS por meio da expressão de vários genes localizados em diferentes cromossomos. As inferências causais realizadas utilizando diferentes metodologias foram capazes de identificar relações causais entre as variáveis em estudo.

Palavras chaves: bovinos de corte, características de carcaça, características de qualidade de carne, locos de características quantitativas, modelos de equações estruturais, modelos gráficos, redes Bayesianas.

CAUSAL LEARNING TECHNIQUES USING MULTI-OMICS DATA FOR CARCASS AND MEAT QUALITY TRAITS IN NELORE CATTLE

ABSTRACT - Quantitative traits and genotypes information have been collected for each animal and used to identify genome regions related to phenotypes variation. However, these investigations are, usually, performed based on correlation or association statistical tests, which do not imply in causation. In order to fully explore these informations, powerful causal inference methods have been developed to estimate causal effects among the variables under study. Despite significant progress in this field infer causal effect among random variables remains a challenge and some few studies have explored causal relationships in quantitative genetics and animal breeding. In this context, two studies were performed with the following objectives: 1) Search for the causal relationship among carcass yield and meat quality traits using a structural equation model (SEM), under linear mixed model context in Nelore cattle, and 2) Reconstruct gene-phenotype networks and perform causal network analysis through the integrating of phenotypic, genotypic, and transcriptomic data in Nelore cattle. For the first study, a total of 4,479 animals with phenotypic information for hot carcass weight (HCW), longissimus muscle area (LMA), backfat thickness (BF), Warner-Bratzler shear force (WBSF), and marbling score (MB) traits were used. Animals were genotyped using BovineHD BeadChip and GeneSeek Genomic Profiler Indicus HD - GGP75Ki. For causal inference using SEM a multistep procedure methodology was used as follow: a) a standard multi-trait model for studied traits was fitted to access the posterior residual (co)variances, b) the Inductive Causation (IC) algorithm was used to infer causal structures between traits using the posterior residual (co)variances, and c) from the selected causal structure retrieved by the IC algorithm the SEM was fitted. Applying 95 %, 90 % and 85 % highest posterior density (HPD) the same graph was detected by the IC algorithm with undirected links between BF with HCW and MB. Extra link between WBSF and HCW and the direction between BF and HWC were identified using narrow HPD interval (80 %), whereas the link between BF and MB remained undirected. Two different SEM were fitted based on the causal network retrieved by the IC algorithm with either arrow BF → MB or BF ← MB. The most feasible SEM comprise the following links between traits: WBSF → LMA, WBSF → HCW, HCW → LMA, BF → HCW, and BF → MB, with structural coefficients posterior means equal -0.29, 0.43, 0.10, 1.92, and 0.03, respectively. The final SEM revealed some interesting relationships among the traits, and the causal effects suggest that interventions on WBSF and BF would direct affect HCW and LMA. For the second study, a total of 4,599 animals with phenotypic (LMA, BF, and WBSF) and genotypic (as previously described) information were used. RNA sequen-cing (RNA-Seq) for 80 Nelore cattle muscle tissue samples was carried out by Illumina HiSeq 2500 System to produce 2x100 base pairs paired-end reads using muscle ti-ssue sample. Gene-phenotype networks and causal network analysis were performed using a three-step approach as follow: a) genome scan analyses to identify the association between genotypic and phenotypic data (pQTL – phenotype quantitative trait loci mapping), and between genotypic and gene expression data (eQTL – expression quantitative trait loci mapping). The markers effects estimated in every single pQTL mapping for the phenotypes studied (LMA, BF, and WBSF) were used to perform a multi-trait analysis. b) significant regions from both QTL mapping (multi-trait and eQTL) were co-localized, and c) network reconstruction using causal structural learning algorithm incorporating LMA, BF, WBSF eQTL and gene expression traits was performed. From the multi-trait analysis, 14 genome regions were significant across LMA, BF, and WBSF and 19 *cis*-eQTL were overlapping five of the genome regions. Based on the *cis*-eQTL position (the most significant in each genome region), thirty-two nearby genes were identified. Integrating phenotypes, genotypes and gene expression data the inferred network indicated that the *rs137704711*, located in chromosome 20, affected the three phenotypes (LMA, BF, and WBSF), and the *rs133894950*, located in chromosome 16, affected BF through the expression of several genes located in different chromosomes. The causal inferences performed using different methodologies were able to identify important causal relationships among the variables under study.

Key words: beef cattle, Bayesian networks, carcass traits, graphical models, meat quality traits, quantitative trait loci, structural equation models.

CHAPTER 1 - GENERAL CONSIDERATIONS

1.1 INTRODUCTION

The whole bovine genome sequencing and advances in technologies have enabled high-density bovine genotyping. High-density arrays of SNP markers (single nucleotide polymorphism) made possible to identify several genome regions, also termed as phenotype quantitative trait loci (pQTL), for carcass and meat quality traits through genome-wide association studies (GWAS) in different breeds (KIM et al., 2011; LU et al., 2013; MAGALHÃES et al., 2016; FERNANDES JÚNIOR et al., 2016; SANTIAGO et al., 2017; HAY; ROBERTS, 2018). However, the identified pQTL for complex traits only account for a small portion of phenotypic variation (MANOLIO et al., 2009) and they are not necessarily true causal variants (AINSWORTH; SHIN; CORDELL, 2017). In addition, the majority pQTL identified were found to reside in non-coding regions (intergenic and intronic) of the genome (NICA; DERMITZAKIS, 2013). One explanation for this pQTL trait association is that such pQTL modify cis-regulatory sequences and thereby change the expression levels of one or more target genes (INNOCENTI et al., 2011).

Gene expression variation is abundant in all organisms and plays essential roles in several important processes responsible for the phenotypic variability (GILAD; SCOTT; JONATHAN, 2008; INNOCENTI et al., 2011). In order to understand this systematic process, it is important to expand the type of traits studied (PEÑA-GARICANO et al., 2015). RNA expression (RNAseq) at the population level, accessed by high-throughput DNA sequencing technology, is one of such type of trait, providing nucleotide-level resolution of gene expression across the entire transcriptome. Gene expression traits combined with genotype data made possible to identify thousands of expression quantitative trait loci (eQTL) through eQTL mapping (BOUWMAN et al., 2018; CESAR et al., 2018; HIGGINS et al., 2018). In summary, eQTL mapping enables to investigate the effect of genotype on gene-expression levels which may affect phenotypes (NICA; DERMITZAKIS, 2013). Moreover, by combining both QTL mapping might unravel genetic architecture of the traits and gene network (GILAD; SCOTT; JONATHAN, 2008; HUANG; ZHENG; PRZYTYCKA, 2010).

Even with the availability of all these biological information and computational resources, causal relationships among variables have not been widely explored (PEÑA-GARICANO et al., 2015; BADSHA; FU, 2019). To infer causation is required a randomized experimental design (FISHER, 1926). However, the randomization of alleles during meiosis (Mendelian randomization) ensures the unidirectional influence of genotype on phenotype (HAGEMAN et al., 2011; ROSA et al., 2011). Mendelian randomization is analogous to a randomized experimental design, providing a set used to infer causality using Fisher's statistical framework (ROSA et al., 2011). In this context, different approaches have been proposed for inferring causal relations using multiple phenotypes and genotypes information, including structural equation models (GIANOLA; SORENSEN, 2004; VALENTE et al., 2010) and Bayesian Networks (PEARL, 1988). The objective of this study was to search for causal network underlying carcass and meat quality traits in Nelore cattle, applying causal learning techniques using phenotypic, genotypic and transcriptomic data.

1.2 LITERATURE REVIEW

1.2.1 Linear mixed models

Linear mixed models (LMM) are an extension of traditional linear models combining fixed and random effects modeled jointly (LAIRD; WARE, 1982). For a single trait LMM can be presented as:

$$y = X\beta + Zu + e \tag{1.1}$$

where *y* is a *n x 1* vector of observations (*n* is the number of observations), β is a *p x 1* vector of fixed parameters (*p* is the number of fixed parameters), *u* is a *q x 1* vector of unknown random effects, *X* and *Z* are known incidence matrices with dimension *n x p* and *n x q* related to β and *u*, respectively, and e is a *n x 1* vector of residual terms. For LMM, usually is assumed that *u* and *e* are independent and normally distributed with mean zero and variance-covariance matrices equal to *G* and *R*, respectively. The prediction of random effects are given by the conditional expectation of *u* given the data, *E(u | y)*. The joint distribution of *y* and *u* is a multivariate normal such as:

$$\begin{bmatrix} y \\ u \end{bmatrix} = N\left(\begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} V & ZG \\ GZ' & G \end{bmatrix}\right), \tag{1.2}$$

where *V = ZGZ ′ + R*, and following multivariate normal distributions properties, *E(u | y)* is given by:

$$\begin{aligned} E(u \mid y) &= E(u) + Cov(u, y')Var^{-1}(y)(y - E(y)), \\ &= Cov(u, y')Var^{-1}(y)(y - E(y)), \\ &= GZ'V^{-1}(y - X\beta). \end{aligned} \tag{1.3}$$

Assuming that *G* and *R* are known and *u* and *e* are normally distributed, the density of the distribution of *y* is given by:

$$f(y;\theta) = \frac{1}{(2\pi)^{n/2}|ZGZ' + R|^{1/2}} \exp\left\{\frac{1}{2}[(y - X\beta)'(ZGZ' + R^{-1})(y - X\beta)]\right\}, \tag{1.4}$$

where θ is the vector of parameters (*u*, β and *G*) and the joint probability density function, *f(y,u) = f(y | u) f(u)* is given by:

$$\begin{aligned} f(y \mid u) = & \frac{1}{(2\pi)^{n/2}|R|^{1/2}} \exp\left\{\frac{1}{2}[(y - X\beta - Zu)'R^{-1}(y - X\beta - Zu)]\right\} \\ x\ & \frac{1}{(2\pi)^{q/2}|G|^{1/2}} \exp\left\{\frac{1}{2}u\,'G^{-1}u\right\} \end{aligned} \tag{1.5}$$

The logarithm of equation (1.5) is given by:

$$\begin{aligned} \ell(y,\, u) = & \frac{1}{2}\, 2n\ log(2\pi) - \frac{1}{2}\ (log|R| + log|G|) \\ & - \frac{1}{2}(y\,'R^{-1}y - 2y\,'R^{-1}X\beta - 2y\,'R^{-1}Zu + 2\beta'X'R^{-1}Zu \\ & + \beta'X'R^{-1}X\beta + u\,'Z'R^{-1}Zu + u\,'G^{-1}u) \end{aligned} \tag{1.6}$$

Deriving this equation in β and *u* and equating to *0* yields

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \tag{1.7}$$

From equation (1.7) is possible to obtain the best linear unbiased estimator (BLUE) of $\hat{\beta}$, given by

$$\begin{aligned} \hat{\beta} = & [X'R^{-1}X - X'R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}Z] \\ x\ & [X'R^{-1}y - X'R^{-1}Z(Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}y], \end{aligned} \tag{1.8}$$

and it is also possible to obtain the best linear unbiased predictor (BLUP) of *u* as

$$\hat{u} = (Z'R^{-1}Z + G^{-1})^{-1}Z'R^{-1}(y - X\hat{\beta}). \tag{1.9}$$

LMM are widely used in quantitative genetics to predict additive genetic effects and estimating genetic parameters, take into account environmental and genetic information (HENDERSON et al., 1959). In this context, animal model has been successfully used in quantitative genetics and animal breeding to predict breeding values for all animals, not only for individuals with known phenotypes. For only one phenotype and a single observation per subject, the animal model can be represented as in equation (1.1). However, the vector *y* represents the observations, β the environmental effects (i.e. contemporary groups, age and others), *u* the breeding values and *e* the residual effects, usually assumed independent across individuals. The residual covariance structure can be expressed as R = I σ_e^2 where I is an identity matrix, and σ_e^2 is the residual variance and the covariance among the breeding values is represented by G. The matrix G is considered as A σ_a^2, where A is the additive genetic relationship matrix. Replacing G^{-1} = A^{-1} σ_a^2 and R^{-1} = $I\sigma_e^2$ in equation (1.7), the mixed model equation (MME) is reduced to:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \tag{1.10}$$

where λ = $\frac{\sigma_e^2}{\sigma_a^2}$ = $\frac{1-h^2}{h^2}$. The h^2 represent the proportion of the total phenotypic variance that is due to additive genetic effects. The matrix A^{-1} can be directly constructed from the pedigree information, and therefore inverting the typically large A is not required (HENDERSON et al., 1959; HENDERSON; QUAAS, 1976).

The animal model can be extended to more than one trait per subject (HENDERSON; QUAAS, 1976; SCHAEFFER, 1984). Considering t traits for each subject as an example, equation (1.1) can be rewritten as:

$$y_k = X_k\beta_k + Z_k u_k + e_k \tag{1.11}$$

where y_k, X_k, β_k, Z_k, u_k and e_k follows the same definitions previously described and k is the index for each trait k = *1, 2, . . . , t*. The LMM that jointly accounts for the t traits is given by:

$$y = X\beta + Zu + e \tag{1.12}$$

where y = [y_1', y_2', ..., y_t'], β = [β_1', β_2', ..., β_t'], u = [u_1', u_2', ..., u_t'] and e = [e_1', e_2', ..., e_t']. The X and Z incidences matrix are:

$$X = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_t \end{bmatrix} \; and \; Z = \begin{bmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_t \end{bmatrix}.$$

In addition, it is assumed that the variance of u and e are:

$$Var\begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G_0 \otimes A & 0 \\ 0 & E \otimes I \end{bmatrix}, \tag{1.13}$$

where G and E are the genetic and residual variance-covariance matrices, respectively.

1.2.2 Causal inference

Statistic framework has been widely used in different fields to make inferences from observational data. Associations among variables, for example, can be inferred using some type of statistical approach, which help researchers to make conclusions about how much the variables under study are connected (VALENTE et al., 2013). These associations are quite often based on correlation and have been efficiently performed by standard statistical models (PEÑAGARICANO et al., 2015). Even though two variables (X and Y), are strongly correlated, it does not express the causal effect of X on Y, for example. This is a well-known proverb concerning scientific inquires and statistics states that "correlation does not imply causation" (ROSA; VALENTE, 2013). Causation goes deeper and the main goal of causal analysis is to infer probabilities under conditions that are changing, such as external interventions (PEARL, 2010).

Sewall Wright (WRIGHT, 1921) was the pioneer inferring causality thought path analysis. After the work of Wright, causation has not been longer explored by researcher until the year 2000, when Judea Pearl (PEARL, 2000) returned the concept of causality in the scientific community (BARROWMAN, 2014). Within the last decades, with development of computers, tremendous progress has been made in developing statistical methods and efficient algorithms for causal inference (WIEDERMANN; DONG; vonEye, 2019). Causality has been inferred in many different ways for economist, epidemiologist, biologist and other sciences (SPIRTES, 2010; BARROWMAN, 2014). In livestock, not different of others areas, there are many scenarios in which the central goal is to investigate causal effects.

Although many statistical methods have been developed, infer causal effects is a challenge because the observed association between a causal variable and an outcome can be due to background confounding factors, do not reflecting causal effects (PEARL, 2010). Randomized experiments (FISHER, 1926) are a powerful approach for dealing with potential confounders (ROSA; VALENTE, 2013). Such controlled experiments impose equal probability of causal variable levels in which subjects are randomly distributed into experimental group, allowing not only test the treatment effects but also estimate their magnitude (FISHER, 1971). However, randomized experiments may be not always feasible, legal, ethical, or logistical constraints (ROSENBAUM, 2010), and, therefore, in such cases, an alternative is to investigate causality from observational data (ROSA; VALENTE, 2013).

Farmers are routinely collecting data for breeding purposes such as phenotypic, environmental, and management variables. From this data we can explore for example the causal effects of environmental and management factors in livestock production as well as how phenotypes are connected with each other (ROSA; VALENTE, 2013). Correction for confounding factors can be achieved by using appropriate statistical methods, and, therefore observational data from farmers might be explored for causal inference (ROSA; VALENTE, 2013). In quantitative genetics many researches have explored causal effect among different phenotypes and genomic information for different species (de los CAMPOS; GIANOLA; HERINGSTAD, 2006; de los CAMPOS et al., 2006; VARONA; SORENSEN; THOMPSON, 2007; WU; HERINGSTAD; GIANOLA, 2008; HERINGSTAD; WU; GIANOLA, 2009; MATURANA et al., 2009; WU; HERINGSTAD; GIANOLA, 2010; VALENTE et al., 2011; INOUE et al., 2016; INOUE; HOSONO; TANIMOTO, 2017; PEÑAGARICANO et al., 2015). Within the methodology available for causal inference, structure equation model and Bayesian network are the most used in the literature.

1.2.2.1 Structural equation model

Structure equation models (SEM) is as extensions of the standard multiple-trait models (MTM) proposed by Henderson et al. (1959), providing a general statistical modeling technique to account for causal associations between variables, which are often not revealed by MTM (WRIGHT, 1921; HAAVELMO, 1943; ROSA et al., 2011). SEM was described in quantitative genetics context by Gianola e Sorensen (2004) to account for possible feedback or recursive relations among traits. Following the work of Gianola e Sorensen (2004), SEM has been applied to different species in multi-trait mixed models settings (de los CAMPOS; GIANOLA; HERINGSTAD, 2006; de los CAMPOS et al., 2006; VARONA; SORENSEN; THOMPSON, 2007; WU; HERINGSTAD; GIANOLA, 2008; HERINGSTAD; WU; GIANOLA, 2009; MATURANA et al., 2009).

To fit a SEM, causal structure coefficients describing qualitatively the causal relationships between variables must be specified a priori. Causal structure is a subset of traits with causal influence on each phenotype under study (VALENTE et al., 2010). The causal structure can be represented as a directed graph, in which nodes are the traits studied and directed edges between nodes represent causal relationships among traits (PEARL, 2000) as depicted in Figure 1.1.

