University of São Paulo
"Luiz de Queiroz" College of Agriculture

Using graphical models to investigate phenotypic networks involving polygenic traits

Renan Mercuri Pinto

Thesis presented to obtain the degree of Doctor in Science. Area: Statistics and Agricultural Experimentation

Piracicaba
2018

Renan Mercuri Pinto
Degree in Mathematics

Using graphical models to investigate phenotypic networks involving polygenic traits

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **ROSELI APARECIDA LEANDRO**

Thesis presented to obtain the degree of Doctor in Science.
Area: Statistics and Agricultural Experimentation

Piracicaba
2018

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Pinto, Renan Mercuri

Using graphical models to investigate phenotypic networks involving polygenic traits/ Renan Mercuri Pinto. – – versão revisada de acordo com a resolução CoPGr 6018 de 2011. – – Piracicaba, 2018 .

99 p.

Tese (Doutorado) – – USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Redes causais 2. Modelos de equações estruturais 3. Redes Bayesianas. I. Título.

DEDICATION

To my family,
for all support and encouragement.

ACKNOWLEDGMENTS

Firstly, I thank my family (Adriana Cristina Mercuri Pinto, Osvail Aparecido Pinto e Laís Mercuri Pinto) for having fed me with good manners and taught me to fight for my dreams with perspicacity and patience. Also, to my girlfriend (Daiane Biscioni Von Zuben) and her family for all support.

I am grateful for the essential guidance of three professors from the University of Wisconsin (UW) - Madison, Dr. Guilherme J. M. Rosa, Dr. Bruno D. Valente and Dr. Fernando L. Brito, without which this work would not be possible.

To my advisor Dr. Roseli A. Leandro, from the University of Sao Paulo (USP) - ESALQ, for all support and for making my sandwich period at UW-Madison possible.

To the professors of the Department of Statistics and Agronomic Experimentation at USP-ESALQ, especially Dr. Taciana V. Savian and Dr. Idemauro R. de Lara, and also to the professors of the Department of Animal Science at UW-Madison.

To my special friends from USP-ESALQ (Ana Julia Righetto, Cesar Augusto Jotta, Hiron Pereira Farias, Luiz Ricardo Nakamura, Pedro Henrique Cerqueira, Rafael Moral, Reginaldo Francisco Hilário, Ricardo Klein, Rodrigo Pescim, Thiago Gentil Ramires) and UW-Madison (André Galo, Arthur Fernandes, Ashley Mikshowsky, Camila Barros, Katrin Topner, Ligia Moreira, Mahmoud Amiri, Marília Ferreira, Mehdi Momen, Nicole Gross, Tiago Passafaro, Vera Ferreira).

To the staff of the Department of Exact Science at USP-ESALQ, in special the secretary Solange Sabadin and the computer technicians Eduardo Bonilha and Jorge Wiendl.

To CAPES and CNPq, for financial support.

CONTENTS

RESUMO

O uso de modelos gráficos para investigar redes fenotípicas envolvendo características poligênicas

Compreender a arquitetura causal subjacente à sistemas biológicos complexos é de grande valia na produção agrícola para o desenvolvimento de estratégias de manejo e seleção genética. Até o momento, a maior parte dos estudos neste contexto utiliza apenas conhecimento prévio para propor redes causais e/ou não considera fatores de confundimento genético na busca de estruturas, fato que pode ocultar relações importantes entre os fenótipos e viesar inferências sobre a rede causal. Nesta tese, exploramos alguns algoritmos de aprendizagem de estruturas e apresentamos um novo, chamado PolyMaGNet (do inglês, Polygenic traits with Major Genes Network analysis), para buscar estruturas causais recursivas entre características fenotípicas poligênicas complexas e permitindo, também, a possibilidade de efeitos de genes maiores que as afetam. Resumidamente, um modelo misto de múltiplas características é ajustado usando abordagem Bayesiana considerando os genes maiores como covariáveis no modelo. Em seguida, amostras posteriores da matriz de covariância residual são usadas como entrada para o algoritmo de causação indutiva para pesquisar estruturas causais putativas, as quais são comparadas usando o critério de informação de Akaike. O desempenho do PolyMaGNet foi avaliado e comparado com outra abordagem bastante utilizada por meio de um estudo simulado considerando uma população de mapeamento de QTL. Os resultados mostraram que, na presença de genes maiores, o método PolyMaGNet recuperou a verdadeira estrutura do esqueleto, bem como as direções causais, com uma taxa de efetividade maior. O método é ilustrado também utilizando-se um conjunto de dados reais de uma população de suínos F2 Duroc $\times$ Pietrain para recuperar a estrutura causal subjacente à características fenotípicas relacionadas a qualidade da carcaça, carne e composição química. Os resultados corroboraram com a literatura sobre as relações de causa-efeito entre os fenótipos e também forneceram novos conhecimentos sobre a rede fenotípica e sua arquitetura genética.