The example presented in Figure 1.1 can be written as a set of structural equations given by:

$$\begin{cases} y_1 = \beta_1 x_1 + e_1 \\ y_2 = \lambda_{21} y_1 + \beta_2 x_2 + e_2 \\ y_3 = \lambda_{31} y_1 + \lambda_{32} y_2 + \beta_3 x_3 + e_3 \end{cases} \tag{1.14}$$

where β's represent the parameters for explanatory variables (fixed effects) and λ's are

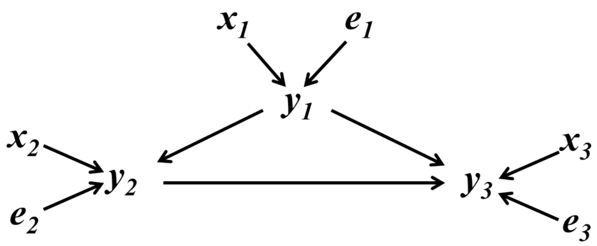

Figure 1.1 – Example of a causal structure for three traits (y's), where x's and e's represent known explanatory variables and residual factors affecting phenotypic traits, respectively (Adapted from Rosa et al. (2011)).

structural coefficients representing the magnitude of the casual effects among *y*'s. The set of structural equations (SEM) can be represented in matrix notation as:

$$y = \Lambda y + X\beta + e \tag{1.15}$$

where *y* is a vector with observations, Λ is a matrix with the coefficients λ, *X* are appropriate matrix with the explanatory variables, β is a vector with model parameters, *e* is a vector of residuals. Extending the equation in (1.15) for the quantitative genetics context, SEM with causal structure and random additive genetic effects for *t* traits (GIANOLA; SORENSEN, 2004) can be written as:

$$y_i = \Lambda y_i + X_i\beta + u_i + e_i \tag{1.16}$$

where y_i is a *(t x 1)* vector of phenotypic observations on subject *i*; Λ is a *(t x t)* matrix with zeroes on the diagonal and with structural coefficients or zeroes on the off-diagonal (the causal structure defines which entries contain free parameters and which entries are constrained to 0); β is a vector of fixed regression coefficient(s), X_i contains the covariates for the i^{th} subject, u_i is a *(t x 1)* vector of random additive genetic effects for the ith subject, and e_i is a *(t x 1)* vector of model residuals for the ith subject. For *n* animals the SEM in (1.16) can be rewritten as:

$$y = (\Lambda \otimes I_n)y + X\beta + Zu + e \tag{1.17}$$

where *y*, *u* and *e* are vectors of phenotypic traits, and additive genetic and residuals effects for *t* traits, sorted by trait and subject within trait, β is a vector containing the fixed effects, and *X* and *Z* are incidence matrices relating effects in β and *u* to *y*. For *u* and *e* are assumed:

$$\begin{bmatrix} u \\ e \end{bmatrix} \sim N\left\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G_0 \otimes A & 0 \\ 0 & \Psi_0 \otimes I_n \end{bmatrix}\right\} \tag{1.18}$$

where G_0 and Ψ_0 are the additive genetic and residual variance-covariance matrices, respectively. Rewritten the equation in (1.17) (VALENTE et al., 2010) the equivalent reduced model as in Gianola e Sorensen (2004) can be obtained as follow:

$$\begin{aligned}[I_{tn}(\Lambda \otimes I_{tn})]y &= X\beta + Zu + e \\ &= [I_{tn} - (\Lambda \otimes I_{tn})]^{-1}X\beta + [I_{tn} - (\Lambda \otimes I_{tn})]^{-1}Zu + [I_{tn} - (\Lambda \otimes I_{tn})]^{-1}e.\end{aligned} \tag{1.19}$$

The resulting sampling distribution of y given the location parameters and the residual covariance matrix is:

$$p(y|\Lambda, \beta, u, \Psi_0) \sim N[I_{tn} - (\Lambda \otimes I_{tn})]^{-1}(X\beta + Zu), [I_{tn} - (\Lambda \otimes I_{tn})]^{-1}\Psi[I_{tn} - (\Lambda \otimes I_{tn})]'^{-1} \tag{1.20}$$

where $\Psi = \Psi_0 \otimes I_n$. The location and dispersion parameters in the reduced SEM (1.19) are transformed into parameters of a standard MTM (VARONA; SORENSEN; THOMPSON, 2007; WU; HERINGSTAD; GIANOLA, 2010; ROSA et al., 2011) as follow:

$$\begin{aligned}y_i &= (I_t - \Lambda)^{-1}X_i\beta + (I_t - \Lambda)^{-1}u_i + (I_t - \Lambda)^{-1}e \\ &= \mu_i^* + u_i^* + e_i^*,\end{aligned} \tag{1.21}$$

where $\mu_i^* = (I_t - \Lambda)^{-1} X_i \beta$, $u_i^* = (I_t - \Lambda)^{-1} u_i$ and $e_i^* = (I_t - \Lambda)^{-1} e_i$. The joint distribution of u_i^* and e_i^* is given by:

$$\begin{bmatrix} u \\ e \end{bmatrix} \sim N\left\{\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G_0^* & 0 \\ 0 & R_0^* \end{bmatrix}\right\} \tag{1.22}$$

with $G_0^* = (I_t - \Lambda)^{-1} G_0 (I_t - \Lambda)'^{-1}$ and $R_0^* = (I_t - \Lambda)^{-1} \Psi_0 (I_t - \Lambda)'^{-1}$. The vectors μ_i^*, u_i^* and e_i^* are the fixed, additive genetic and residuals effects, respectively, and G_0^* and R_0^* are respectively the genetic and residual covariance matrices of an MTM (ROSA et al., 2011). As showed in equation (1.17), SEM has additional parameters in Λ, which result in an unidentifiable likelihood function (ROSA et al., 2011). In order to achieve the identifiable likelihood function it is possible to introduce constraints in SEM through coercing the residual covariance matrix Ψ_0 to be diagonal (WU; HERINGSTAD; GIANOLA, 2010; ROSA et al., 2011).

Even fitting SEM with few traits, the space of possible causal hypotheses is typically very large (SHIPLEY, 2002). The authors that followed the work of Gianola e Sorensen (2004), pre-selecting the causal structure based on biological knowledge and not exploring the full space of possible structures (VALENTE et al., 2010; VALENTE et al., 2011). In order to overcome the problem of causal structure selection, Valente et al.

(2010) adapted the inductive causation (IC) algorithm (VERMA; PEARL, 1990; PEARL, 2000) to mixed-models scenarios. The IC algorithm uses the notion of *d-separation* to explorer the space of causal hypotheses so as to arrive to a causal structure (PEARL, 2000). Based on a given correlation matrix, this algorithm perform conditional independence tests between variables, which return a partially oriented graph as output used as *a prior* in SEM (ROSA et al., 2011). However, the IC algorithm cannot be applied directly to the joint distribution of the phenotypes because genetic correlations, for example, may act as confounders (VALENTE et al., 2010). By this reason the author proposed to select the causal structure based on the residual variance-covariance draw from a MTM (see more details in Valente et al. (2010).

In summary, the overall statistical approach proposed by Valente et al. (2010) to search for a causal structure in a mixed models context, using samples from the posterior distribution of R_0^* as input to the IC algorithm consists of three steps:

- Fit a Bayesian MTM in order to obtain posterior samples of R_0^*.
- The IC algorithm is applied on R_0^* matrix to make the statistical decisions required. Specifically, for each query about the statistical independence between variables *a* and *b* given a set of variables *S* and, implicitly, the genetic effects:

  a) Obtain the posterior distribution of residual partial correlation $\rho_{(a,b|S)}$. These partial correlations are functions of R_0^*. Therefore their posterior distribution can be obtained by computing the correlation at each sample drawn from the posterior distribution of R_0^*.

  b) Compute the 95% Highest Posterior Density (HPD) interval for the posterior distribution of $\rho_{(a,b|S)}$.

  c) If the HPD interval contains *0*, declare $\rho_{(a,b|S)}$ as null. Otherwise, declare a and b as conditionally dependent.
- From the selected causal structure fit the SEM as in Gianola e Sorensen (2004), such that causal relationships (i.e., recursive effects) can be estimated.

After select the causal structure and achieving parameter identifiability, inferences about model parameters may be performed applying standard statistical methodologies as proposed by Sorensen e Gianola (2002).

1.2.2.2 Bayesian network

Bayesian network (BN) is an annotated directed acyclic graph (DAG) - arrows must be directed without circles, which combines the rigor of a probabilistic distribution with the intuitive representation of relationships among variables, given by a graph (PEARL, 1988). BN is a class of graphical model composed by two parts: a set *X* = X_1, X_2, ..., X_p of random variables describing the quantities of interest, and a graph *G = (V, E)* where each vertex *v* $\in$ *V*, also called node, is associated with one of the random variables in *X*, and each edge *e* $\in$ *E*, also called arc, is used to express the dependence structure of the data, i.e., dependence relationships among *X*'s. In summary, BN is graphical representation of (in)dependencies among random variables (ZHANG; POOLE, 1996).

The dependence structure of the data expressed by arcs and its graphical representation is given in terms of conditional dependence and graphical separation (PEARL, 1988). Considering the example in Figure 1.2, C and D are the parents of node E, whereas F and G are children due to the directed edges among these nodes. In the example, each node is conditionally independent of its non-descendants given its parents. Probability distributions in BN are represented by the Markov condition (NEAPOLITAN, 2003). The Markov condition is a result descending from the definitions of directed separation or *d-separation* (PEARL, 1988). Based on the Markov condition, the joint probability distribution for all random variables can be decomposed into a product of conditional probabilities (KORB; NICHOLSON, 2010). The chain rule of probability for continuous random variable is

$$\begin{aligned} f_x(x_1, x_2, ..., x_n) = \prod_{i=1}^{p} fx_i(x_i | x_1, x_2, ..., x_{x-1}) \\ f_x(X) = \prod_{i=1}^{p} fx_i(X_i | X_{pa(i))}), \end{aligned} \tag{1.23}$$

where local distribution is associated with a single node X_i and depends only on the joint distribution of its parents $X_{pa(i)}$ in *G*. This decomposition holds for any Bayesian network, regardless of its graph structure.

In the BN context statistic modeling, also called learning, is performed in two different steps, which correspond to model selection (structure learning) and parameters estimation (parameter learning) based on the structure retrieved in the first step (SCUTARI; STRIMMER, 2011). Structure learning consists in to find the graph structure that encodes the conditional independencies present in the data. Parameter learning estimates the parameters of the local distribution, since the network structure is known from the previous step (KOLLER; FRIEDMAN, 2009; KORB; NICHOLSON, 2010). To learn the structure of BN from the data, many algorithms classified as constraint-based and score-based have been proposed. Here only constraint-based structure learning

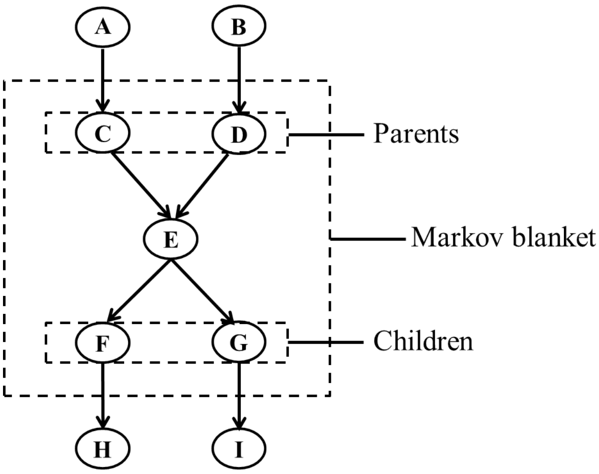

Figure 1.2 – Directed acyclic graph for nine random variables, where C and D represent the parents, F and G are the children, and the Markov blanket of node E (Adapted from Scutari e Strimmer (2011)).

approach will be covered.

Constraint-based algorithms estimate the conditional independence relationships among variables assuming that the graph underlying the probability distribution is able to determine the correct network structure (SCUTARI; STRIMMER, 2011). Based on the observed data this constraint-based algorithms start by determining a skeleton (Markov network) of the underlying network. However, as the number of random variables typically used is large and the independence test is performed for each node, the search space to determine the skeleton has an exponential growth (MOROTA et al., 2012). For this reason Tsamardinos, Aliferis e Statnikov (2003) proposed to impose a constraint in the search space by restricting up to the Markov blanket (MB) of a node. The MB of a node E, denoted by MB(E), is a minimal set of nodes consisting of *E*'s parents and its children (PEARL, 1988) as the example in Figure 1.2.

The knowledge of the MB(E) for example is enough to determine the probability distribution of E, which means that the values of the remaining variables are superfluous (TSAMARDINOS; ALIFERIS; STATNIKOV, 2003). Several algorithms have been proposed to identify the MB including Koller–Sahami (KS) (KOLLER; SAHAMI, 1996), grow–shrink (GS) (MARGARITIS; THRUN, 1999), incremental association Markov blanket (IAMB) (TSAMARDINOS; ALIFERIS; STATNIKOV, 2003) and fast-IAMB (YARAMAKALA; MARGARITIS, 2005). IAMB algorithm has its foundation in the IC algorithm (VERMA; PEARL, 1990) and consists of two phases, a forward and a backward one. The forward or growing phase starts with an empty set and added candidates of MB(E) to a current Markov blanket (CMB) that maximizes a heuristic function (TSAMARDINOS; ALIFERIS; STATNIKOV, 2003). For every variable that is a member of the MB, variable J for example, the heuristic function should return a non-zero value, which is a measure of association between J and E given CMB. According to Tsamardinos, Aliferis e Statnikov (2003) the heuristic function needs to be informative and effective in order to have a small set of candidate variables as possible after growing phase. Using this heuristic function the algorithms do not spend time considering irrelevant variables and do not require sample larger than necessary to perform conditional tests of independence.

Once the algorithm found a variable that is associated with the target node, based on the conditional independence test, it will include this candidate in the CMB and start again from the first variable in the data set. The backward or shrinking phase is a refinement step that removes one-by-one false-positive node from the CMB by a series of conditional independence tests. After identified the MB(E), the algorithm needs to compute the network structure given the Markov blanket determining the nodes in the Markov blanket that are actually direct parents and children of node E (TSAMARDINOS; ALIFERIS; STATNIKOV, 2003). Details about the heuristic function as well as the IAMB algorithms can be found in Tsamardinos, Aliferis e Statnikov (2003) and Morota et al. (2012).

1.3 OBJECTIVES

1.3.1 General objective

The objective of this study was to search for causal network underlying carcass and meat quality traits in Nelore cattle, applying causal learning techniques using phenotypic, genotypic and transcriptomic data.

1.3.2 Specific objectives

- Search for causal relationship among carcass and meat quality traits using structural equation model, under linear mixed model context.
- Reconstruct gene-phenotype networks and perform causal network analysis by integrating phenotypic, genotypic, and transcriptomic data.

CHAPTER 2 - CAUSAL RELATIONSHIPS AMONG CARCASS AND MEAT QUALITY TRAITS USING STRUCTURAL EQUATION MODEL IN NELORE CATTLE

2.1 ABSTRACT

Knowledge regarding any potential causal relationships among carcass and meat quality traits is important to improve Nelore cattle productivity. However, these traits have been studied in terms of linear association, without considering the recursive and simultaneous relationships among them. The objective of this study was to investigate the causal relationships among carcass and meat quality traits using structural equation model, under linear mixed model context, in Nelore cattle. A total of 4,405 animals with phenotypic information for hot carcass weight (HCW), longissimus muscle area (LMA), backfat thickness (BF), Warner-Bratzler shear force (WBSF), and marbling score (MB) traits were used. Causal structures were investigated applying the Inductive Causation (IC) algorithm on the posterior distribution of the residual (co)variance matrix, draw in a standard Bayesian multi-trait model (MTM). Applying 95 %, 90 % and 85 % highest posterior density (HPD) the same graph was detected by the IC algorithm, which included undirected links between BF with HCW and MB. Extra link between WBSF and HCW, and the direction between BF and HWC were identified using HPD interval of 80 %, however, the link between BF and MB remained undirected. Two structural equation models (SEM) were fitted based on the causal network retrieved by the IC algorithm, with either the arrow BF → MB or the arrow BF ← MB. The most feasible SEM comprises the following links between traits: WBSF → LMA, WBSF → HCW, HCW → LMA, BF → HCW, and BF → MB, with structural coefficients posterior means equal to -0.29, 0.43, 0.10, 1.92, and 0.03, respectively. The final SEM revealed causal relationships among the traits, and the causal effects suggest that interventions on WBSF and BF would direct affect HCW and LMA.

Key words: beef cattle, causal effect, inductive causation, structural coefficients

2.2 INTRODUCTION

Nelore cattle (*Bos taurus indicus*), the most important breed in Brazil, present low carcass and meat quality grade compared to *Bos taurus taurus*, affecting the Brazilian beef industry (O'CONNOR et al., 1997; ELZO et al., 2012; CASTRO et al., 2014; PEREIRA et al., 2015). With the consumer markets and beef industry placing more emphasis on these traits (DELGADO et al., 2006; SMITH et al., 2007), studies have been performed to estimate genetic variability and correlations necessary to design a scheme to improve Nelore cattle carcass and meat quality grade (CASTRO et al., 2014; TONUSSI et al., 2015; GORDO et al., 2016; GORDO et al., 2018). But, multiple traits are analyzed through multi-trait mixed model (MTM) (HENDERSON, 1976), allowing to infer only the symmetric linear association among random variables (VALENTE et al., 2010; ROSA et al., 2011). Linear associations are not the only relationship present in biological systems, which can be also recursive and simultaneous and, therefore, these potential causal relationships among traits need to be investigated (ROSA et al., 2011).

Causal associations can be explored by fitting structural equation models (SEM) proposed by Wright (1921) and Haavelmo (1943), and adapted to the quantitative genetics mixed models context by Gianola e Sorensen (2004). More than to identify recursive and simultaneous relationships among phenotypes, SEM also allows predicting the behavior of complex biological systems (GIANOLA; SORENSEN, 2004; ROSA; VALENTE, 2013). Following the work of Gianola e Sorensen (2004), Valente et al. (2010) proposed a methodology to search for recursive causal structures in MTM, allowing to qualitatively describe the causal influence of a subset of phenotypes on each studied trait (VALENTE et al., 2011). This approach has been used by many authors to fit SEM in different species and traits (VALENTE et al., 2011; BOUWMAN et al., 2014; INOUE et al., 2016; INOUE; HOSONO; TANIMOTO, 2017), and they have identified important causal relationships among the studied traits. Thus, the objective of this study was to investigate the causal relationships among carcass and meat quality traits using a structural equation model, under linear mixed model context, in Nelore cattle.

2.3 MATERIAL AND METHODS

All animal procedures were approved by the São Paulo State University (Unesp), School of Agricultural and Veterinary Science Ethical Committee (Approval No. 18.340/16).

2.3.1 Data collection and editing procedure

Phenotype and pedigree information from commercial herds located in the southeast, mid-west and northeast of Brazil were used. The animals (bulls), born between 2008 and 2014, were raised on pasture conditions and finished in feedlot system for around 90 days to be slaughtered at, approximately, 2 years of age in commercial slaughterhouse. Hot carcass weight (HCW), longissimus muscle area (LMA), backfat thickness (BF), Warner-Bratzler shear force (WBSF) and marbling score (MB) traits were studied. Briefly, at slaughter, HCW was recorded for each animal and after 24 to 48 hours chill, from the slaughter, samples from *longissimus thoracis* muscle between 12/13th ribs were collected and frozen at -20 °C. From steaks of 2.54 cm thickness and using a plastic grid (squared with 1 cm^2) placed on the sample, LMA was measured as the sum of all counted squares. Using a caliper, the layer of subcutaneous fat on the steaks was measured and the total of millimeters was used as BF trait. The degree of marbling (MB) was determined, according to the United States Standards for Grades of Carcass Beef (USDA, 1997), as a score on a scale from 1 (practically absent) to 10 (very abundant). For WBSF the steaks were cooked in an oven (180 °C) to an internal temperature of 71 °C. After 24 hours cooling to 2 °C, eight 1.27 mm meat cylinders were obtained from each cooked sample and sheared with a V blade attached to a WBSF machine (WHEELER; KOOHMARAIE; SHACKELFORD, 1995). Average of the eight meat cylinders were used as WBSF trait. Contemporary group (CG) was defined by year and farm of birth, farm and management group at yearling and slaughter date. Observations with three standard deviations above or below to the mean of their CG and CG with less than three animals were removed. A summary of the data structure used is shown in Table 2.1.

Table 2.1 – Descriptive statistical for the traits studied.

Traits	Mean	SD	Min.	Max.
HCW (kg)	279.30	27.74	181.8	374.9
LMA (cm^2)	68.01	7.87	40.0	96.0
BF (mm)	4.75	2.17	1.0	14.0
WBSF (Kg)	6.44	1.90	1.8	11.9
MB (score)	2.82	0.48	1.9	4.8

Hot carcass weight (HCW), longissimus muscle area (LMA), backfat thickness (BF), Warner-Bratzler shear force (WBSF), and marbling score (MB) traits. Standard deviation (SD), minimum (Min), and maximum (Max) for 4,405 animals sourced in 148 contemporary groups.

A total of 5,542 animals (1,128 sires and 4,414 bulls) were genotyped using BovineHD BeadChip (Illumina®, Inc., San Diego, CA, USA) and GeneSeek® Genomic Profiler Indicus HD - GGP75Ki (Neogen Corporation, Lincoln, NE, USA) which contains 777,962 and 74,677 SNP markers distributed across the genome, respectively. Animals genotyped with GGP75Ki were imputed to BovineHD panel using FImpute software (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014), considering pedigree information with an expected accuracy of imputation equal to 0.992 (CARVALHEIRO et al., 2014). After imputation a quality control was performed excluding markers with call rate lower than 0.98, deviations from the Hardy-Weinberg equilibrium (p-value $< 10^{-5}$), with minor allele frequency lower than 0.03 and markers located in non-autosomal chromosomes. Samples with call rate lower than 0.90 were also excluded. After quality control 5,533 animals (4,405 with phenotype and genotype information) and 412,904 SNPs markers were used in the statistical analysis.

2.3.2 Statistical analysis

Searching for causal structures in a mixed model context were performed by following (VALENTE et al., 2010) in three steps: 1) fit a standard MTM to access the posterior residual (co)variance (corrected for cofounder effects); 2) apply the IC algorithm (VERMA; PEARL, 1990; PEARL, 2000) on the residual (co)variance posterior samples to infer causal structures among traits; and 3) fit the SEM from the selected causal structure retrieved by the IC algorithm. Genetic and residual (co)variances were estimated by fitting a standard Bayesian MTM as follow:

$$y = X\beta + Zu + e \tag{2.1}$$

where *y* is a vector of observations, β is a vector of systematic effect of CG and linear terms for slaughter age (linear and quadratic effects), *u* is a vector of random additive effects, *e*is a vector of random residuals and *X* and *Z* are known incidence matrices. Random effects were assumed to be normally distributed $u \sim N(0, G \otimes H)$ and $e \sim N(0, R \otimes I)$, where *G* and *R* are the additive genetic and residual (co)variances matrices, respectively, *H* is the relationship matrix combining pedigree and genomic information (AGUILAR et al., 2010; CHRISTENSEN; LUND, 2010) and *I* is an identity matrix with suitable dimensions. The inverse of the modified relationship matrix *H* (AGUILAR et al., 2010) is defined as:

$$H^{-1} = A^{-1} \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix} \tag{2.2}$$

where A^{-1} is the inverse numerator of the pedigree-based relationship matrix for all animals, G^{-1} represents the inverse of genomic relationship matrix and A_{22}^{-1} is the inverse of the pedigree-based relationship matrix for genotyped animals. The *G* matrix was created as proposed by VanRaden (2008): $G = (M - P)(M - P)' / 2 \sum_{j=1}^{m} p_j(1 - p_j)$, where *M* is is a matrix of genotypes for each animal (coded according to the numbers of copies for the B allele) and *P* is a matrix with the second allele (p_j) frequency, expressed as 2_{pj}. In order to facilitate inversion a weighted *G* was used as proposed by VanRaden (2008): $G = 0.95G_0 + 0.05A_{22}$. In addition, to make *G* proportional to A_{22}, *G* was scaled based on A_{22} considering the diagonal mean of *G* equal to the diagonal mean of A_{22}, and the off-diagonal mean of *G* equal to the off-diagonal mean of A_{22}.

To select the causal structure, the IC algorithm was applied to the residual (co)variances accessed using the MTM. The residual (co)variances draw by the MTM were corrected for the confounding issues caused by additive genetic and fixed effects (i.e. CG and age at slaughter) as described by Valente et al. (2010). In a Bayesian approach, decisions about declaring partial correlation as null or not were made based on HPD intervals (correlation was declared null if the interval contained the value 0). Four different HPD content magnitudes (80, 85, 90 and 95%) were applied to compare the final causal structures, and to observe the structures that were more sensitive to changes in HPD (VALENTE et al., 2010). The IC algorithm was implemented in R program (R Core Team, 2017) by Valente e Rosa (2013).

Finally, using the causal structure inferred by IC algorithms, the structure equation model was fitted as proposed by Gianola e Sorensen (2004):

$$y = (\Lambda \otimes I_n)y + X\beta^* + Zu^* + e \tag{2.3}$$

where *y*, β^*, *u**, *e**, *X* and *Z* have meanings similar to those described for multi-trait model. The vectors β^*, *u** and *e** are effects (systematic and random effects) related to each trait in *y*, however, they are not effects mediated by other traits (GIANOLA; SORENSEN, 2004; ROSA et al., 2011; VALENTE et al., 2013). In addition Λ is a *t x t* matrix with zeroes on the diagonal and structure coefficients or 0 on the off-diagonal, where t is the number of traits used. For structure equation model the join distribution of vectors *u** and *e** were assumed to be normally distributed $u^* \sim N(0, G^* \otimes H)$ and $e^* \sim N(0, R^* \otimes I)$, where G* is the structural equation model additive genetic (co)variances matrix and R* is a diagonal matrix with the structure equation model residual variances (residual covariances assumed to be zero). Such assumption on the residual covariance matrix confers identifiability to the structure coefficients in the likelihood function (INOUE et al., 2016).

For both models (MTM and SEM), marginal posterior distributions of genetic and residual (co)variances were obtained by integrating a multivariate density function in Gibbs2f90 program (MISZTAL et al., 2018). A Gibbs sampling chain with 300,000 samples was generated, with initial 20,000 samples discarded as burn-in and taken each 2 iterates as thinning interval. The Gibbs chain convergence was verified by visual inspection of the sample trace plots and by *coda* package in R (R Core Team, 2017) using Heidelberger and Welch, and Geweke statistics convergence tests (PLUMMER et al., 2006). The remaining 140,000 samples were used as posterior distribution of the variance and (co)variance components in both models (MTM and SEM) and also to characterize the structural equation model.

2.4 RESULTS AND DISCUSSION

Heidelberger and Welch, and Geweke statistical tests and trace plot visual inspection, indicated that Markov chains reached the convergence (results not shown). The posterior means, modes and medians of the heritability estimates were similar for all traits showing symmetrical posterior distributions (Table 2.2). Another indicative of the Markov chains convergence is associated to the Monte Carlo errors (MCE), which were low for all the traits, indicating that chain size was suitable to obtain precise estimates for the posterior parameters. Thus, the mean can satisfactorily represent the properties of the parameters, reflecting the measures of central tendency of the posterior marginal distribution. The posterior means and standard deviations of the heritabilities estimated using MTM for carcass and meat quality traits are shown in Table 2.2. Heritabilities for carcass traits were of low (0.17 for HCW) to moderate magnitude (0.26 and 0.38 for BF and LMA, respectively) and for meat quality traits all the estimates were of low magnitude (0.11 and 0.18 for WBSF and MB, respectively). The estimated heritabilities for LMA, BF and MB traits were higher than those reported by Gordo et al. (2016) and Gordo et al. (2018) using part of the same data. These author reported lower heritabilities for HCW trait than those estimated in our study and similar for WBSF trait. Heritabilities reported by Riley et al. (2002) and Smith et al. (2007) for carcass traits in Brahman cattle were higher than those estimated in our study. For meat quality traits, higher heritabilities than those estimated in this study have also been reported for different beef cattle breeds (RILEY et al., 2002; DIKEMAN et al., 2005; SMITH et al., 2007). The differences between heritabilities observed throughout the studies, may be due to genetic aspect of the breeds, fitted model and also due to environmental conditions that animals were submitted (i.e commercial and research herds).

Table 2.2 – Posterior heritabilities (Mean) and standard deviation (PSD), mode, median, Monte Carlo error (MCE) and high posterior distributions (HPD) interval for multi-trait model.

Traits	Mean (PSD)	Mode	Median	MCE	HPD(95%)
HCW (kg)	0.17 (0.02)	0.16	0.17	0.00010	0.12 to 0.22
LMA (cm^2)	0.38 (0.03)	0.38	0.38	0.00012	0.31 to 0.44
BF (mm)	0.26 (0.03)	0.26	0.26	0.00012	0.20 to 0.32
WBSF (Kg)	0.11 (0.02)	0.11	0.11	0.00008	0.08 to 0.16
MB (score)	0.18 (0.03)	0.17	0.18	0.00011	0.12 to 0.24

Hot carcass weight (HCW), longissimus muscle area (LMA), backfat thickness (BF), Warner-Bratzler shear force (WBSF) and marbling score (MB) traits.

Total genetic and residual correlations estimated using MTM and direct genetic correlations estimated using SEM are depicted in Figure 2.1. The highest total genetic correlations estimated using MTM were between HCW with LMA and WBSF, whereas using SEM the highest direct genetic correlations were between HCW with LMA, BF, and WBSF. These estimates suggest that genes playing important roles to produce heavier animals are partially the same or are linked to the set of genes that play roles to produce larger LMA and more tender meat. The remaining genetic correlations among the traits studied were of low magnitude (-0.21 to 0.19), indicating that selection for one trait will not result in a significant change on the other one. Overall, the genetic correlations (total and direct) are different from the ones previously reported by our group (GORDO et al., 2018) between carcass and meat quality traits, using part of the same data. These differences might be due to the different number of animals used in the studies. Smith et al. (2007) in Brahman cattle and Reverter et al. (2003) in adapted (temperate and tropical) beef breeds reported different genetic correlations compared with the results in this study. The highest residual correlation was observed between HCW and LMA (0.29), whereas low to weak residual correlations were estimated among the remaining traits, ranging from -0.06 to 0.17. Nonetheless, due to the wide amplitude of the highest posterior density region (results not shown), estimated genetic correlations should be treated with caution. The causal structures used as prior information to fit SEM were accessed based on the residual (co)variances estimated in the MTM, using IC algorithm. After applying the described approach to search for causal structures, based on different HPD interval contents, two almost fully directed acyclic graphs were retrieved (Figure 2.2). Using 95, 90 and 85 % of HPD interval the same graph was detected by the algorithm with undirected link of BF with HCW (BF **–** HCW) and BF with MB (BF **–** MB) (Figure 2.2A). Narrower HPD interval (80 %) resulted in an extra link between WBSF and HWC (WBSF → HCW) and directed the arrow between BF with HCW (BF → HCW), whereas the link between BF and MB remained undirected (BF **–** MB) (Figure 2.2B). Using narrower HPD interval one could expect more links being recovered and unshielded colliders should be detected by the IC algorithm in Step 2 (see more details in Valente et al. (2010)). The edges conveyed by the graphs shown in Figure 2.2 (A and B) were stable, since they were present for every HPD interval (except for the extra link between WBSF and HCW). The direction between BF and MB could be determined based on biological prior knowledge as in Gianola e Sorensen (2004). But, regardless of the direction, in this case, there is no more reasonable direction and this decision cannot be made on a statistical basis, once both models are statistically equivalent (VALENTE et al., 2010). Therefore, based on the graph depicted in Figure 2.2B, we fitted two SEM conditioned on the causal structures presented in Figure 2.2C (model A) and Figure 2.2D (model B) with a directed link between BF and MB (BF → MB or BF ← MB). The posterior variance components (genetic and residual) and standard deviations for each trait from MTM and SEM (model A and B), are presented in Table 2.3. Genetic and residual variance posterior means from both SEM were similar. Posterior mean of SEM variances assigned for LMA were smaller than those in MTM, since LMA was conditioned on WBSF and HCW, as shown in Figure 2.2C and D. Smaller SEM variance was inferred for MB,

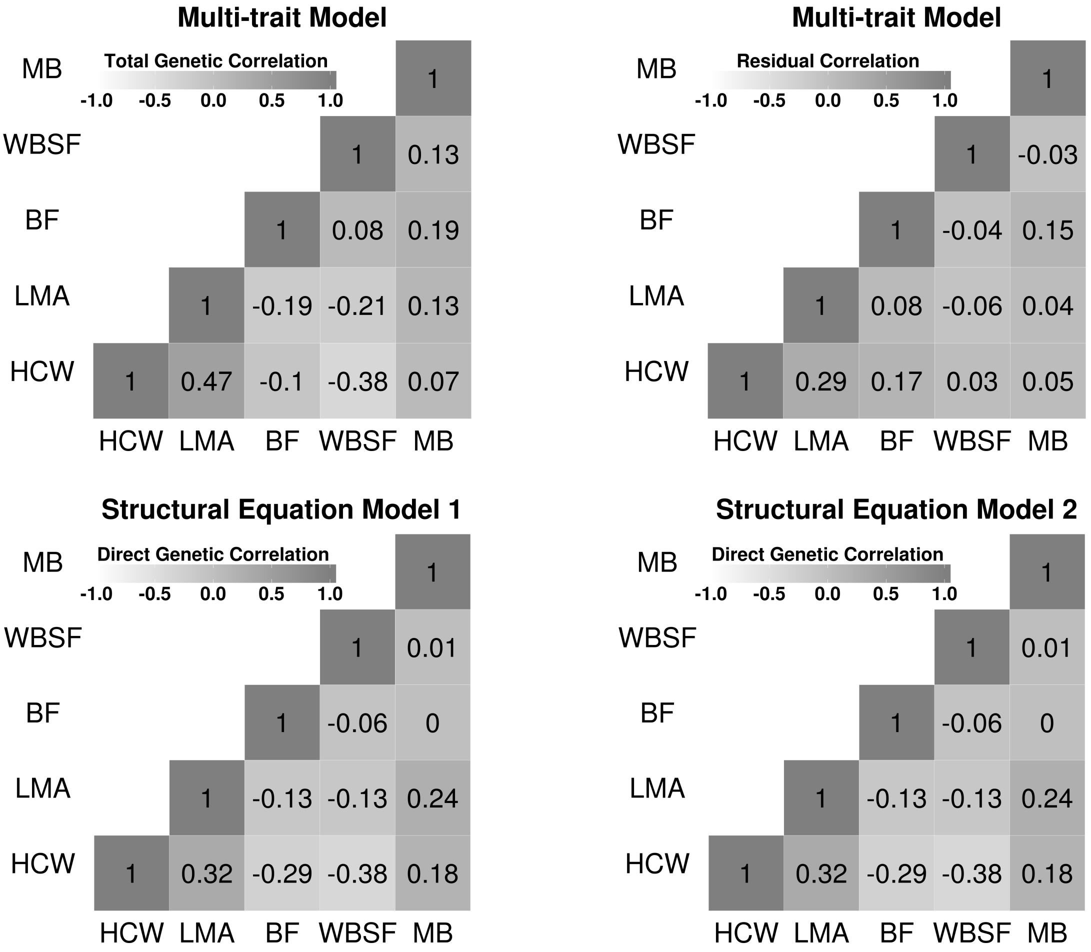

Figure 2.1 – Posterior genetic correlation means for hot carcass weight (HCW), longissimus muscle area (LMA), backfat thickness (BF), Warner-Bratzler shear force (WBSF) and marbling score (MB) traits, obtained using Multi-Trait (top left) and Structure equation models 1 (bottom left) and 2 (bottom right), representing the two models fitted using structure coefficients inferred in Figure 2.2 (C and D). Residual correlation means (top right) are shown only from multi-trait model.