Palavras-chave: Redes causais; Modelos de equações estruturais; Redes Bayesianas

ABSTRACT

Using graphical models to investigate phenotypic networks involving polygenic traits

Understanding the causal architecture underlying complex systems biology has a great value in agriculture production for the development of optimal management strategies and selective breeding. So far, most studies in this area use only prior knowledge to propose causal networks and/or do not consider the possible genetic confounding factors on the structure search, which may hide important relationships among phenotypes and also bias the resulting inferred causal network. In this dissertation, we explore many structural learning algorithms and present a new one, called PolyMaGNet (Polygenic traits with Major Genes Network analysis), to search for recursive causal structures involving complex phenotypic traits with polygenic inheritance and also allowing the possibility of major genes affecting the traits. Briefly, a multiple-trait animal mixed model is fitted using a Bayesian approach considering major genes as covariates. Next, posterior samples of the residual covariance matrix are used as input for the Inductive Causation algorithm to search for putative causal structures, which are compared to each other using the Akaike information criterion. The performance of PolyMaGNet was evaluated and compared with another widely used approach in a simulated study considering a QTL mapping population. Results showed that, in the presence of major genes, our method recovered the true skeleton structure as well as the causal directions with a higher rate of true positives. The PolyMaGNet approach was also applied to a real dataset of an F2 Duroc $\times$ Pietrain pig resource population to recover the causal structure underlying on carcass, meat quality and chemical composition traits. Results corroborated with the literature regarding the cause-effect relationships between these traits and also provided new insights about phenotypic causal networks and its genetic architectures in complex systems biology.

Keywords: Causal networks; Structural equation models; Bayesian networks

1 INTRODUCTION

In agriculture, relationships among phenotypic traits are commonly studied using the standard Multiple Trait Model (MTM; Henderson and Quaas, 1976; Mrode, 2005). Although such models can be used to infer how likely events are, they are not stable enough to predict how the probabilities could vary as a result of external interventions (Pearl, 2000; Rosa et al., 2011; Shipley, 2016). For example, a correlation detected between two traits T1 and T2 may be due to a direct effect of T1 on T2 (or vice versa) or to a latent variables affecting both together. Knowledge of the underlying phenotypic causal network is essential to predict the effect of management practices applied to both traits. That is, if T1 affects T2, but T2 has no effect on T1, an intervention on T1 would change T2, but the reverse would not happen.

Similar scenarios occur in genetic improvement, where genetic correlation is defined as the proportion of variance that two traits share due to genetic causes (Rosa et al., 2011). In classical genetics many genes are known to have multiple effects, this action is called pleiotropy. For example, the vestigial gene in Drosophila is responsible for affecting not only the bristles and the wings, but also the fecundity (Mode and Robinson, 1959).

According to Schadt et al. (2005) there are three possible causal relationships (Figure 1.1) involving a gene (G) and two phenotypic traits (T1 and T2), which are explored in details by Li et al. (2006). In the first case (Figure 1.1a), the gene G affects phenotype T1, and the phenotypic change on T1 affects T2; in the second (Figure 1.1b), the gene G acts on T2, and the phenotypic change on T2 changes T1; and, in the third (Figure 1.1c) the gene G changes both traits directly, which may or may not have a causal effect between them.

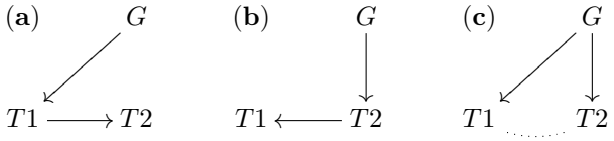

Figure 1.1. Possible gene-phenotypic networks involving one gene (G) and two phenotypic traits (T1 and T2) - [adapted from Rosa et al., 2011]

The traditional MTM mentioned above might detect a correlation between two traits and possibly the pleiotropic effect of the gene, but would not be able to distinguish the configuration of the paths that connect them. As an alternative, graphical models can be used to study recursive and simultaneous relationships among variables in multivariate systems, offering a different interpretation of the relationships between traits, such that one trait can be considered as a predictor of another, providing a causal path between them (Haavelmo, 1943; Wright, 1921; Rosa et al., 2011).

Graphical models provide a qualitative representation of biological systems along with the quantitative analysis. The qualitative part is a graphical representation of dependences among variables, expressed, for instance, by a directed acyclic graph (DAG), an undirected graph (UG) or a partially directed acyclic graph (PDAG), that is, a graph which may have both directed and undirected edges. Knowledge about the quantitative dependences between variables is added in the graph by means of path coefficients (parameters). There are many kinds of graphical models in the literature, in this thesis we will cover the most popular ones: structural equation models (SEM) and Bayesian networks (BN).

Gianola and Sorensen (2004) described SEMs in the context of mixed models in quantitative genetics, and since then many authors have used such approach (De los Campos et al., 2006; de Maturana et al., 2009), but usually causal structures are pre-selected using some kind of prior knowledge. Here, we propose to use the notion of directional separation (d-separation; Spirtes et al., 2000; Pearl, 2000;

Shipley, 2016) to explore the space of causal hypotheses and thus arrive at a causal structure (or a class of equivalent structures) that is capable of generating the observed pattern of conditional probabilities among variables.