when this trait was conditioned on BF, and for BF when conditioned on MB, according to Figure 2.2C and D, respectively. Higher directed genetic variance was estimated in SEM (model A and B) than in MTM for HCW trait. HCW are conditioned on other two traits (WBSF and BF), which contribute for the genetic variability of HCW. As discussed in Valente et al. (2013), divergence in terms of variance components between MTM and SEM is expected, since MTM estimate overall genetic effects (direct and indirect effects mediated by other phenotypic traits) and the SEM estimate only direct effects (i.e. not mediated by other traits in the causal network). The posterior means of genetic and residual variances estimated for WBSF and BF traits through the different

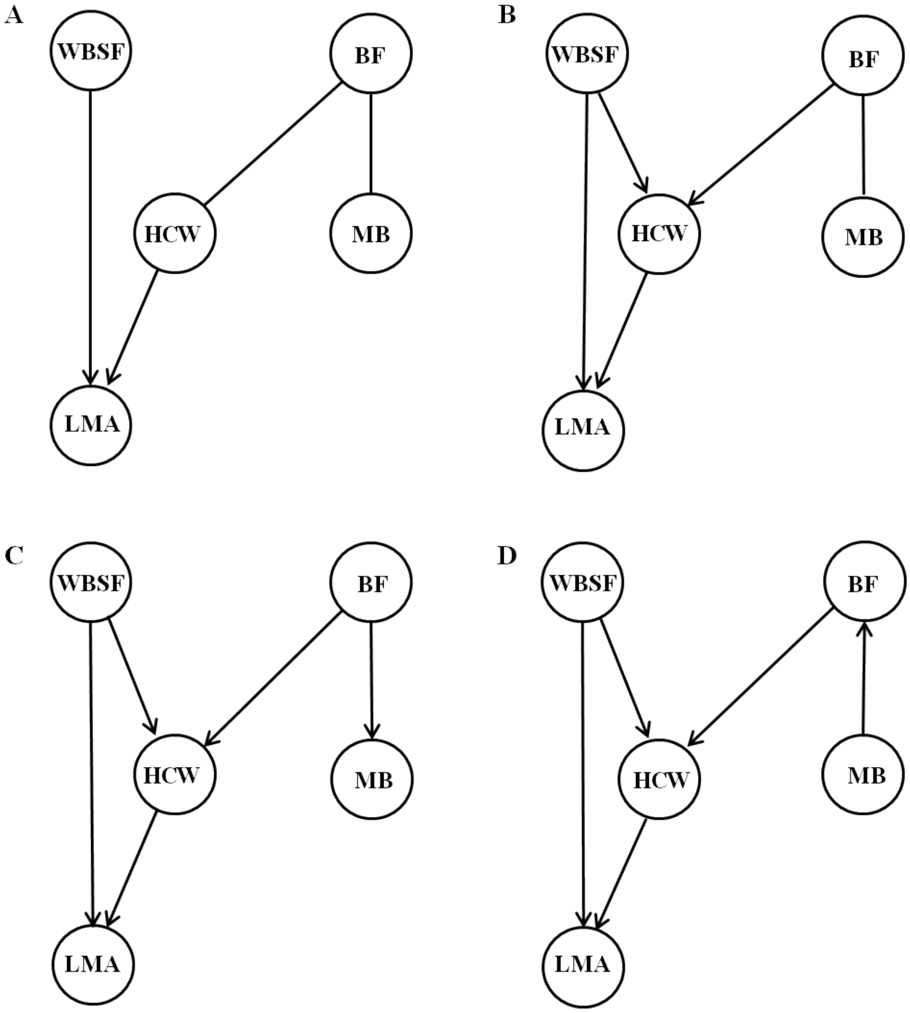

Figure 2.2 – Phenotype network with 95, 90, 85 (A) and 80 % (B) of highest posterior density intervals for hot carcass weight (HCW), longissimus muscle area (LMA), backfat thickness (BF), Warner-Bratzler shear force (WBSF) and marbling score (MB) traits. Network C and D are two structures between BF and MB representing unshielded link observed in B.

models were similar, because these two traits were not conditioned on any other trait, as shown in Figure 2.2C and D. Therefore, for these two traits, the equations in MTM and SEM were similar. Differences in the variance components for the downstream traits (i.e. LMA, HCW and MB or BF) were also observed by Bouwman et al. (2014) and Inoue et al. (2016) for bovine milk fat acid and meat quality traits, respectively, using the same approach proposed by Valente et al. (2010). The posterior means and standard deviations, as well as the 95 % HPD intervals of the structural coefficients inferred using SEM are presented inTable 2.4. Fitting model A and B (based on graphs depicted in Figure 2.2C and Figure 2.2D, respectively) resulted in the same structural coefficients. The deviance information criterion (DIC) proposed by Spiegelhalter et al.

Table 2.3 – Posterior means and standard deviation (PSD) of the variance components for multi-trait model (MTM) and structural equation model (SEM)

Traits	MTM	SEM[1]	SEM[2]
	Mean (PSD)	Mean (PSD)	Mean (PSD)
Genetic			
HCW	57.762 (9.284)	63.122 (9.613)	63.515 (10.074)
LMA	18.203 (1.831)	15.395 (1.568)	15.426 (1.575)
BF	0.791 (0.101)	0.788 (0.103)	0.753 (0.100)
WBSF	0.210 (0.032)	0.203 (0.038)	0.201 (0.039)
MB	0.024 (0.004)	0.023 (0.004)	0.024 (0.004)
Residual			
HCW	283.831 (9.581)	276.131 (9.448)	275.752 (9.626)
LMA	29.938 (1.392)	27.518 (1.216)	27.498 (1.215)
BF	2.264 (0.088)	2.267 (0.089)	2.227 (0.087)
WBSF	1.567 (0.046)	1.569 (0.046)	1.569 (0.046)
MB	0.110 (0.004)	0.107 (0.004)	0.110 (0.004)

Hot carcass weight (HCW), longissimus muscle area (LMA), backfat thickness (BF), Warner-Bratzler shear force (WBSF) and marbling score (MB) traits. SEM1 represent the model considering BF → MB and SEM2 as BF ← MB.

(2002) takes the trade-off between model goodness-of-fit and corresponding complexity of model into account, in which smaller DIC is preferable. Smaller DIC was observed for model A (63,813.80) than for B (63,879.86) and MTM (63,883.89), suggesting the directed acyclic graph with directed link of BF → MB (Figure 2.2C) was more feasible. Therefore, the results from model A will be discussed from now on in terms of structure coefficients $\lambda_{i,j}$ ($\lambda_{i,j}$ denotes a structural coefficient from the jth trait to the ith trait) and biological reasons related to the causal effect among the studied traits. Structural coefficients inferred based on the causal structure selected indicated that WBSF imposes a positive causal effect over HCW ($\lambda_{HCW,WBSF}$). The posterior mean of the magnitude of change in HCW due to a 1 kg increase in WBSF was inferred as 0.43 kg (Table 2.4). In turn, HCW imposes a positive effect on LMA ($\lambda_{LMA,HCW}$), with a posterior mean of 0.10 cm^2. This structure implies that WBSF presents also an indirect positive causal effect on LMA. The causal structure indicates also that WBSF has a negative causal effect on LMA ($\lambda_{LMA,WBSF}$), with posterior mean of -0.29 cm^2. The sign of the three structural coefficients was the same as the sign of residual covariance among these traits estimated in the MTM.

One possible biological explanation for the causal effect between WBSF and HCW might be related to *longissimus thoracis* muscle fiber type. In Nelore cattle the *longissimus thoracis* muscle have high proportion of MyHC-IIx and MyHC-IIa fiber types, which is related to the increase in muscle mass during postnatal growth as well as tough meat (MARTYN; BASS; OLDHAM, 2004; OLIVEIRA et al., 2011; CHRIKI

Table 2.4 – Posterior means, standard deviation (SD) and 95% highest posterior density (HPD) intervals of the structural coefficients

Structural coefficients	Structural Equation Model 1			Structural Equation Model 2		
	Mean	SD	HPD(95%)	Mean	SD	HPD(95%)
$\lambda_{LMA,HCW}$	0.10	0.006	0.08 to 0.11	0.10	0.008	0.08 to 0.11
$\lambda_{LMA,WBSF}$	-0.29	0.09	-0.48 to -0.10	-0.29	0.09	-0.49 to -0.10
$\lambda_{HCW,WBSF}$	0.43	0.29	-0.12 to 1.00	0.43	0.27	-0.10 to 0.96
$\lambda_{HCW,BF}$	1.92	0.28	1.39 to 2.46	1.92	0.28	1.37 to 2.49
$\lambda_{MB,BF}$	0.03	0.006	0.02 to 0.04	-	-	-
$\lambda_{BF,MB}$	-	-	-	0.66	0.11	0.45 to 0.88

Hot carcass weight (HCW), longissimus muscle area (LMA), backfat thickness (BF), Warner-Bratzler shear force (WBSF) and marbling score (MB) traits. Here, $\lambda_{i,j}$ denotes a structural coefficient from the j^{th} trait to the i^{th} trait.

et al., 2012; PICARD et al., 2014). In addition, Guillemin et al. (2011) and Picard et al. (2014) have reported association between muscle proteins and meat tenderness in *Bos Taurus*. Another hypothesis may be due to the mechanism involving suppression of protein degradation which, according to Koohmaraie et al. (2002), increases muscle deposition and decreases meat tenderness.

The high proportion of MyHC-IIx and MyHC-IIa fiber types present in Nelore cattle *longissimus thoracis* muscle, which increase muscle mass during postnatal growth, may be a biological explanation also to the causal relationship between HCW and LMA (MARTYN; BASS; OLDHAM, 2004; OLIVEIRA et al., 2011; CHRIKI et al., 2012; PICARD et al., 2014). In addition, animals with high HWC have presented larger LMA in different breeds (BRONDANI et al., 2004; REZENDE et al., 2012). For the relationship between WBSF and LMA, no biological explanation was found in the literature. But the edge between WBSF and LMA was stable, as it was present regardless the HPD interval contents (Figure 2.2), and, therefore, supported by the data, i.e. the statistical consequences of their association were found by the IC algorithm applied on the posterior distribution of residual (co)variances (VALENTE et al., 2011).

Inferences for the remaining edges indicate that BF has a positive causal effect over HCW ($\lambda_{HCW,BF}$) and MB ($\lambda_{MB,BF}$), with posterior mean of 1.92 kg and 0.03 score grade, respectively. The sign of the coefficients $\lambda_{HCW,BF}$ and $\lambda_{MB,BF}$ were the same as the sign of residual covariances inferred in MTM. The causal effect of BF on HWC was expected once animals were placed in a feedlot system in which they are, usually, fed with a high energetic diet (BOITO et al., 2018). During this period animals have enough energy to convert into muscle mass and depot as BF (BRONDANI et al., 2004), such that increasing the BF will also increase HCW (BOITO et al., 2018). In addition, specific tissues changes with animal age by reducing muscle and bone growth rates and increasing fat deposition (BERG; BUTTERFIELD, 1976; NURNBERG; WEGNER; ENDER, 1998). Although weak, the causal effect of BF on MB can be explained based on the process of muscle tissue development in cattle. Fat deposition in cattle follows a chronological order during animal's life, where intermuscular fat is the first fraction of adipose tissue that is accumulated in the carcass followed by subcutaneous and intramuscular fat or marbling (REZENDE et al., 2012).

Causal relationships change the focus from marginal associations among traits and allow predicting changes when external interventions are applied (ROSA et al., 2011). For the set of trait studied the most important interventions are that affecting LMA, HCW, and MB, because of their economic importance. For example, if an intervention is made on BF, such as controlling the diet energetic level, only the direct genetic effects would influence MB and HCW, as the intervention would block the indirect genetic effect through BF (VALENTE et al., 2013). This intervention also would indirectly influence LMA through the effect of HCW on LMA. External intervention on WBSF may also influence selection for LMA and HCW. Animals slaughter age or feeding management, for example, has been stressed out as important factors influencing meat tenderness in cattle (CHRIKI et al., 2013). By controlling these factors, indirect effects through WBSF could be blocked, and LMA and HCW would only be affected by direct genetic effects (VALENTE et al., 2013). Thus, if an external intervention exists on traits presenting causal effects among them, a breeding strategy based only on MTM analysis could lead to wrong selection decisions (INOUE et al., 2016). Regardless of the structure coefficients magnitude inferred here, SEM produced interesting and useful results, generating causality hypotheses for further research and investigation for carcass and meat quality traits (ROSA et al., 2011). But, the causal relationships among carcass and meat quality traits identified in this study might be considered with caution and confirmed using a larger number of records.

2.5 CONCLUSIONS

Potential causal relationships were detected among carcass and meat quality traits in Nelore cattle. Using structure equation model, Warner-Bratzler shear force had negative and positive causal effects on longissimus muscle area and hot carcass weight, respectively; hot carcass weight had positive causal effect on longissimus muscle area; and backfat thickness had positive effect on hot carcass weight and marbling score. These findings suggest that interventions on Warner-Bratzler shear force and backfat thickness would direct affect hot carcass weight, longissimus muscle area, and marbling score.

2.6 ACKNOWLEDGMENTS

The authors would like to acknowledge DeltaGen, PAINT and Cia de Melhoramento breeding programs for providing the data. We also would like to acknowledge the São Paulo Research Foundation (FAPESP) for supporting this work (grants: #2009/16118-5, #2016/02366-0 and #2017/18779-5). The present study was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Resources supplied by the Center for Scientific Computing (NCC/GridUNESP) of the São Paulo State University (UNESP) also supported this research.

CHAPTER 3 - INTEGRATION OF MULTI-OMICS DATA TO INVESTIGATE CAUSAL NETWORK FOR CARCASS AND MEAT QUALITY TRAITS IN NELORE CATTLE

3.1 ABSTRACT

Information regarding molecular networks can be used to better understand phenotype expression. In this context, the integration of heterogeneous omic data has the potential to uncover gene networks and the causal relationships among variables under study. The objective of this study was to reconstruct gene-phenotype networks and to perform a causal network analysis by integrating phenotypic, genotypic, and transcriptomic data in Nelore cattle. Longissimus muscle area (LMA, cm^2), backfat thickness (BF, mm) and Warner-Bratzler shear force (WBSF, kg) traits were used. Phenotypes and genotypes information from 4,599 bulls were used and gene expressions were accessed for 80 animals. In order to identify genomic regions associated with phenotypes, two genome scan analyses were performed: exploring association between genotypic and phenotypic data (pQTL – phenotype quantitative trait loci mapping), and between genotypic and gene expression data (eQTL – expression quantitative trait loci mapping). For both genome scan analysis, a mixed linear model was used applying the framework living-one-chromosome-out. A multi-trait analysis was carried out using markers effects from each single genome scan analysis for the phenotypes studied (LMA, BF, and WBSF). Co-localized genome regions identified by integrating multi-omic data were used to reconstruct gene network and causal inference through structural learning algorithm. Fourteen genome regions showed significant associations with LMA, BF, and WBSF in the multi-trait analysis and 19 *cis*-eQTL were overlapping five of the genome regions. Based on the five *cis*-eQTL position (the most significant in each genome region), thirty-two nearby genes were identified. Integrating phenotypes, genotypes and gene expression data the inferred network indicated that the *rs137704711*, located in chromosome 20, affected the three phenotypes (LMA, BF, and WBSF), and the *rs133894950*, located in chromosome 16, affected BF through the expression of several genes located in different chromosomes.

Key-words: beef cattle, causal inference, gene expression, graphical models, quantitative trait loci

3.2 INTRODUCTION

Carcass and meat quality traits such as LMA, BF and WBSF have an important impact on consumer satisfaction and meat product pricing. Despite their importance, selection programs have not fully explored these traits due to costly and difficult to measure, they are observed later in the animal's life and are mediated by many genes and environmental factors (HOCQUETTE et al., 2012; FONSECA et al., 2017). Genome-wide association studies (GWAS) have identified several genomic regions, also termed as pQTL, for carcass and meat quality traits (KIM et al., 2011; LU et al., 2013; MAGALHÃES et al., 2016; FERNANDES JÚNIOR et al., 2016). However, pQTL identified in a typical GWAS explaining only a small fraction of the genetic variability, they are not necessarily true causal variants and the majority pQTL fall in non-coding genomic regions (MACKAY; STONE; AYROLES, 2009; MONTGOMERY; DERMITZAKIS, 2011; AINSWORTH; SHIN; CORDELL, 2017).

Advances in sequencing technologies have enabled high-throughput measurement of transcriptome at the population level, which combined with genotype markers have made possible to map thousands eQTL (BOUWMAN et al., 2018; CESAR et al., 2018; HIGGINS et al., 2018). Indeed, the eQTL mapping is a widely and powerful tool to identify regulatory mechanisms involved in the phenotypic expression (MONTGOMERY; DERMITZAKIS, 2011). Combining eQTL and pQTL information may has the potential to uncover gene networks, the genetic control of gene activity and unravel the genetic architecture of complex traits, as well as may help to shed light on the non-coding variants that might play important role in phenotype expressions (KADARMIDEEN; VON ROHR; JANSS, 2006; HUANG; ZHENG; PRZYTYCKA, 2010; STEIBEL et al., 2011; NICA; DERMITZAKIS, 2013; YANG; RONG; KUI, 2017).