Seeking to offer new methods to researchers who use only prior knowledge to propose causal networks, in chapter 2 we merge structural learning algorithms along with prior knowledge to investigate the causal networks underlying phenotypic traits of two fruit species belonging to the Sapotaceae family. We use a constrain- and a score-based algorithm for structure learning of BNs and select a putative causal network from the equivalence class of the score-based output, which are fitted using maximum likelihood and evaluated through some fit indices of SEM.

When considering only phenotypic data in the structural search, as in Chapter 2, we are subject to obtain networks biased by genetic confounding factors. For instance, if there is a genetic variable not considered in the search that affects two traits simultaneously (Figure 1.1c), we may find a causal path between the two traits that does not actually exist. Thus, if high density molecular marker data is available, more reliable causal networks can be obtained through more efficient genetic prediction approaches.

The use of high-density single nucleotide polymorphism (SNP) panels has increased significantly the quantitative trait loci (QTL) mapping resolution and its applications have extended to outbred populations. In Chapter 3, using 35 phenotypic traits measured in a F2 Duroc $\times$ Pietrain pig resource population with genetic data available, we compare three different methodologies for genome-wide association studies (GWAS): a single-marker regression, a ridge regression BLUP, and a Bayes Cπ. Initially, two specific chromosomes were chosen to compare these methods in terms of the highest SNP peaks detected. In addition, we also included in Appendix B a more detailed analysis in which the three largest SNPs peaks, and their respective chromosomes, were recorded for all phenotypic traits analyzed.

There are many structure learning algorithms for BN in the area of quantitative genetics (Logsdon and Mezey, 2010; Neto et al., 2008, 2010; Valente et al., 2010; Wang and Van Eeuwijk, 2014), which allow to explore the structure space compatible with the joint distribution function of the variables studied. Among those, Valente et al. (2010) proposed a constraint-based algorithm to search for recursive causal structures among phenotypes conditionally to unobservable polygenic effects, which act as confounders. Basically, a standard Bayesian multiple-trait model is fitted using a Markov chain Monte Carlo implementation to obtain samples from the posterior distribution of a residual covariance matrix, which are then used as input for the inductive causation (IC) algorithm (Verma and Pearl, 1990; Pearl, 2009). However, their method is based on an infinitesimal model and, as such, it does not take into account the possibility of major genes affecting the traits. Often, their method produces a partially directed network that represents a class of possible equivalent solutions, and the use of prior biological knowledge is necessary to determine a final, fully directed graph.

Thus, in Chapter 4, we reach the main goal of this research, which was to develop a hybrid method, called PolyMaGNet (Polygenic traits with Major Genes Network analysis), to search for recursive causal structures among complex phenotypic traits with polygenic inheritance, but allowing also the possibility of major genes affecting the traits. Major genes, previously detected via GWAS, are used as instrumental variables in a final step of the algorithm, providing a fully oriented acyclic graph. This method can be seen as an extension of the one proposed by Valente et al. (2010) and it is especially useful for the analysis of QTL mapping population data involving crosses of outbred populations. We also provide a detailed description of the proposed method and illustrate it with a simulation study, as well as with the real dataset explored in Chapter 3, to recover the causal structure underlying carcass and meat quality traits.

Finally, in Chapter 5 we present a summary of the methods proposed in this dissertation and suggest extensions and refinements for them, which can be used as tools for improvement of economically

important traits as well as to aid the development of breeding programs and optimal decision-making strategies.

References

DE LOS CAMPOS, G. et al. A structural equation model for describing relationships between somatic cell score and milk yield in dairy goats. *Journal of Animal Science*, v. 84, n. 11, p. 2934-2941, 2006.

DE MATURANA, E. L. et al. Exploring biological relationships between calving traits in primiparous cattle with a Bayesian recursive model. *Genetics*, v. 181, n. 1, p. 277-287, 2009.

GELMAN, A. A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-fit Testing. *International Statistical Review*, v. 71, n. 2, p. 369-382, 2003.

GIANOLA, D.; SORENSEN, D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics*, v. 167, n. 3, p. 1407-1424, 2004.

HAAVELMO, T. The statistical implications of a system of simultaneous equations. Econometrica, *Journal of the Econometric Society*, p. 1-12, 1943.

HENDERSON, C. R.; QUAAS, R. L. Multiple trait evaluation using relatives' records. *Journal of Animal Science*, v. 43, n. 6, p. 1188-1197, 1976.

LI, R. et al. Structural model analysis of multiple quantitative traits. *PLoS Genetics*, v. 2, n. 7, p. e114, 2006.

LOGSDON, B. A.; MEZEY, J. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Computational Biology*, v. 6, n. 12, p. e1001014, 2010.

MODE, C. J.; ROBINSON, H. F. Pleiotropism and the genetic variance and covariance. *Biometrics*, v. 15, n. 4, p. 518-537, 1959.

MRODE, R. et al. Random regression model for the genetic evaluation of production traits of dairy cattle in the UK. *Interbull Bulletin*, n. 33, p. 211, 2005.

NETO, E. C. et al. Inferring causal phenotype networks from segregating populations. *Genetics*, v. 179, n. 2, p. 1089-1100, 2008.

NETO, E. C. et al. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics*, v. 4, n. 1, p. 320, 2010.

PEARL, J. Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, v. 95, n. 450, p. 428-431, 2000.

PEARL, J. Causality. Cambridge university press, 2009.

ROSA, G. J. M. et al. Inferring causal phenotype networks using structural equation models. *Genetics Selection Evolution*, v. 43, n. 1, p. 6, 2011.

SCHADT, E. E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, v. 37, n. 7, p. 710-717, 2005.

SHIPLEY, B. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R. Cambridge University Press, 2016.

SORENSEN, D.; GIANOLA, D. Likelihood, Bayesian and MCMC Methods in Genetics. Springer, 2002.

SPIRTES, P.; GLYMOUR, C. N.; SCHEINES, R. Causation, Prediction, and Search. MIT press, 2000.

VALENTE, B. D. et al. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, v. 185, n. 2, p. 633-644, 2010.

VERMA, T. S.; PEARL, J. Equivalence and synthesis of causal models [Technical report R-150]. Department of Computer Science, University of California, Los Angeles, 1990.

WANG, H.; VAN EEUWIJK, F. A. A new method to infer causal phenotype networks using QTL and phenotypic information. *PloS One*, v. 9, n. 8, p. e103997, 2014.

WRIGHT, S. Correlation and causation. *Journal of Agricultural Research*, v. 20, n. 7, p. 557-585, 1921.

2 MODELING CAUSAL PHENOTYPIC NETWORKS OF TWO FRUIT SPECIES OF THE SAPOTACEAE FAMILY: *POUTERIA SAPOTA (JACQ.) H.E. MOORE & STEARN* AND *CHRYSOPHYLLUM CAINITO L.*

Abstract: Understanding the causal architecture underlying complex biological systems, such as in plants and fruits, has a great value in agriculture production for the development of optimal management strategies and selective breeding. So far, most studies use only prior knowledge to propose structural models, fact that may hide important relationships among phenotypes. In this study, we merged structural learning algorithms and prior biological knowledge to investigate the causal networks underlying phenotypic traits of two fruit species belonging to the Sapotaceae family: mamey sapote (*Pouteria sapota (Jacq.) H.E. Moore & Stearn*) and star apple (*Chrysophyllum cainito L.*). We used a constrain- and a score-based algorithm for structure learning of Bayesian networks and selected a putative causal network from the equivalence class of the score-based output using prior knowledge. The decision by the search method was based in the stability evaluation via Jackknife resampling and sample size. The final causal networks for both fruits were fitted using maximum likelihood and evaluated through some fit indices of structural equation models. Statistical tests showed good fit of the models, and common paths in both fruits were similar regarding the intensity of their causal effects. Despite being species from different genus, remarkable similarities can be observed on the inferred causal structures. It is therefore likely that the presented findings here are transferable to other species of the Sapotaceae family, but further investigations relating other fruits are required. In addition, the methods used in this study have the potential to unravel causal phenotypic networks in complex biological systems.
Keywords: Bayesian networks; Structural equation models; Causal networks; Mamey sapote; Star apple; Structure learning.

2.1 Introduction

The Sapotaceae family belongs to the extended order Ericales, a clade of morphologically variable angiosperm families, wherein relationships among species are not fully resolved (Anderberg and Swenson, 2003). This family is subdivided into five tribes with 53 genera and approximately 1,250 species, mostly originating from tropical and subtropical regions of Asia and South America (Pennington, 1991; Govaerts et al., 2001; Swenson and Anderberg, 2005). Two of these species were used in this study, the *Pouteria sapota (Jacq.) H.E. Moore and Stearn* and *Chrysophyllum cainito L.*

Pouteria sapota (Jacq.) H.E. Moore and Stearn is a tropical fruit tree native to Mexico and Central America (Popenoe, 1920; Martínez and Martínez, 1959), nowadays cultivated in several regions of the world, due to its high economic value (Arias et al., 2015). It is commonly called mamey sapote, although is also referred to with other names such as zapote (Gazel Filho et al., 1999) and zapote mamey (Gómez-Jaimes et al., 2012). The fruit can be consumed directly or its pulp can be used to make candy, ice cream and milkshakes. Its seed oil is explored by cosmetic industries for producing shampoos, hair dyes, medicines, and other products. The wood of the tree is used in manufacturing fine furniture and its latex to treat skin infections, while leaves provide fungicides and insecticides (Alia-Tejacal et al., 2007; Pinto et al., 2016).

Chrysophyllum cainito L. is typically considered to be originally from Central America, although many botanists suggest that it is actually native to the West Indies (Morton, 1987). The fruit is mainly known as caimito, a local name (Gazel Filho, 1995; Parker et al., 2010), or star apple (Luo et al., 2002; Pino et al., 2002). It is not explored for multiple purposes as the mamey sapote, but it is used in the development of conventional and ayurvedic medicines (Li et al., 2015).