The integration of different layers of information might be used to elucidate causative changes that lead to phenotypes variation (HASIN; SELDIN, 2017). However, genetical genomic studies have not focused on the causal relationship between the variables under study (CHAIBUB NETO et al., 2010; PEÑAGARICANO et al., 2015). Investigating causal relationships among phenotype, genotype, and gene expression are justified by the Mendelian randomization of alleles and the unidirectional effect of genotype on gene expression and phenotype (ROSA et al., 2011; CHEN, 2012; PEÑAGARICANO et al., 2015). Even with the availability of a large number of variables and omic layers, learn causality remains a challenge and few studies have been conducted in animal livestock. The objective of this study was to reconstruct gene-phenotype networks and perform a causal network analysis by integrating phenotypic, genotypic, and transcriptomic data in Nelore cattle.

3.3 MATERIAL AND METHODS

To reconstruct gene-phenotype networks and perform causal inference the multistep procedure proposed by Peñagaricano et al. (2015) was used. Briefly, genomic regions associated with LMA, BF, WBSF and expression traits were identified through genome scan analyses (pQTL and eQTL mapping), and then, significant regions from both QTL mapping were co-localized to perform network reconstruction using causal structural learning algorithms. For the data set used here, all animal procedures were approved by the São Paulo State University (Unesp), School of Agricultural and Veterinary Science Ethical Committee (Approval No. 18.340/16).

3.3.1 Phenotypic data collation

Data were collected from 4,599 Nelore bulls born between 2008 and 2014. The animals were raised under pasture conditions and finished in feedlot system (for around three months), at different farmers located in the southeast, mid-west, and northeast of Brazil. These animals were slaughtered at an average age of 24 months in commercial slaughterhouse. After 24 to 48 hours chilling, from the slaughter, a sample from *longissimus thoracis* muscle (between 12 and 13th ribs) were collected for each animal. Using a plastic grid (squared with 1 cm^2) placed on a steak of 2.54 cm thickness, LMA trait was measured. BF trait, defined as the layer of subcutaneous fat on the steak, was measured using a caliper. Steaks were cooked to an internal temperature of 71 °C and cooled at 2 °C for 24 hours as proposed by Wheeler, Koohmaraie e Shackelford (1995). The mean of eight cooked meat cylinders (1.27 mm), sheared with a V blade attached to a Warner-Bratzler shear force machine (G-R Electric, Manhattan, KS), was used as WBSF trait. For the analyses, contemporary groups (CG) were defined as year and farm of birth, farm and management group at yearling and slaughter date. Phenotypes with three standard deviations above or below to the mean of their CG and CG with less than three animals were removed from the data set. Further details for the data set used in the analyses are shown in Table 3.1.

Table 3.1 – Descriptive statistical for longissimus muscle area (LMA), backfat thickness (BF) and Warner-Bratzler shear force (WBSF) traits in Nelore cattle.

Traits	Mean	SD	Min	Max
LMA (cm^2)	68.01	7.87	40.0	96.0
BF (mm)	4.75	2.17	1.0	14.0
WBSF (Kg)	6.44	1.90	1.8	11.9

Standard deviation (SD), minimum (Min) and maximum (Max) for 4,599 animals sourced in 156 contemporary groups.

3.3.2 Genotypic data

DNA was isolated for genotyping using muscle tissues *longissimus thoracis* from 4,599 animals described previously. The DNeasy Blood & Tissue Kit (Qiagen GmbH, Hilden, Germany) was used to extract DNA as manufacturer's instructions. Once DNA was isolated, samples were analyzed for quality and quantity using a Nanodrop spectrophotometer. Genotyping was performed using BovineHD BeadChip (Illumina®, Inc., San Diego, CA, USA) and GeneSeek® Genomic Profiler Indicus HD - GGP75Ki (Neogen Corporation, Lincoln, NE, USA) which contains 777,962 and 74,677 SNP markers distributed across the genome, respectively. The FImpute software (SARGOLZAEI; CHESNAIS; SCHENKEL, 2014) was used to carry out genotype imputation from GGP-75Ki to BovineHD, including pedigree information, with an accuracy of imputation equal to 0.992 (CARVALHEIRO et al., 2014).

3.3.3 Gene expression data

Longissimus thoracis muscle tissue samples were collected from 80 bulls (previously described) during slaughter. RNA sequencing (RNA-seq) was carried out by Illumina HiSeq 2500 System to produce 2x100 base pairs paired-end reads. Details regarding tissue sample collection, and RNA extraction and sequencing were reported by Fonseca et al. (2017). In order to improve mapping specificity, reads were trimmed to remove contaminated adaptor sequences using Trimmomatic 0.36 (Bolger et al., 2014). Based on the Ensembl Bos_taurus UMD3.1 (version 92), reads were mapped and counted using STAR software (DOBIN et al., 2013). Gene count data were normalized using Trimmed Mean of M-values (TMM) method and log_2 transformed using *edgeR* (ROBINSON; MCCARTHY; SMYTH, 2010) R package (R Core Team, 2017). Principal component analysis (PCA) was performed on genes expression data to identify confounding factors within gene expression data (ELLIS et al., 2013) using *prcomp* function in R software (R Core Team, 2017).

3.3.4 Quality control

Genotype quality control for pQTL mapping was carried out by removing genotypes unmapped to autosomes SNP markers or sex-linked, with call rate lower than 0.98, minor allele frequency lower than 0.05 and those that deviated from the Hardy-Weinberg equilibrium (p-value < 10^{-5}). Samples were removed from analysis if they had call rate lower than 0.90. After quality control, a total of 4,599 samples and 410,019 SNP markers remained for further analyses. Similar quality control previously described was applied for genotypes and samples used in the eQTL mapping, except for minor allele frequency, in which SNP markers with minor allele frequency lower than 0.04 were removed. Additionally, genotypes with less than 2 animals in all genotypes were also removed. For transcriptomic data, expressed genes with non-zero counts in more than 20% of all animals were retained in the data set. Seventy-eight samples, 302,829 SNP markers, and 12,863 genes expression were suitable for further analysis. All quality control were performed using *snpStats* (CLAYTON, 2015) R package (R Core Team, 2017).

3.3.5 Phenotype and expression QTL mapping

Genome scan analyses between each phenotype (LMA, BF, and WBSF) and SNP makers (pQTL mapping) were performed by using a mixed linear model implemented in GCTA software (YANG et al., 2011). The model fitted was:

$$y = X\beta + Wu + g + e \tag{3.1}$$

where *y* is an *n x 1* vector of phenotypes with *n* being the number of animals, β is a vector of fixed effects (CG and slaughter age), *u* is a vector of SNP marker effects, *g* is an *n x 1* vector of the polygenic effect (captured by the genomic relationship matrix calculated using all SNP), *e* is a vector of residual effects, *X* is an incidence matrix relating β in *y* and *W* is a standardized genotype matrix with the ij^{th} element w_{ij} = $(x_{ij}$ - $2p_i)$ / $\sqrt{2p_i(1-p_i)}$ where x_{ij} is the number of copies of the reference allele for the i^{th} SNP of the j^{th} individual and p_i is the frequency of the reference allele. The assumptions for markers (*u*), genetic (*g*) and residual (*e*) effects were: $u \sim N(0, I\,\sigma^2_u)$, $g \sim N(0, G\,\sigma^2_g)$, and $e \sim N(0, I\,\sigma^2_e)$, where *I* is an *n x n* identity matrix, σ^2_u is the SNP variance, σ^2_g is the variance explained by all the SNPs, σ^2_e is the residual variance, and *G* is the genomic relationship matrix. For the ease of computation, σ^2_g is estimated based on the null model (i.e. $y = X\beta + g + e$), and then fixed while testing for the association between each SNP marker and the trait. Genetic relationship between individual *j* and *k* was estimated by the following equation:

$$G_{ik} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij}-2p_i)(x_{ik}-2p_i)}{2p_i(1-p_i)}, \qquad \textit{i = 1, 2, ..., N markers.} \tag{3.2}$$

The significance of the additive marker effect on each trait was tested using the probability value (p-value) test by comparing the full model to the null model without marker effects.

From the SNP markers effect estimated by each single-trait GWAS, we applied a multi-trait statistic test to determine the effect of i^{th} SNP (i = 1, 2, . . . , 410,019) across LMA, BF and WBSF traits as proposed by Bolormaa et al. (2014). Following a *chi-square* distribution (χ^2) with *t* degrees of freedom (*t* is the number of traits) this approach test for each SNP marker, based on a null hypothesis that a SNP maker has no effect on any trait. Multi-trait significant level for each SNP marker was calculated as follow:

$$\chi^2_{Multi-trait} = t_i^{'} V^{-1} t_i, \tag{3.3}$$

where t_i is a *3 x 1* vector of signed *t-values* of i^{th} SNP for all the traits, $t_i^{'}$ is a transposed of vector t_i and V^{-1} is an inverse of *3 x 3* correlation matrix where the correlation between two traits is the correlation over the 410,019 *t-values* of the two traits. The *t-values* were calculated as $t_i = u_i / SE(u_i)$, where *SE* is the standard error for the i^{th} SNP markers. False discovery rate (FDR) of 1 % was used to control for multiple testing (BENJAMINI; HOCHBERG, 1995).

For eQTL mapping, the same mixed linear model described for pQTL mapping were applied. However, gene expression data expressed in log_2 scale were used as target variable (*y*), the fixed effect of sequencing date, and the first six principal components (to account for confounding factors in the gene expression data) and age at slaughter (linear effect) as covariates in the model ($X\beta$). The CG was not included as fixed effect because the animals were raised under the same environmental condition (i. e. sex, management, farm, and year). Local eQTL (*cis*-eQTL) were identified when the significant SNPs was located within 1 Mb of the associated gene. The *cis*-eQTL p-values were corrected for multiple testing across all expression traits using FDR of 5% (Benjamini and Hochberg, 1995). The eQTL mapping was carried out by our group in a study in preparation (BRAZ et al.). The Gene Ontology (GO) and biological pathways annotations of the genes were retrieved using the *biomaRt* package (DURINCK et al., 2009) in R (R Core Team, 2017). The positions of the significant genome regions were compared with positions of know QTL on the *Bos taurus* UMD3.1 reference genome according to the Animal QTL database (HU et al., 2013).

3.3.6 Co-localized genome regions

Given a particular significant pQTL identified in the multi-trait analysis, delimited by a 250 kb interval to each side of the peak, all significant *cis*-eQTL overlapping this region were picked. However, for the causal inference we selected only the most significant *cis*-QTL as well as the gene expressions where these *cis*-eQTL are located. Based on the *cis*-eQTL position, a window (500 kb downstream and upstream) was opened and gene expressions for all nearby genes were used. The co-localized genome regions were carried out in R (R Core Team, 2017).

3.3.7 Causal inference

Co-localized genome regions identified by integrating multi-omic data were used to perform causal inference through Bayesian network (BN). A BN is a special case of graphical model, in which all the edges are directed (directed acyclic graph – DAG), that encodes a joint probability distribution over a set of random variables (PEARL, 1988). BN are composed of two parts: a set *X* = (x$_1$, x$_2$, ..., x$_p$) of random variables describing the quantities of interest, and a graph DAG = (*V*, *E*), in which each vertex *v* $\in$ *V*, also called node, is associated with one of the random variables in *X*. Each edges *e* $\in$ *E*, also called arcs or links, are used to express the dependence structure of the data (SCUTARI; STRIMMER, 2011). The *X* variables under investigation here include *t* traits (x$_{t1}$, x$_{t2}$, . . . , x$_{tn}$), *q cis*-eQTLs (x$_{q1}$, x$_{q2}$, . . . , x$_{qn}$), and *g* genes expressions traits (x$_{g1}$, x$_{g2}$, . . . , x$_{gn}$). Phenotypes (LMA, BF, and WBSF) and gene expressions used in the causal learning structure were pre-adjusted for the systematic effects described for pQTL and eQTL mapping (see *Phenotype and expression QTL mapping* section). BN was performed in two steps: first, the DAG structure was identified (structure learning) and second, the parameters were estimated (parameter estimation) based on the structure obtained in the first step. For structure learning we used the constraint-based structure learning algorithm (TSAMARDINOS; ALIFERIS; STATNIKOV, 2003). The conditional independence tests were based on *Fisher's Z* test, which is a transformation of the linear correlation coefficients between *X* and *Y* given *Z* (ρ*XY* | *Z*) and defined as:

$$Z(X, Y \mid Z) = log\left(\frac{1+\rho_{XY|Z}}{1-\rho_{XY|Z}}\right)\frac{\sqrt{N- \mid Z \mid -3}}{2} \tag{3.4}$$

where *n* is the number of observations and | *Z* | is the number of nodes belonging to *Z*. This test has an approximate normal distribution *Z(X, Y*| *Z)* $\sim$ *N (0, 1)*. In the structure learning *a priori* biological knowledge was used, excluding the possibilities that *t* $\rightarrow$ *q*, *t* $\rightarrow$ *g*, and *g* $\rightarrow$ *q*; that is, phenotypes cannot affect *cis*-eQTLs and gene expression as genomic variables are measured at baseline and the phenotypes is measured at follow-up times, and gene expression cannot affect *cis*-eQTL since according to the central dogma of molecular biology messenger RNA is produced by transcription from segments of DNA (NI; STINGO; BALADANDAYUTHAPANI, 2014). After structure learning the causal parameters estimation, under the context of normal distributed variables, were performed using maximum likelihood. The structure stability of the causal networks was evaluated using Jackknife resampling by leaving out one observation per time from the dataset. For each sampling we evaluate the presence (presence or absence within samples) and direction (same direction as the original arrow, opposite direction, or undirected arc within samples) of the arc based on original graph. All the analyses were performed using *bnlearn* package (SCUTARI, 2010) implemented in R (R Core Team, 2017).

3.4 RESULTS AND DISCUSSION

Following the procedure proposed by Peñagaricano et al. (2015), the first step was to identify genome regions (pQTL) associated with phenotypes (LMA, BF, and WBSF). From classical genome scan analysis we performed a multi-trait analyses, in which 96 SNP markers were found significant (FDR < 1 %) across the three traits (Figure 3.1). The multi-trait analysis is a powerful procedure to identify pQTL positions that affect all the associated traits (BOLORMAA et al., 2014). Significant SNP markers are located in fourteen genome regions (pQTL) distributed over 10 chromosomes: 1, 3, 5, 8, 12, 13, 14, 16, 20 and 22. A total of 39 genes were identified within regions tagged by the pQTL including *RAP2B*, *NDUFA10*, *PCED1B*, *AMIGO2*, *bta-mir-1251*, *MYBPC1*, *CHPT1*, *SYCP3*, *GNPTAB*, *DRAM1*, *WASHC3*, *TMC1*, *ALDH1A1*, *ANXA1*, *MAFB*, *RF00026*, *LYN*, *RPS20*, *PLAG1*, *CHCHD7*, *SDR16C5*, *SDR16C6*, *PENK*, *MOS*, *RF01277*, *RF00003*, *TGS1*, *TMEM74*, *PARK7*, *TNFRSF9*, *VAMP3*, *CAMTA1*, *ERRFI1*, *CCNB1*, *SLC30A5*, *CENPH*, *TGFBR2*, *GADL1*, and RF00026. Overall, these candidate genes are related to zinc ion transport (GO:0006829), lipid metabolic process (GO:0006629), negative regulation of cell population proliferation (GO:0008285), regulation of gene expression (GO:0010628), regulation of keratinocyte differentiation (GO:0045616), muscle contraction (GO:0006936), insulin secretion (GO:0030073), myoblast migration involved in skeletal muscle regeneration (GO:0014839), positive regulation of MAPK cascade (GO:0043410), regulation of growth (GO:0040008), and cell differentiation (GO:0030154). Some of these identified genome regions have already been associated with LMA, BF and WBSF in Nelore cattle (FERNANDES JÚNIOR et al., 2016; MAGALHÃES et al., 2016; SILVA et al., 2017). These findings provide evidence about the existence of genome regions with additive pleiotropic effects on LMA, BF and WBSF traits.

The second step of the procedure proposed by Peñagaricano et al. (2015) was to identify eQTL through the integration of genotype and gene expression data, and was performed by our group in a study in preparation (BRAZ et al.). Overall, the mapping of SNP markers associated with variation in Nelore muscle tissue RNAs identified 10,026 significant (FDR < 5 %) putative *cis*-eQTL, affecting the expression of 1,343 genes on all autosomal chromosomes. Many SNPs in non-coding regions were identified and their identification might be crucial to the understanding of the molecular mechanisms underlying economically important traits in Nelore cattle.