Although these fruit trees are classified into different genera, they are similar in many physiological aspects due to their common genetic origin. Here, we investigate the underlying phenotypic architecture in their fruits, i.e., we explore the relationships among quantitative phenotypic traits using algorithms to search for causal structures of Bayesian networks (BN) (Pearl, 2000; Spirtes, Glymour and Scheines, 2000). Afterwards, we verify if a structural equation model (SEM) (Grace and Bollen, 2005) with a causal structure based on the search algorithm's output and on biological prior knowledge fits well the data evidence. In addition, in the first part, we also compare results from a constraint- and a score-based algorithm, assessing their stability via Jackknife resampling; and, in the second part, we estimate and test the magnitude of path coefficients by maximum likelihood and Wald test, respectively.

2.2 Methods

2.2.1 Data

The *Pouteria sapota (Jacq.) H. Moore and Stern* and *Chrysophyllum cainito L.* datasets considered in this study were provided by Gazel Filho (Gazel Filho, 1995). Data were collected from a Sapotaceae family tree collection located in Cabiria 6 Botanical Garden of Centro Agronómico Tropical de Investigación y Enseñanza (CATIE), Costa Rica. More specifically, the geographical location is on north 9°53' and east 83°39', 602 meters above sea level. The average annual temperature is 22.3°C. Physical and chemical characteristics of the soil are presented in Table 2.1.

Table 2.1. Physical and chemical characteristics of the Cabiria 6 (CATIE) soil.

Depth (cm)	**0-20**	**20-40**
Sand[1] (%)	39.1	45.0
Silt[1] (%)	41.3	32.5
Clay[1] (%)	19.6	22.5
Stoniness[1] (%)	1-2	1-2
pH[2]	5.6	5.7
Ca[2] (meq/100ml)	2.53	1.64
Mg[2] (meq/100ml)	0.81	0.35
K[2] (meq/100ml)	0.36	0.25
P[2] (mg/l)	4.4	7.0
Cu[2] (mg/l)	9.0	5.7
Zn[2] (mg/l)	2.0	1.7
Mn[2] (mg/l)	24.0	8.4

[1]physical characteristics [2]chemical characteristics [3]Table adapted from Gazel Filho (1995)

Seeds collected from Mexico to Panama were introduced into the CATIE genbank between 1977 and 1983. They were planted at a distance of 8 x 6.5m from each other. The trees with abundant fruits or flowers were tagged to the study between November 1994 and January 1995. After harvesting the fruits at the usual time, they were wrapped in paper and stored until they reached the appropriate point of maturity to be evaluated. The fruits and their peel were weighed on an electronic scale. Fruits lengths and diameters were also recorded. Subsequently, the fruits were cut, so that their pulp and seeds could be extracted. Data were collected from fruits of 112 trees (63 of mamey sapote and 49 of star apple) for the phenotypic traits listed in Table 2.2 – for more details, see (Gazel Filho, 1995).

2.2.2 Causal structure learning

The BN can be seen as a factorization of the joint (global) distribution $P(X_1, ..., X_p)$ of the variables under study consisting on the product of all (local) distributions of the variables X_i, $i \in \{1, ..., p\}$ conditionally to their parents according to a directed acyclic graph (DAG). The possibility

Table 2.2. Phenotypic traits

Variables	Description (unit)
FRW	Fruit weight (g)
FRL	Fruit length (mm)
FRD	Fuit diameter (mm)
PET	Peel thickness (mm)
PEW	Peel weight (g)
SEL	Seed length (mm)
SED	Seed diameter (mm)
SEW	Seed weight (g)

of such factorization involves assuming the Markov property. A sequence of variables $X_1, ..., X_n, ...$ is a Markov chain if $X_{(n+1)}$ is conditionally independent of $X_1, ..., X_{(n-1)}$ given X_n and can be written mathematically as $X_{(n+1)} \perp\!\!\!\perp (X_1, ..., X_{(n-1)}) | X_n$ or $X_{(n+1)} \perp\!\!\!\perp_p (X_1, ..., X_{(n-1)}) | X_n$ to emphasize that the statement is relative to a given probability distribution P. If parents of the variable X_i are denoted by $pa(X_i)$, the BN can be represented by

$$P(X_1, ..., X_p) = \prod_{i=1}^{p} P(X_i | pa(X_i)) \tag{2.1}$$

The DAG G can be represented as $G = (V, E)$ with $V = (X_1, X_2, ..., X_p)$ as the set of all nodes (variables) and E the set of arrows connecting all X_i to each element of $pa(X_i)$. Arrows in E are directed from parents to children. Given any two nodes X_i and $X_j \subset V$ and a set of nodes $Z \subset V$, such X_i and $X_j \not\subset Z$, the conditional independence between X_i and X_j given Z can be tested applying the d-separation criterion. The set Z is said to d-separate X_i from X_j if and only if Z blocks every path from X_i to X_j (Pearl, 2000).

Several algorithms have been proposed in the literature for BN structure learning (Nagarajan, Scutari and Lèbre, 2013). They are typically classified into three broad categories: constraint-based, score-based, and hybrid algorithms. Constraint-based algorithms are based on the concepts about causal graphical models introduced by Judea Pearl (Pearl, 2000). His inductive causation (IC) algorithm (Verma and Pearl, 1990) provides a framework for learning structures using conditional independence tests. Score-based algorithms (commonly known as search-and-score algorithms) involve the application of optimization techniques. Each candidate network is assigned a network score reflecting its goodness of fit and the algorithm then attempts to find its maximum in a network space. Finally, the hybrid structure learning algorithms combine features of both constraint- and score-based methods, mostly by finding edges with a constraint-based approach and orienting them afterwards by a score-based approach. Here, we search for causal structures among the studied traits with two different types of algorithms: the incremental association Markov blanket (IAMB) (Tsamardinos et al., 2003) and the Tabu search (Bouckaert, 2001), which are constraint- and score-based methods, respectively.

The stability of the selected structures was subsequently evaluated via Jackknife resampling (Peñagaricano et al., 2015). For large samples, a bootstrap approach can yield more insight into the stability of a network structure (Topner et al., 2017), but the Jackknife method is more suitable to smaller samples as the one studied here. Considering that N is the number of observations in the dataset, the Jackknife resampling consists on applying the search method N times and leaving one observation out each time (i.e., on subsamples of size $N - 1$). This allows one to investigate the influence of each experimental unit on the network learning.

2.2.3 Incremental Association Markov Blanket

IAMB consists of two steps referred to as forward and backward phases (Tsamardinos et al., 2003). The forward phase starts by searching the Markov blanket (MB) of each variable X_i (i.e., $MB(X_i)$). This concept corresponds to the minimal set of variables conditioned on which all other variables are independent of X_i. The backward phase involves identifying and removing false positives. Exploring MBs reduces the number of independence tests in the search. As a result, the algorithm provides computationally and time-efficient search without compromising the accuracy.

The IAMB was not the first algorithm to explore the concept of MB, which was employed earlier by Koller and Sahami (KS) (Koller and Sahami, 1996). Later, the grow-shrink (GS) algorithm (Margaritis and Thrun, 2000) was proposed, already involving a forward and a backward phase as described here. Based on GS, Tsamardinos et al. (2003) developed IAMB along with some variations. The main advantage of IAMB compared to GS and KS is the use of an efficient heuristic function in the first phase.

All these algorithms perform series of conditional independence tests of some type. In this study, we used the exact Student's t test with a type I error rate of $\alpha = 0.05$ and $\alpha = 0.20$, which is suitable for testing associations among normally distributed variables. The choice for this test was based for its robustness in relation to the normality assumption, since we have small sample sizes.

2.2.4 Tabu search

The Tabu search is implemented by an optimization algorithm (Bouckaert, 2001) that starts with an arbitrary point in the DAG space and recursively selects a new solution in the neighborhood of the previous one, increasing the score for a pre-defined function in each step. For example, given a structure G_1 and the dataset D, the score is typically described by the probability of the structure given the data, that is

$$Score(G_1|D) = P(G_1|D) = \frac{P(D|G_1)P(G_1)}{P(D)}. \tag{2.2}$$

The algorithm stops after a pre-defined number of steps or when a stop criterion is satisfied – for more details, see (Bouckaert, 2001).

Several metrics have been proposed as score functions, for example, (Morota et al., 2012) used the Bayesian Dirichlet equivalent (BDe) (Heckerman, 1995), a suitable choice for a data set including discrete variables. In this study, the Tabu search was combined with a score equivalent Gaussian posterior density (BGe) (Geiger and Heckerman, 1994) that is a posterior probability of the graph's structure given the data for settings consisting of Gaussian variables.

2.2.5 Equivalent structures

In most cases, constraint-based algorithms return a partially directed acyclic graph (PDAG), that only assign directions to edges whose d-separations are supported. PDAGs represent classes of statistically equivalent BN structures (same joint probability distributions), with no cycle, containing directed edges only for nodes participating in a v-structure (also called unshielded collider: two converging arrows whose tails are not connected, e.g., an ordered triple of nodes $X_1 \rightarrow X_2 \leftarrow X_3$ such X_1 and X_3 are not adjacent).

Noteworthy, if the edges are directed in a PDAG, all the possible solutions in the equivalence class agree with their orientation. Otherwise, for each undirected edge, there are at least two DAGs in the equivalence class whose corresponding edges point into opposite directions. Although score-based algorithms do not result in PDAGs, is straightforward to detect the equivalent class responsible for the

DAG provided. DAGs are Markov equivalent if and only if they have the same skeletons and the same set of v-structures (Verma and Pearl, 1990).

According to Cheng, Bell and Liu (1997), in small sample size and noisy data, score-based algorithms are more accurate since they search the whole model space to find the optimal model. Therefore, we recovered the equivalence class of the output provided by the score-based algorithm and used prior knowledge to select the most likely architecture for the biological system within this class. Afterwards, the resulting structures for both fruits were evaluated using SEM.

2.2.6 Structural Equation Models

A SEM can be seen as a multiple equation system where response variables of one equation can be included as covariates in another one. The system may be used to express how each variable in the left hand side is causally affected by their causal parents in the right hand side. The construction of SEM can be guided by DAGs, as they can express how variables are causally related. These models can be used to study recursive and simultaneous relationship among phenotypes in multivariate systems.