Nineteen putative *cis*-eQTL (FRD < 5%) were detected overlapping five particular pQTL (Table 3.2), and they were associated with the level of expression of the following genes: *NDUFA10*, *WASHC3*, *VAMP3*, *TAF9*, and *GADL1*, located in chromo-

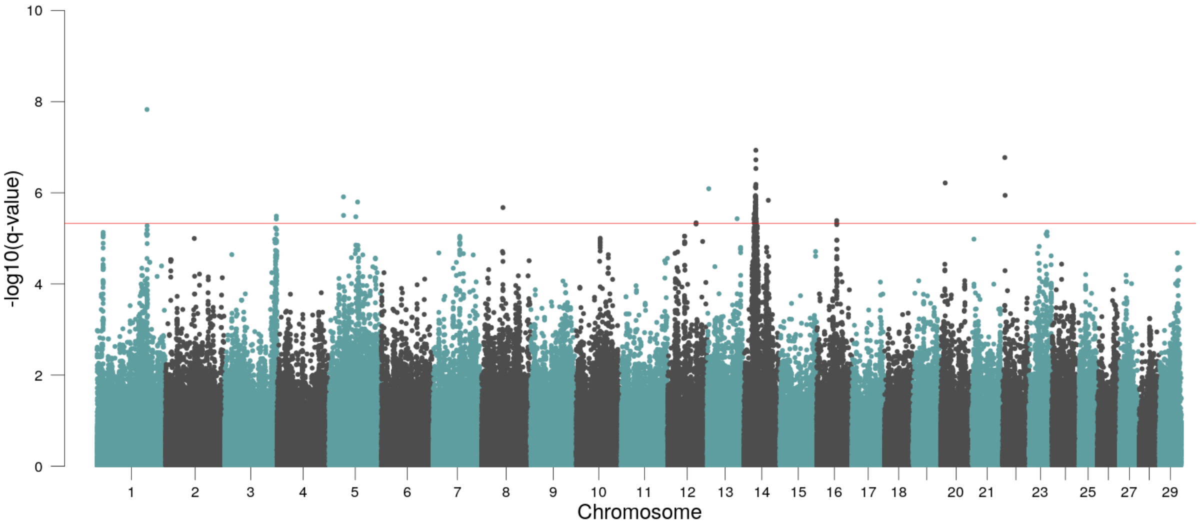

Figure 3.1 – Manhattan plot based on the *–log*$_{10}$ *p-values* of marker throughout the whole genome using multi-trait analysis. Red line indicates significance at FDR 1%

somes 3, 5, 16, 20, and 22, respectively. Genome scan analysis showed that there are at least two phenotypes and five different gene expression traits significantly associated with the same genome regions on chromosomes 3, 5, 16, 20, and 22 for the populations studied. In addition, twenty-seven nearby genes (within a window of 500 kb downstream and upstream) were considered co-localized with the *cis*-eQTL including *CSF2RA*, *COPS9*, *OTOS*, *MYBPC1*, *CHPT1*, *SYCP3*, *GNPTAB*, *DRAM1*, *NUP37*, *PARPBP*, *IGF1*, *ERRFI1*, *PARK7*, *CAMTA1*, *NAIP*, *GTF2H2*, *OCLN*, *MARVELD2*, *RAD17*, *AK6*, *CCDC125*, *CDK7*, *MRPS36*, *CENPH*, *CCNB1*, *SLC30A5*, and *TGFBR2*.

Table 3.2 – Co-localized significant genome regions by chromosome (BTA).

BTA	pQTL (start-end position)	*cis*-eQTL
3	119591217 - 120091217	***rs136982136***
5	65733035 - 66233035	***rs135207526***; *rs133451333; rs136853608*
16	46203881 - 46703881	***rs133894950***; *rs133860779; rs133046724; rs132642057; rs137177225*
20	10422865 - 10922865	***rs137704711***
22	4924587 - 5424587	***rs41992695***; *rs109064338; rs134054023; rs136516831; rs41991944; rs41991942; rs41991225; rs41991219; rs110378126*

In bold is the most significant *cis*-eQTL.

For causal structure learning three pre-adjusted phenotypes, five *cis*-eQTL (the most significant *cis*-eQTL) and 32 genes expression traits (pre-adjusted for the systematic effects) were included in the analyses. The causal network inferred by the IAMB algorithm is depicted in Figure 3.2. The algorithm allowed to reconstructing a partially directed acyclic graph, without using any prior information, with only one undirected edge (Figure 3.2a). The link between *TAF9* and *NAIP* genes remained unresolved (i.e. undirected). *TAF9* is a transcription factor and may have causal effect on *NAIP*. From this information (prior knowledge) is possible to set direct link between them (*TAF9* → *NAIP*), which allow the algorithm to reconstruct a fully directed acyclic graph (Figure 3.2b). Based on the causal graphical model, genetic markers are marginally associated with the remaining variables under study (i.e., phenotypic and expression traits). The causal network (Figure 3.2b) indicated that the effects of the genotypes on the phenotypes are mediated by the expression of several genes located in different genome regions and chromosomes. In fact, these results were expected once the traits studied here are considered complex been modulated by many genes.

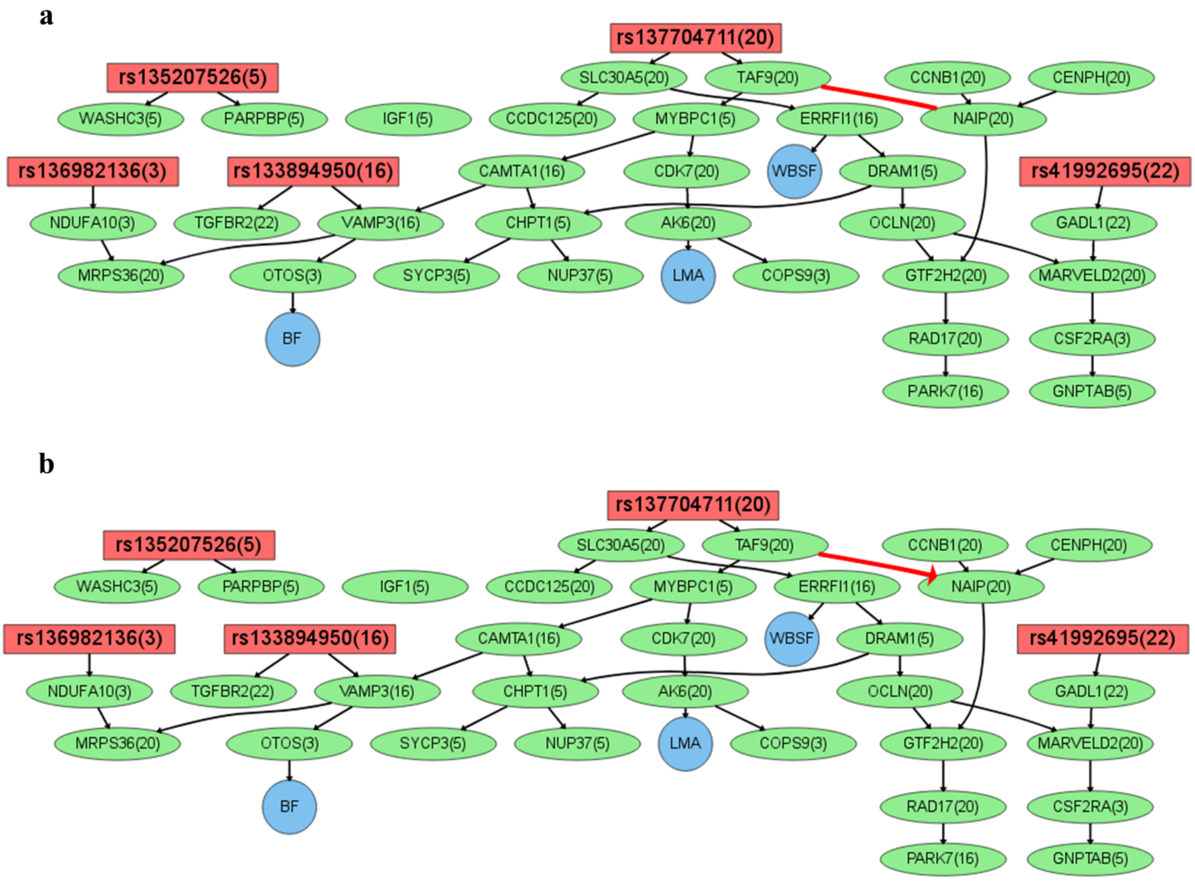

Figure 3.2 – Causal networks integrating phenotypic (blue), *cis*-eQTL (red) and transcriptomic (green) data. Causal network inferred without using any prior information (a) and causal network inferred with the prior knowledge TAF9(20) → NAIP(20) (b). Between parentheses are the chromosomes for each *cis*-QTL and expression level.

The SNP *rs137704711*, located on chromosome 20 at 10,627,752 bp, stood out as influencing LMA, BF and WBSF. Many QTL for growth and meat quality traits including body weight (QTLs 11100, 11101, and 11102), hump length and width (QTLs 3423, and 3427, respectively), and fatty acid content (QTL 12250), were identified in beef cattle in the same region. In addition, Fernandes Júnior et al. (2016) reported a QTL for LMA in this region using animals from the same population as used in this study. The *rs137704711* influence WBSF through the expressions of the genes *SLC30A5* and *ERRFI1* (Figure 3.2b). *SLC30A5* encodes a protein involved in zinc transport into beta cells in order to produce insulin (KARISA et al., 2013), which contributes to ontogenesis of skeletal muscles (GOTOH et al., 2014). Karisa et al. (2013) reported that *SLC30A5* was associated with carcass traits in beef cattle. According to Liu et al. (2015), *ERRFI1* plays role in the development of muscular system. *ERRFI1* is involved in keratinocyte proliferation and differentiation, which may play an important role in myoblast differentiation (LEAL-GUTIÉRREZ et al., 2018). Genes related to keratin filament were reported to be associated with WBSF in Nelore steers (TIZIOTO et al., 2013).

The genes in the pathway between the *rs137704711* and LMA (*TAF9*, *MYBPC1*, *CDK7*, and *AK6*) are involved in muscle development processes. *TAF9* encodes the TATA-box binding protein (TBP) associated factor 9 which is a transcription factor that may regulate the expression of several genes influencing LMA. Malecova et al. (2016) showed that TBP is required for skeletal muscle differentiation since they regulate the transcription of the MyoD gene family during this process in mice. MyoD gene family regulates the number of muscle fibers at birth and plays key roles in growth and muscle development (te PAS et al., 1999; HANDEL; STICKLAND, 1988). Du et al. (2013) reported that a member of the MyoD gene family was associated with LMA in beef cattle. *MYBPC1*, a myosin binding protein C, interacts with muscle-type creatine kinase, allowing it to regulate energy homeostasis during muscle contraction by coupling to the myofibril (CHEN et al., 2011). Tong et al. (2015) assumed that *MYBPC1* might lead to high growth performance through enhancing muscle satellite cell proliferation. *MYBPC1* protein was differentially expressed between high and low-quality meat in *longissimus thoracis* muscle of beef cattle (WEI et al., 2019). Moreover, Tong et al. (2015) reported that expression levels of the *MYBPC1* gene was significantly higher in the high LMA group than in low LMA group, and identified a SNP in the *MYBPC1* that was associated with LMA in Japanese Black beef cattle. *CDK7* is involved in cell cycle regulation process (GO:0051726), which may influence growth, differentiation, and tissue formation during development (HEUVEL, 2005). Fernandes Júnior et al. (2016) reported that genes involved in cell cycle regulation were associated with LMA using animals from the same population as used in the present study. *AK6* is located in the same region of *TAF9* gene and they share two exons. Therefore, *AK6* and *TAF9* may interact with each other and/or may carry out the same biological functions.

Two pathways were identified influencing BF, one coming from *rs137704711* (*TAF9*, *MYBPC1*, *CAMTA1*, *VAMP3*, and *OTOS*) as previously mentioned, and other from *rs133894950* (*VAMP3*, and *OTOS*). The SNP *rs133894950* is located on chromosome 16 at 45,612,936 bp where a QTL for backfat thickness (QTL 11712) has been reported in beef cattle. *TAF9* is a transcription factor and may regulate the expression of genes that influence BF. *MYBPC1* encodes a protein that plays an important role in efficient energy metabolism and homeostasis during muscle contraction (CHEN et al., 2011). *MYBPC1* has been associated with intramuscular fat of *longissimus dorsi* in beef cattle (TONG et al., 2015). The protein encoded by *CAMTA1* participates in the calcium/calmodulin signaling process (BAS-ORTH et al., 2016), which may regulate adipocyte development (YANG et al., 2015). Silva et al. (2017) reported that a calmodulin family gene was associated with BF in Nelore cattle. *VAMP3* was shown to be expressed in rat adipocyte (CAIN; TRIMBLE; LIENHARD, 1992). No evidence of a biological link or mechanism connecting *OTOS* gene to BF was found in the literature.

Parameter estimation (causal parameters) and the stability of the network were performed conditioned on the DAG structure depicted in Figure 3.2b, and presented here only for the variables with a paths from *rs137704711* and *rs133894950* to the phenotypes (LMA, BF, and WBSF), showed in Figure 3.3a. The *rs137704711*, located in chromosome 20, had a positive total effect on LMA and WBSF, whereas *rs137704711* and *rs133894950* had a negative total effect on BF. These effects were mediated by the expression of several genes located in different chromosomes. The majority of the links and directions showed high stability, except for the relationship VAMP3 → OTOS and MYBPC1 → CDK7 → AK6 (Figure 3.3b). Removing a single data point may have caused instability in the network for some relationships among the variables studied (PEÑAGARICANO et al., 2015). However, the arrows between the genes affecting directly the phenotypes remained, in general, unchanged. The procedure applied here to reconstruct the gene-phenotype causal network may help shed light on the molecular mechanism underlying LMA, BF, and WFSF traits.

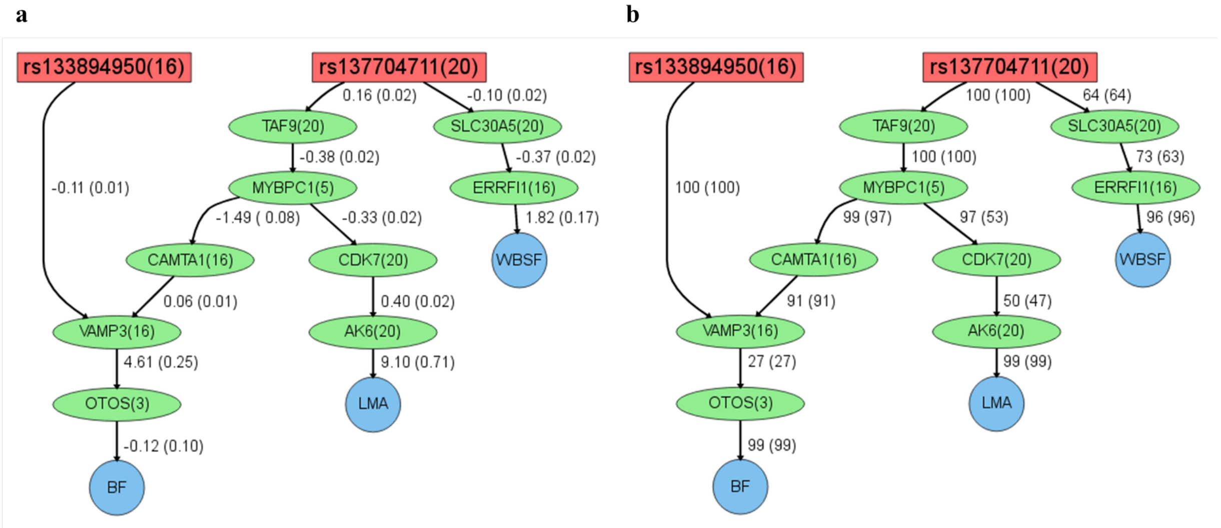

Figure 3.3 – Causal parameters and the stability of the network integrating phenotypic (blue), *cis*-eQTL (red) and transcriptomic (green) data. Point estimates (standard errors), conditional on the inferred DAG structure (Figure 3.2b), estimated by Maximum Likelihood (a), and the network stability evaluated using Jackknife resampling expressed in percentage (b) that a given arc was presented (with the same direction).

3.5 CONCLUSIONS

Integrating phenotypes, genotypes and transcriptomic data allowed identifying five co-localized genome regions that may modulate the expression of important genes that influence longissimus muscle area, backfat thickness and meat tenderness traits. The *rs137704711* affected longissimus muscle area, backfat thickness and meat tenderness traits and the *rs133894950* affected backfat thickness through the expression of many genes located in different chromosomes. Longissimus muscle area and meat tenderness were affected positively by *rs137704711*, whereas backfat thickness was affected negatively by *rs137704711* and *rs133894950*.

3.6 ACKNOWLEDGMENTS

The authors would like to acknowledge DeltaGen, PAINT and Cia de Melhoramento breeding programs for providing the data. We also would like to acknowledge the São Paulo Research Foundation (FAPESP) for supporting this work (grants: #2009/16118-5, #2016/02366-0 and #2017/18779-5). The present study was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Resources supplied by the Center for Scientific Computing (NCC/GridUNESP) of the São Paulo State University (UNESP) also sup-

ported this research.

REFERENCES

AGUILAR, I.; MISZTAL, I.; JOHNSON, D. L.; LEGARRA, A.; TSURUTA, S.; LAWLOR, T. J. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of Dairy Science**, v. 93, p. 743–752, 2010.

AINSWORTH, H. F.; SHIN, S.-Y.; CORDELL, H. J. A comparison of methods for inferring causal relationships between genotype and phenotype using additional biological measurements. **Genetic Epidemiology**, v. 41, p. 577–586, 2017.

BADSHA, M. B.; FU, A. Q. Learning causal biological networks with generalized Mendelian randomization. **Frontiers in Genetics**, v. 10, p. 1–16, 2019.

BARROWMAN, N. Correlation, causation, and confusion. **The New Atlantis Journal**, v. 43, p. 1–22, 2014.

BAS-ORTH, C.; TAN, Y. W.; OLIVEIRA, A. M.; BENGTSON, C. P.; BADING, H. The calmodulin-binding transcription activator CAMTA1 is required for long-term memory formation in mice. **Learning & Memory**, v. 23, p. 313–321, 2016.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal Royal Statistic Society**, v. 57, p. 289–300, 1995.

BERG, R. T.; BUTTERFIELD, R. M. **New concepts of cattle growth**. Sydney: Sydney University Press, 1976. 240 p.