A general SEM, with $p + q$ observed variables, such that q are exogenous (whose values are completely influenced by factors ignored by the model) and p are endogenous (values influenced by exogenous and other endogenous variables), can be expressed mathematically as

$$\mathbf{y} = \Lambda\mathbf{y} + \Gamma\mathbf{x} + \mathbf{e}, \tag{2.3}$$

where $\mathbf{y}$ is a (px1) vector of endogenous observed variables; Λ is a square (pxp) matrix with zeros in the diagonal and structural coefficients in the off-diagonals defining the relations among endogenous variables; Γ is a (pxq) matrix defining the relations from exogenous to endogenous variables; $\mathbf{x}$ is a (qx1) vector of exogenous observed variables; and, $\mathbf{e}$ is a (px1) vector of residuals. For more details on the model and its applications in the context of quantitative genetic and biology, see (Rosa et al., 2011; Shipley, 2002).

Fitting a SEM involves minimizing the difference between observed and predicted patterns of covariation among variables (Shipley, 2002). In other words, conditionally on a given causal structure among variables, the values for the free parameters should be chosen in a way that makes the predicted and observed covariance matrices as similar as possible. This is usually performed by using maximum likelihood (ML) estimation, which is equivalent to minimizing the following criterion (Grace and Bollen, 2005):

$$F_{ML} = ln|(\mathbf{\Sigma}(\mathbf{\Theta}))| + trace[\mathbf{S}\mathbf{\Sigma}^{-1}(\mathbf{\Theta})] - ln|\mathbf{S}| - (p+q), \tag{2.4}$$

where $\mathbf{\Sigma}(\mathbf{\Theta})$ and $\mathbf{S}$ are the predicted and observed covariance matrices, respectively, involving p endogenous and q exogenous variables.

The ML estimation method assumes a multivariate normality distribution for the variables. Considering a sample of N observations, the asymptotic distribution of $(N-1)F_{ML}$ is a χ^2 with $s - t$ degrees of freedom, where s is the number of non-redundant elements in the symmetrical matrix $\mathbf{S}$ and t is the number of parameters to be estimated, under the assumption that the model is correct (Shipley, 2002). Therefore, this statistic is used to test if the model fits the data; i.e., the null hypothesis $H_0 : \mathbf{S} - \mathbf{\Sigma}(\mathbf{\Theta}) = 0$. Notice that the interpretation of the test result is the opposite of that normally used in typical statistical tests as a consequence of the different meaning of the null hypothesis, i.e., here it would be interesting not to reject the null hypothesis, since it would guarantee the proximity between observed and predicted covariance matrices.

The χ^2 statistical significance test is sensitive to sample size (Hooper, Coughlan and Mullen, 2003), such that the model tends to be rejected more when samples are large (Bentler and Bonett, 1980).

On the other hand, when sample sizes are small, it is susceptible to type II error (incorrectly keeping a false null hypothesis), leading to difficulty to distinguish good and poor models (Kenny and McCoach, 2003). To avoid this dilemma, sample size should be evaluated relative to the number of free parameters to be estimated in the model (Shipley, 2002). One rule of thumb is that there should be at least five times more experimental units than free parameters (Bentler, 1990).

Evaluating the fit of a SEM it is not a simple task, because there is no benchmark statistical test that identifies whether a model is correctly adjusted. Therefore, it is necessary to consider several criteria simultaneously to evaluate the quality of adjustment. According to Schermelleh-Engel, Moosbrugger and Müller (2003), model evaluations can be assessed inferentially by the χ^2 test or descriptively by other fit indices as goodness-of-fit (GFI), adjusted goodness-of-fit (AGFI), root mean-squared error approximation (RMSEA) and many others.

The RMSEA evaluates the fit of a studied non-saturated model relative to a saturated model, i.e., a model that is complex enough to be compatible to any covariance pattern among variables. Small values for this criterion indicate that the tested model fits nearly as well as a saturated model (Fox, 2002). RMSEA values ≤ 0.05 can be considered as a good fit, values between 0.05 and 0.08 are deemed as an adequate fit, and values between 0.08 and 0.10 correspond to a mediocre fit. Values superior to 0.10 are not acceptable (Browne and Cudeck, 1992). The GFI are an alternative to the χ^2 test, calculating the proportion of variance that is accounted by the model (Hooper, Coughlan and Mullen, 2003). The AGFI adjusts the GFI based upon degrees of freedom. Several cutoffs for the GFI and AGFI have been proposed, but consensus indicates it should be close to 1 (Fox, 2002).

In this study, path coefficients were estimated by ML and their magnitudes tested by Wald test, which is based on the ratio of each regression coefficient estimate to its standard error, that is distributed as a z statistic. This test potentially locates the path coefficients that can be considered zero without impairing the fit of the model (Bentler, 1990). As recommended by Iriondo, Albert and Escudero (2003), non-significant parameters can be eliminated from the model in order to improve it, especially if their theoretical interpretation is weak. On the other hand, when the model is considered to be poor by the criteria of goodness-of-fit, it is possible to use the modification indices proposed by Sorbom (1989) to include as much as possible of what is known about the dataset. This method is a chi-square statistic, each on one degree of freedom (Lagrange multiplier), and can be regarded as an estimate of the improvement in the likelihood-ratio chi-square statistic for the model if one respective parameter is considered in the model as a free parameter.