BOITO, B.; KUSS, F.; MENEZES, L. F. G.; PARIS, E. L. M.; CULLMANN, J. R. Influence of subcutaneous fat thickness on the carcass characteristics and meat quality of beef cattle. **Ciência Rural**, v. 48, p. 1–7, 2018.

BOLORMAA, S.; PRYCE, J. E.; REVERTER, A.; ZHANG, Y.; BARENDSE, W.; KEMPER, K.; TIER, B.; SAVIN, K.; HAYES, B. J.; GODDARD, M. E. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. **Plos Genetics**, v. 10, p. 1–23, 2014.

BOUWMAN, A. C.; DAETWYLER, H. D.; CHAMBERLAIN, A. J.; PONCE, C. H.; SARGOLZAEI, M.; SCHENKEL, F. S.; SAHANA, G.; GOVIGNON-GION, A.; BOITARD, S.; DOLEZAL, M.; PAUSCH, H.; BRONDUM, R. F.; BOWMAN, P. J.; THOMSEN, B.; GULDBRANDTSEN, B.; LUND, M. S.; SERVIN, B.; GARRICK, D. J.; REECY, J.; VILKKI, J.; BAGNATO, A.; WANG, M.; HOFF, J. L.; SCHNABEL, R. D.; TAYLOR, J. F.; VINKHUYZEN, A. A. E.; PANITZ, F.; BENDIXEN, C.; HOLM, L.-E.; GREDLER, B.; HOZé, C.; BOUSSAHA, M.; SANCHEZ, M.-P.; ROCHA, D.; CAPITAN, A.; TRIBOUT, T.; BARBAT, A.; CROISEAU, P.; DRÖGEMÜLLER, C.; JAGANNATHAN, V.; JAGT, C. V.; CROWLEY, J. J.; BIEBER, A.; PURFIELD, D. C.; BERRY, D. P.; EMMERLING, R.; GÖTZ, K.-U.; FRISCHKNECHT, M.; RUSS, I.; SÖLKNER, J.; TASSELL, C. P. V.; FRIES, R.; STOTHARD, P.; VEERKAMP, R. F.; BOICHARD, D.; GODDARD, M. E.; HAYES, B. J. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. **Nature Genetics**, v. 50, p. 362–367, 2018.

BOUWMAN, A. C.; VALENTE, B. D.; JANSS, L. L. G.; BOVENHUIS, H.; ROSA, G. J. M. Exploring causal networks of bovine milk fatty acids in a multivariate mixed model context. **Genetics Selection Evolution**, v. 46, p. 1–12, 2014.

BRONDANI, I. L.; SAMPAIO, A. A. M.; RESTLE, J.; BERNARDES, R. A. L. C.; PACHECO, P. S.; FREITAS, A. K.; KUSS, F.; PEIXOTO, L. A. O. Aspectos quantitativos de carcaças de bovinos de diferentes raças, alimentados com diferentes níveis de energia. **Revista Brasileira de Zootecnia**, v. 33, p. 978–988, 2004.

CAIN, C. C.; TRIMBLE, W. S.; LIENHARD, G. E. Members of the VAMP family of synaptic vesicle proteins are components of glucose transporter-containing vesicles from rat adipocytes. **Journal of Biological Chemistry**, v. 267, p. 11681–11684, 1992.

CARVALHEIRO, R.; BOISON, S. A.; NEVES, H. H. R.; SARGOLZAEI, M.; SCHENKEL, F. S.; UTSUNOMIYA, Y. T.; O'BRIEN, A. M. P.; SÖLKNER, J.; MCEWAN, J. C.; TASSELL, C. P. V.; SONSTEGARD, T. S.; GARCIA, J. F. Accuracy of genotype imputation in Nelore cattle. **Genetics Selection Evolution**, v. 46, p. 1–11, 2014.

CASTRO, L. M.; MAGNABOSCO, C. U.; SAINZ, R. D.; FARIA, C. U.; LOPES, F. B. Quantitative genetic analysis for meat tenderness trait in Polled Nellore cattle. **Revista Ciência Agronômica**, v. 45, p. 393–402, 2014.

CESAR, A. S. M.; REGITANO, L. C. A.; REECY, J. M.; POLETI, M. D.; OLIVEIRA, P. S. N.; 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.; COUTINHO1, L. L. Identification of putative regulatory regions and transcription factors associated with intramuscular fat content traits. **BMC Genomics**, v. 19, p. 1–20, 2018.

CHAIBUB NETO, E.; KELLER, M. P.; ATTIE, A. D.; YANDELL, B. S. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. **Annals of Applied Statistics**, v. 4, p. 320–339, 2010.

CHEN, L. **Using eQTLs to reconstruct gene regulatory networks**: Quantitative Trait Loci (QTL). New York, US: Humana Press, 2012. 175-189 p.

CHEN, Z.; ZHAO, T. J.; LI, J.; GAO, Y. S.; MENG, F. G.; YAN, Y. B.; ZHOU, H. M. Slow skeletal muscle myosin-binding protein-c (MYBPC1) mediates recruitment of muscle-type creatine kinase (CK) to myosin. **Biochemical Journal**, v. 436, p. 437–445, 2011.

CHRIKI, S.; PICARD, B.; JURIE, C.; REICHSTADT, M.; BRUN, D. M. adn J. P.; JOURNAUX, L.; HOCQUETTE, J. F. Meta-analysis of the comparison of the metabolic and contractile characteristics of two bovine muscles: *Longissimus thoracis* and *semitendinosus*. **Meat Science**, v. 91, p. 423–429, 2012.

CHRIKI, S.; RENAND, G.; PICARD, B.; MICOL, D.; JOURNAUX, L.; HOCQUETTE, J. F. Meta-analysis of the relationships between beef tenderness and muscle characteristics. **Livestock Science**, v. 155, p. 424–434, 2013.

CHRISTENSEN, O. F.; LUND, M. S. Genomic prediction when some animals are not genotyped. **Genetics Selection Evolution**, v. 42, p. 1–8, 2010.

CLAYTON, D. **snpStats: SnpMatrix and XSnpMatrix classes and methods**. [S.l.], 2015. R package version 1.28.0.

de los CAMPOS, G.; GIANOLA, D.; BOETTCHER, P.; MORONI, P. A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. **Journal of Animal Science**, v. 84, p. 2934–2941, 2006.

de los CAMPOS, G.; GIANOLA, D.; HERINGSTAD, B. A structural equation model for describing relationships between somatic cell score and milk yield in first lactation dairy cows. **Journal of Dairy Science**, v. 89, p. 4445–4455, 2006.

DELGADO, E. F.; AGUIAR, A. P.; ORTEGA, E. M. M.; SPOTO, M. H. F.; CASTILLO, C. J. C. Brazilian consumers' perception of tenderness of beef steaks classified by shear force and taste. **Science Agriculture.**, v. 63, p. 232–239, 2006.

DIKEMAN, M. E.; POLLAK, E. J.; ZHANG, Z.; MOSER, D. W.; GILL, C. A.; DRESSLER., E. A. Phenotypic ranges and relationships among carcass and meat palatability traits for fourteen cattle breeds, and heritabilities and expected progeny differences for warner-bratzler shear force in three beef cattle breeds. **Journal of Animal Science**, v. 83, p. 2461–2467, 2005.

DOBIN, A.; DAVIS, C. A.; SCHLESINGER, F.; DRENKOW, J.; ZALESKI, C.; BATUT, S. J. P.; CHAISSON, M.; GINGERAS, T. R. Star: ultrafast universal rna-seq aligner. **Bioinformatics**, v. 29, p. 15–21, 2013.

DU, X. H.; GAN, Q. F.; YUAN, Z. R.; GAO, X.; ZHANG, L. P.; GAO, H. J.; LI, J. Y.; XU, S. Z. Polymorphism of MyoD1 and Myf6 genes and associations with carcass and meat quality traits in beef cattle. **Genetics Molecular Research**, v. 12, p. 6708–6717, 2013.

DURINCK, S.; SPELLMAN, P.; BIRNEY, E.; HUBER, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. **Nature Protocols**, v. 4, p. 1184–1191, 2009.

ELLIS, S. E.; GUPTA, S.; ASHAR, F. N.; BADER, J. S.; WEST, A. B.; ARKING, D. E. Rna-seq optimization with eqtl gold standards. **BMC Genomics**, v. 892, p. 1–11, 2013.

ELZO, M. A.; JOHNSON, D. D.; WASDIN, J. G.; DRIVER, J. Carcass and meat palatability breed differences and heterosis effects in an Angus–Brahman multibreed population. **Meat Science**, v. 90, p. 87–92, 2012.

FERNANDES JÚNIOR, G. A.; COSTA, R. B.; CAMARGO, G. M. F.; CARVALHEIRO, R.; ROSA, G. J. M.; BALDI, F.; GARCIA, D. A.; GORDO, D. G. M.; ESPIGOLAN, R.; TAKADA, L.; aES, A. F. B. M.; BRESOLIN, T.; FEITOSA, F. L. B.; CHARDULO, L. A. L.; OLIVEIRA, H. N.; ALBUQUERQUE, L. G. Genome scan for postmortem carcass traits in Nellore cattle. **Journal of Animal Science**, v. 94, p. 4087–4095, 2016.

FISHER, R. A. The arrangement of field experiments. **Journal of the Ministry of Agriculture**, v. 33, p. 503–513, 1926.

FISHER, R. A. **The design of experiments**. New York: Hafner Publishing Company, 1971.

FONSECA, L. F. S.; GIMENEZ, D. F. J.; SILVA, D. B. S.; BARTHELSON, R.; BALDI, F.; FERRO, J. A.; ALBUQUERQUE, L. G. Differences in global gene expression in muscle tissue of Nellore cattle with divergent meat tenderness. **BMC Genomics**, v. 18, p. 1–12, 2017.

GIANOLA, D.; SORENSEN, D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. **Genetics**, v. 167, p. 1407–1424, 2004.

GILAD, Y.; SCOTT, A. R.; JONATHAN, K. P. Revealing the architecture of gene regulation: the promise of eqtl studies. **Trends in Genetics**, v. 24, p. 408–415, 2008.

GORDO, D. G. M.; ESPIGOLAN, R.; BRESOLIN, T.; JúNIOR, G. A. F.; aES, A. F. B. M.; BRAZ, C. U.; FERNANDES, W.; BALDI, F.; ALBUQUERQUE, L. G. Genetic analysis of carcass and meat quality traits in Nelore cattle. **Journal of Animal Science**, v. 96, p. 3558–3564, 2018.

GORDO, D. G. M.; ESPIGOLAN, R.; TONUSSI, R. L.; JúNIOR, G. A. F.; BRESOLIN, T.; aES, A. F. B. M.; FEITOSA, F. L.; BALDI, F.; CARVALHEIRO, R.; TONHATI, H.; OLIVEIRA, H. N.; CHARDULO, L. A. L.; ALBUQUERQUE, L. G. Genetic parameter estimates for carcass traits and visual scores including or not genomic information. **Journal of Animal Science**, v. 94, p. 1821–1826, 2016.

GOTOH, T.; TAKAHASHI, H.; NISHIMURA, T.; KUCHIDA, K.; MANNEN, H. Meat produced by Japanese Black cattle and Wagyu. **Animal Frontiers**, v. 4, p. 46–54, 2014.

GUILLEMIN, N.; BONNET, M.; JURIE, C.; PICARD, B. Functional analysis of beef tenderness. **Journal of Proteomics**, v. 75, p. 352–365, 2011.

HAAVELMO, T. The statistical implications of a system of simultaneous equations. **Econometrica**, v. 11, p. 1–12, 1943.

HAGEMAN, R. S.; LEDUC, M. S.; KORSTANJE, R.; PAIGEN, B.; CHURCHILL, G. A. A bayesian framework for inference of the genotype–phenotype map for segregating populations. **Genetics**, v. 187, p. 1163–1170, 2011.

HANDEL, S. E.; STICKLAND, N. C. Catch-up growth in pigs: relationship with muscle cellularity. **Animal Production**, v. 47, p. 291–295, 1988.

HASIN, Y.; SELDIN, M. L. Multi-omics approaches to disease. **Genome Biology**, v. 83, p. 1–15, 2017.

HAY, E.-H.; ROBERTS, A. Genome-wide association study for carcass traits in a composite beef cattle breed. **Livestock Science**, v. 213, p. 35–43, 2018.

HENDERSON, C. R. Multiple trait sire evaluation using the relationship matrix. **Journal of Dairy Science**, v. 59, p. 769–774, 1976.

HENDERSON, C. R.; KEMPTHORNE, O.; SEARLE, S. R.; vonKrosigk, C. M. The estimation of environmental and genetic trends from records subject to culling. **Biometrics**, v. 15, p. 192–218, 1959.

HENDERSON, C. R.; QUAAS, R. L. Multiple trait evaluation using relative records. **Journal of Animal Science**, v. 43, p. 1188–1197, 1976.

HERINGSTAD, B.; WU, X. L.; GIANOLA, D. Inferring relationships between health and fertility in Norwegian Red cows using recursive models. **Journal of Dairy Science**, v. 92, p. 1778–1784, 2009.

HEUVEL, S. v. **Cell-cycle regulation**: Wormbook: The online review of c. elegans biology. Pasadena, CA: WormBook, 2005.

HIGGINS, M. G.; FITZSIMONS, C.; MCCLURE, M. C.; MCKENNA, C.; CONROY, S.; KENNY, D. A.; MCGEE, M.; WATERS, S. M.; MORRIS, D. W. Gwas and eQTL analysis identifies a SNP associated with both residual feed intake and GFRA2 expression in beef cattle. **Plos One**, v. 8, p. 1–12, 2018.

HOCQUETTE, J. F.; BOTREAU, R.; PICARD, B.; PETHICK, A. J. D. W.; SCOLLAN, N. D. Opportunities for predicting and manipulating beef quality. **Meat Science**, v. 92, p. 197–209, 2012.

HU, Z.-L.; PARK, C. A.; WU, X.-L.; REECY, J. M. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. **Nucleic Acids Research**, v. 41, p. D871–D879, 2013.

HUANG, Y.; ZHENG, J.; PRZYTYCKA, T. M. **Discovery of regulatory mechanisms from gene expression variation by eQTL analysis**: Biological data mining. Boca Rotan, London, NY: CRC Press, 2010.

INNOCENTI, F.; COOPER, G. M.; STANAWAY, I. B.; GAMAZON, E. R.; SMITH, J. D.; MIRKOV, S.; RAMIREZ, J.; LIU, W.; LIN, Y. S.; MOLONEY, C.; ALDRED, S. F.; TRINKLEIN, N. D.; SCHUETZ, E.; NICKERSON, D. A.; THUMMEL, K. E.; RIEDER, M. J.; RETTIE, A. E.; RATAIN, M. J.; COX, N. J.; BROWN, C. D. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. **PLoS Genetics**, v. 7, p. 1–16, 2011.

INOUE, K.; HOSONO, M.; TANIMOTO, Y. Inferring causal structures and comparing the causal effects among calving difficulty, gestation length and calf size in Japanese black cattle. **Animal**, v. 11, p. 2120–2128, 2017.

INOUE, K.; VALENTE, B. D.; SHOJI, N.; HONDA, T.; OYAMA, K.; ROSA, G. J. M. Inferring phenotypic causal structures among meat quality traits and the application of a structural equation model in Japanese black cattle. **Journal of Animal Science**, v. 94, p. 4133–4142, 2016.

KADARMIDEEN, H. N.; VON ROHR, P.; JANSS, L. L. From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. **Mammalian Genome**, v. 17, p. 548–564, 2006.

KARISA, B. K.; THOMSON, J.; WANG, Z.; BRUCE, H. L.; PLASTOW, G. S.; MOORE, S. S. Candidate genes and biological pathways associated with carcass quality traits in beef cattle. **Canedian Journal of Animal Science**, v. 93, p. 295–306, 2013.

KIM, Y.; RYU, J.; WOO, J.; KIM, J. B.; KIM, C. Y.; LEE, C. Genome-wide association study reveals five nucleotide sequence variants for carcass traits in beef cattle. **Animal Genetics**, v. 42, p. 361–365, 2011.

KOLLER, D.; FRIEDMAN, N. **Probabilistic Graphical Models: Principles and Techniques**. London, England: MIT Press, 2009. 1215 p.

KOLLER, D.; SAHAMI, M. Toward optimal feature selection. In: THIRTEENTH INTERNATIONAL CONFERENCE IN MACHINE LEARNING, 1996. **Proceeding...** [S.l.], 1996.

KOOHMARAIE, M.; KENT, M. P.; SHACKELFORD, S. D.; VEISETH, E.; WHEELER, T. L. Meat tenderness and muscle growth: is there any relationship? **Meat Science**, v. 62, p. 345–52, 2002.

KORB, K. B.; NICHOLSON, A. E. **Bayesian Artificial Intelligence**. London, New York: CRC Press, 2010. 491 p.

LAIRD, N. M.; WARE, J. H. Random effects models for longitudinal data. **Biometrics**, v. 38, p. 963–974, 1982.

LEAL-GUTIÉRREZ, J. D.; ELZO, F. M. R. M. A.; JOHNSON, D.; PEÑAGARICANO, F.; MATEESCU, R. G. Structural equation modeling and whole-genome scans uncover chromosome regions and enriched pathways for carcass and meat quality in beef. **Frontiers in Genetics**, v. 11, p. 1–13, 2018.

LIU, X.; DU, Y.; TRAKOOLJUL, N.; BRAND, B.; MURÁNI, E.; KRISCHEK, C.; WICKE, M.; SCHWERIN, M.; WIMMERS, K.; PONSUKSILI, S. Muscle transcriptional profile based on muscle fiber, mitochondrial respiratory activity, and metabolic enzymes. **International Journal of Biological Sciences**, v. 11, p. 1348–1362, 2015.

LU, D.; SARGOLZAEI, M.; KELLY, M.; VOORT, G. V.; WANG, Z.; MANDELL, I.; MOORE, S.; PLASTOW, G.; MILLER, S. P. Genome-wide association analyses for carcass quality in crossbred beef cattle. **BMC Genetics**, v. 14, p. 1–10, 2013.

MACKAY, T. F. C.; STONE, E. A.; AYROLES, J. F. The genetics of quantitative traits: challenges and prospects. **Nature Reviews Genetics**, v. 10, p. 565–577, 2009.

MAGALHÃES, A. F. B.; CAMARGO, G. M. F.; JÚNIOR, G. A. F.; GORDO, D. G. M.; TONUSSI, R. L.; COSTA, R. B.; ESPIGOLAN, R.; SILVA, R. M. O.; BRESOLIN, T.; ANDRADE, W. B. F.; TAKADA, L.; FEITOSA, F. L. B.; BALDI, F.; CARVALHEIRO, R.; CHARDULO, L. A. L.; ALBUQUERQUE, L. G. Genome-wide association study of meat quality traits in Nellore cattle. **Plos One**, v. 1, p. 1–12, 2016.

MALECOVA, B.; DALL’AGNESE, A.; MADARO, L.; GATTO, S.; TOTO, P. C.; ALBINI, S.; RYAN, T.; TORA, L.; PURI, P. L. TBP/TFIID-dependent activation of MyoD target genes in skeletal muscle cells. **eLife**, v. 5, p. 1–18, 2016.

MANOLIO, T. A.; COLLINS, F. S.; COX, N. J.; GOLDSTEIN, D. B.; HINDORFF, L. A.; HUNTER, D. J.; MCCARTHY, M. I.; RAMOS, E. M.; CARDON, L. R.; CHAKRAVARTI, A.; CHO, J. H.; GUTTMACHER, A. E.; KONG, A.; KRUGLYAK, L.; MARDIS, E.; ROTIMI, C. N.; SLATKIN, M.; VALLE, D.; WHITTEMORE, A. S.; BOEHNKE, M.; CLARK, A. G.; EICHLER, E. E.; GIBSON, G.; HAINES, J. L.; MACKAY, T. F. C.; MCCARROLL, S. A.; VISSCHER, P. M. Finding the missing heritability of complex diseases. **Nature**, v. 461, p. 747–753, 2009.

MARGARITIS, D.; THRUN, S. Bayesian network induction via local neighborhoods. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEM, 1999. **Proceeding...** [S.l.], 1999.

MARTYN, J. K.; BASS, J. J.; OLDHAM, J. M. Skeletal muscle development in normal and double-muscled cattle. **The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology**, v. 281, p. 1363–1371, 2004.

MATURANA, E. L.; WU, X.-L.; GIANOLA, D.; WEIGEL, K. A.; ROSA, G. J. M. Exploring biological relationships between calving traits in primiparous cattle with a bayesian recursive model. **Genetics**, v. 181, p. 277–287, 2009.

MISZTAL, I.; TSURUTA, S.; LOURENCO, D. A. L.; MASUDA, Y.; AGUILAR, I.; LEGARRA, A.; VITEZICA, Z. **Manual for BLUPF90 family programs**. Georgia, USA, 2018. Disponível em: <http://nce.ads.uga.edu/wiki/doku.php?id=documentation>.

MONTGOMERY, S. B.; DERMITZAKIS, E. T. From expression qtls to personalized transcriptomics. **Nature Reviews Genetics**, v. 12, p. 277–282, 2011.

MOROTA, G.; VALENTE, B. D.; ROSA, G. J. M.; WEIGEL, K. A.; GIANOLA, D. An assessment of linkage disequilibrium in holstein cattle using a bayesian network. **Journal of Animal Breeding and Genetics**, v. 129, p. 474–487, 2012.

NEAPOLITAN, R. E. **Learning Bayesian Networks**. Chicago, Illinois: Northeastern Illinois University, 2003. 686 p.

NI, Y.; STINGO, F. C.; BALADANDAYUTHAPANI, V. Integrative bayesian network analysis of genomic data. **Cancer Informatics**, v. 13, p. 39–48, 2014.

NICA, A. C.; DERMITZAKIS, E. T. Expression quantitative trait loci: presente and future. **Philosophical Transaction of the Royal Society**, v. 368, p. 1–6, 2013.

NURNBERG, K.; WEGNER, J.; ENDER, K. Factors influencing fat composition in muscle and adipose tissue of farm animals. **Livestock Production Science**, v. 56, p. 145–156, 1998.

OLIVEIRA, I. M.; PAULINO, P. V. R.; MARCONDES, M. I.; FILHO, S. C. V.; DETMANN, E.; CAVALI, J.; DUARTE, M. S.; MEZZOMO, R. Pattern of tissue deposition, gain and body composition of Nellore, F1 Simmental x Nellore and F1 Angus x Nellore steers fed at maintenance or *ad libitum* with two levels of concentrate in the diet. **Revista Brasileira de Zootecnia**, v. 40, p. 2886–2893, 2011.

O'CONNOR, S. F.; TATUM, J. D.; WULF, D. M.; GREEN, R. D.; SMITH, G. C. Genetic effects on beef tenderness in *Bos indicus* composite and *Bos taurus* cattle. **Journal of Animal Science**, v. 75, p. 1822–1830, 1997.

PEARL, J. **Probabilistic Reasoning in Intelligent System: Networks of Plausible Inference**. San Francisco, CA: Morgan Kaufmann Publishers, INC, 1988.

PEARL, J. **Causality: Models, Reasoning and Inference**. Cambridge, UK: Cambridge University Press, 2000.

PEARL, J. An introduction to causal inference. **The International Journal of Biostatistics**, v. 6, p. 1–61, 2010.

PEÑAGARICANO, F.; VALENTE, B. D.; STEIBEL, J. P.; ERNST, R. O. B. C. W.; ROSA, H. K. G. J. M. Exploring causal networks underlying fat deposition and muscularity in pigs through the integration of phenotypic, genotypic and transcriptomic data. **BMC Systems Biology**, v. 9, p. 1–9, 2015.

PEREIRA, A. S. C.; BALDI, F.; SAINZ, R. D.; UTEMBERGUE, B. L.; CHIAIA, H. L. J.; MAGNABOSCO, C. U.; MANICARDI, F. R.; ARAUJO, F. R. C.; GUEDES, C. F.; MARGARIDO, R. C.; LEME, P. R.; SOBRAL, P. J. A. Growth performance, and carcass and meat quality traits in progeny of Poll Nellore, Angus and Brahman sires under tropical conditions. **Animal Production Science**, v. 55, p. 1295–1302, 2015.

PICARD, B.; GAGAOUA, M.; MICOL, D.; CASSAR-MALEK, I.; HOCQUETTE, J. F.; TERLOUW, C. E. M. Inverse relationships between biomarkers and beef tenderness according to contractile and metabolic properties of the muscle. **Journal of Agricultural and Food Chemistry**, v. 40, p. 9808–9818, 2014.

PLUMMER, M.; BEST, N.; COWLES, K.; VINES, K. Coda: Convergence diagnosis and output analysis for MCMC. **R News**, v. 6, p. 7–11, 2006.

R Core Team. **R: A language and environment for statistical computing**. Vienna, Austria, 2017. Disponível em: <http://www.R-project.org/>.

REVERTER, A. A.; JOHNSTON, D. J. A.; FERGUSON, D. M. B.; PERRY, D. C.; GODDARD, M. E. A.; BURROW, H. M. E.; ODDY, V. H. F.; THOMPSON, J. M. G.; BINDON, B. M. H. Genetic and phenotypic characterisation of animal, carcass, and meat quality traits from temperate and tropically adapted beef breeds. 4. correlations among animal carcass and meat quality traits. **Australian Journal of Agricultural Research**, v. 54, p. 149–158, 2003.

REZENDE, P. L. P.; RESTLE, J.; FERNANDES, J. J. R.; NETO, M. D. F.; PRADO, C. S.; PEREIRA, M. L. R. Carcass and meat characteristics of crossbred steers submitted to different nutritional strategies at growing and finishing phases. **Ciência Rural**, v. 45, p. 875–881, 2012.

RILEY, D. G.; JR, C. C. C.; HAMMOND, A. C.; WEST, R. L.; JOHNSON, D. D.; OLSON, T. A.; COLEMAN, S. W. Estimated genetic parameters for carcass traits of Brahman cattle. **Journal of Animal Science**, v. 80, p. 955–962, 2002.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edger: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, v. 26, p. 139–140, 2010.

ROSA, G. J. M.; VALENTE, B. D. Breeding and genetics symposium: Inferring causal effects from observational data in livestock. **Journal of Animal Science**, v. 91, p. 553–564, 2013.

ROSA, G. J. M.; VALENTE, B. D.; CAMPOS, G. de los; WU, X.-L.; GIANOLA, D.; SILVA, M. A. Inferring causal phenotype networks using structural equation models. **Genetics Selection Evolution**, v. 43, p. 1–13, 2011.

ROSENBAUM, P. R. **Design of Observational Studies**. New York: Springer, 2010. 382 p.

SANTIAGO, G. G.; SIQUEIRA, F.; CARDOSO, F. F.; REGITANO, L. C. A.; VENTURA, R.; SOLLERO, B. P.; SOUZA, J. M. D.; MOKRY, F. B.; FERREIRA, A. B. R.; TORRES, R. A. A. j. Genomewide association study for production and meat quality traits in canchim beef cattle. **Journal of Animal Science**, v. 95, p. 3381–3390, 2017.

SARGOLZAEI, M.; CHESNAIS, J. P.; SCHENKEL, F. S. A new approach for efficient genotype imputation using information from relatives. **BMC Genomics**, v. 15, p. 1–12, 2014.

SCHAEFFER, L. R. Sire and cow evaluation under multiple trait models. **Journal of Dairy Science**, v. 67, p. 1567–1580, 1984.

SCUTARI, M. Learning bayesian networks with the bnlearn R package. **Journal of Statistical Software**, v. 35, p. 1–22, 2010.

SCUTARI, M.; STRIMMER, K. **Introduction to graphical modelling**: Handbook of statistical systems biology. Chichester, UK: John Wiley & Sons, Ltd, 2011.

SHIPLEY, B. **Cause and Correlation in Biology**. Cambridge, London, New York: Cambridge University Press, 2002. 316 p.

SILVA, R. M. O.; STAFUZZA, N. B.; FRAGOMENI, B. O.; CAMARGO, G. M. F.; CEACERO, T. M.; CYRILLO, J. N. S. G.; BALDI, F.; BOLIGON, A. A.; MERCADANTE, M. E. Z.; LOURENCO, D. L.; MISZTAL, I.; ALBUQUERQUE, L. G. Genome-wide association study for carcass traits in an experimental Nelore cattle population. **Plos One**, v. 12, p. 1–14, 2017.

SMITH, T.; DOMINGUE, J. D.; PASCHAL, J. C.; FRANKE, D. E.; BIDNER, T. D.; WHIPPLE, G. Genetic parameters for growth and carcass traits of Brahman steers. **Journal of Animal Science**, v. 85, p. 1377–1384, 2007.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC Methods in Quantitative Genetics**. New York: Springer, 2002.

SPIEGELHALTER, D. J.; BEST, N. G.; CARLIN, B. P.; LINDE, A. van der. Bayesian measures of model complexity and fit (with discussion). **Journal of the Royal Statistical Society**, v. 64, p. 583–639, 2002.

SPIRTES, P. Introduction to causal inference. **Journal of Machine Learning Research**, v. 11, p. 1643–1662, 2010.

STEIBEL, J. P.; BATES, R. O.; ROSA, G. J. M.; TEMPELMAN, R. J.; RILINGTON, V. D.; RAGAVENDRAN, A.; RANEY, N. E.; RAMOS, A. M.; CARDOSO, F. F.; EDWARDS, D. B.; ERNST, C. W. Genome-wide linkage analysis of global gene expression in loin muscle tissue identifies candidate genes in pigs. **Plos One**, v. 6, p. 1–11, 2011.

te PAS, M. F. W.; SOUMILLION, A.; HARDERS, F. L.; VERBURG, F. J.; van den Bosch, T. J.; GALESLOOT, P.; MEUWISSEN, T. H. E. Influences of myogenin genotypes on birth weight, growth rate, carcass weight, backfat thickness, and lean weight of pigs. **Journal of Animal Science**, v. 77, p. 2352–2356, 1999.

TIZIOTO, P. C.; DECKER, J. E.; TAYLOR, J. F.; SCHNABEL, R. D.; MUDADU, M. A.; SILVA, F. L.; MOURÃO, G. B.; COUTINHO, L. L.; THOLON, P.; SONSTEGARD, T. S.; ROSA, A. N.; ALENCAR, M. M.; TULLIO, R. R.; MEDEIROS, S. R.; NASSU, R. T.; FEIJÓ, G. L. D.; SILVA, L. O. C.; TORRES, R. A.; SIQUEIRA, F.; HIGA, R. H.; REGITANO, L. C. A. Genome scan for meat quality traits in Nelore beef cattle. **Physiology Genomics**, v. 45, p. 1012–1020, 2013.

TONG, B.; XING, Y. P.; MURAMATSU, Y.; OHTA, T.; KOSE, H.; ZHOU, H. M.; YAMADA, T. Association of expression levels in skeletal muscle and a SNP in the MYBPC1 gene with growth-related trait in Japanese Black beef cattle. **Journal of Genetics**, v. 94, p. 135–137, 2015.

TONUSSI, R. L.; ESPIGOLAN, R.; GORDO, D. G. M.; MAGALHÃES, A. F. B.; VENTURINI, G. C.; BALDI, F.; OLIVEIRA, H. N.; CHARDULO, L. A. L.; TONHATI, H.; ALBUQUERQUE, L. G. Genetic association of growth traits with carcass and meat traits in Nellore cattle. **Genetics and Molecular Research**, v. 14, p. 18713–18719, 2015.

TSAMARDINOS, I.; ALIFERIS, C. F.; STATNIKOV, A. Algorithms for large scale markov blanket discovery. In: 16TH INTERNATIONAL FLORIDA ARTIFICIAL INTELLIGENCE RESEARCH SOCIETY CONFERENCE, 2003, California. **Proceeding...** California, 2003.

VALENTE, B. D.; ROSA, G. J. M. **Mixed effects structural equation models and phenotypic causal networks**: Genome-wide association studies and genomic prediction. New York, US: Humana Press, 2013. 449-464 p.

VALENTE, B. D.; ROSA, G. J. M.; CAMPOS, G. de los; GIANOLA, D.; SILVA, M. A. Searching for recursive causal structures in multivariate genetics mixed models. **Genetics**, v. 185, p. 633–644, 2010.

VALENTE, B. D.; ROSA, G. J. M.; GIANOLA, D.; WU, X. L.; WEIGEL, K. Is structural equation modeling advantageous for the genetic improvement of multiple traits? **Genetics**, v. 194, p. 561–572, 2013.

VALENTE, B. D.; ROSA, G. J. M.; SILVA, M. A.; TEIXEIRA, R. B.; TORRES, R. A. Searching for phenotypic causal networks involving complex traits: An application to European quail. **Genetics Selection Evolution**, v. 43, p. 1–12, 2011.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, p. 4414–4423, 2008.

VARONA, L.; SORENSEN, D.; THOMPSON, R. Analysis of litter size and average litter weight in pigs using a recursive model. **Genetics**, v. 177, p. 1791–1799, 2007.

VERMA, T.; PEARL, J. Equivalence and synthesis of causal models. In: 6TH CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 1990, Cambridge. **Proceeding...** Cambridge, 1990.

WEI, Y.; LI, X.; ZHANG, D.; LIU, Y. Comparison of protein differences between high- and low-quality goat and bovine parts based on iTRAQ technology. **Food Chemistry**, v. 289, p. 240–249, 2019.

WHEELER, T. L.; KOOHMARAIE, M.; SHACKELFORD, S. D. **Standardized Warner-Bratzler shear force procedures for meat tenderness measurement**. [S.l.], 1995.

WIEDERMANN, W.; DONG, N.; vonEye, A. Advances in statistical methods for causal inference in prevention science: Introduction to the special section. **Prevention Science**, v. 20, p. 390–393, 2019.

WRIGHT, S. Correlation and causation. **Journal of Agricultural Research**, v. 201, p. 557–585, 1921.

WU, X.-L.; HERINGSTAD, B.; GIANOLA, D. Exploration of lagged relationships between mastitis and milk yield in dairy cows using a Bayesian structural equation Gaussian-threshold model. **Genetic Selection Evolution**, v. 40, p. 333–357, 2008.

WU, X.-L.; HERINGSTAD, B.; GIANOLA, D. Bayesian structural equation models for inferring relationships between phenotypes: a review of methodology, identifiability, and applications. **Journal of Animal Breeding and Genetics**, v. 127, p. 3–15, 2010.

YANG, J.; LEE, S. H.; GODDARD, M. E.; VISSCHER, P. M. Gcta: a tool for genome-wide complex trait analysis. **American Journal of Human Genetics**, v. 88, p. 76–82, 2011.

YANG, Y.; SONG, J.; FU, R.; SUN, Y.; WEN, J. The expression of can and camk is associated with lipogenesis in the muscle of chicken. **Brazilian Journal of Poultry Science**, v. 17, p. 287–292, 2015.

YANG, Y. I.; RONG, Z.; KUI, L. Future livestock breeding: Precision breeding based on multi-omics information and population personalization. **Journal of Integrative Agriculture**, v. 16, p. 2784–2791, 2017.

YARAMAKALA, S.; MARGARITIS, D. Speculative markov blanket discovery for optimal feature selection. In: FIFTH IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 2005, Washington, DC. **Proceeding...** Washington, DC, 2005.

ZHANG, N. L.; POOLE, D. Exploiting causal independence in bayesian network inference. **Journal of Artificial Intelligence Research**, v. 5, p. 301–328, 1996.