All analyzes were performed using the R statistical software (The R Core Team, 2013), including the packages “bnlearn” (Scutari, 2009) and “sem” (Fox, Nie and Byrnes, 2012).

2.3 Results

2.3.1 Exploratory data analysis

A descriptive exploratory analysis was conducted (Table 2.3). Coefficients of variation (CV) of the phenotypic traits of star apple [mamey sapote] varied from 8.21% [9.28%] for seed diameter (SED) up to 43.85% [33.68%] for fruit weight (FRW). The phenotypic difference between these two species is evident by the divergent averages. The mamey sapote is longer with ellipsoidal shape while star apple is spherical, what reflects on fruit length (FRL) and fruit diameter (FRD). Seed length and diameter follow the same trend as fruit shape in mamey sapote, and seed weight is ten times higher for mamey sapote than for star apple. However, CVs of traits are relatively similar for both fruit types.

An important assumption associated with many structure learning algorithms and the use of maximum likelihood estimation methods in SEM is the presence of multivariate normality among observed variables. Royston’s multivariate normality test (Royston, 1983) indicated multivariate normality for the

Table 2.3. Mean followed by standard error (in parentheses), min – max and coefficient of variation (%, in square brackets) for all phenotypic traits.

Variable	Sapotaceae family	
	Mamey sapote	**Star apple**
FRW (g)	380.81 (16.16) 173.6 – 693.6 [33.68%]	100.10 (6.27) 42.80 – 244.7 [43.85%]
FRL (mm)	98.38 (2.02) 64.6 – 135.7 [16.30%]	56.06 (1.16) 40.41 – 74.3 [14.54%]
FRD (mm)	83.36 (1.28) 62.8 – 106.2 [12.19%]	54.86 (1.14) 42.51 – 78.8 [14.55%]
PET (mm)	1.93 (0.05) 1.0 – 2.9 [21.27%]	2.71 (0.10) 1.23 – 4.6 [25.42%]
PEW (g)	52.38 (2.11) 24.4 – 96.2 [31.96%]	33.58 (1.99) 14.90 – 81.0 [41.49%]
SEL (mm)	62.22 (0.90) 45.1 – 76.6 [11.55%]	18.48 (0.29) 12.92 – 22.7 [10.85%]
SED (mm)	33.17 (0.39) 23.0 – 41.5 [9.28%]	12.24 (0.14) 10.28 – 16.3 [8.21%]
SEW (g)	40.28 (1.17) 18.4 – 59.9 [23.03%]	4.85 (0.20) 2.01 – 8.9 [28.75%]

[1]FRW: Fruit weight; FRL: Fruit length; FRD: Fruit diameter; PET: Peel thickness; PEW: Peel weight; SEL: Seed length; SED: Seed diameter; SEW: Seed weight.

mamey sapote ($H = 11.65; p = 0,096$) and star apple ($H = 9.28; p = 0,079$) dataset. Before this test, as recommended by Box-Cox, a logarithmic transformation has been applied to the phenotypic values of star apple, besides SEL and FRL.

2.3.2 Structure learning of Bayesian networks

Structures resulting from learning algorithms are shown in Figure 2.1 for both fruits. Each row represents a different method: (I) IAMB with $\alpha = 0.05$; (II) IAMB with $\alpha = 0.20$; and, (III) Tabu search. The stability of the structures was evaluated via Jackknife resampling, computing the frequency of each edge and respective orientation during the process. These values are shown beside the edges (percentage of occurrence of the connection/ percentage with same orientation). Black edges depict the original output provided by the algorithms, while gray edges represent pathways that only appeared during Jackknife process. Structures recovered by Tabu search were 100% stable.

Although the original connections have remained stable ($> 80\%$) in the structures from the IAMB algorithm (except the paths $FRL \rightarrow PEW$ [65.3%; $\alpha = 0.05$] and $FRW \rightarrow PEW$ [53.1%; $\alpha = 0.20$] in star apple), there was no evident agreement regarding the orientation in most cases, especially when the type I error rate was fixed to 0.20. In addition, the Jackknife process revealed some extra edges in the IAMB output, such that all of them were detected using Tabu search, except the pathway FRW–FRL in mamey sapote.

Bold edges (Figure 2.1) highlight fixed edges in the equivalence class of the results provided by Tabu search, i.e., of all possible solutions with the same joint probability distribution, these edges always have fixed orientation. In mamey sapote, two v-structures were recovered by Tabu search: $FRW \rightarrow FRD \leftarrow FRL$ and $SEL \rightarrow SEW \leftarrow SED$. The latter one was only detected in the IAMB with $\alpha = 0.20$. The structure with a shielded collider, $FRD \rightarrow SEL \leftarrow FRL$, is also a fixed path since $SEL \rightarrow FRD$ is not possible because it would create a new v-structure ($FRW \rightarrow FRD \leftarrow SEL$). Consequently, we set $FRL \rightarrow SEL$ to avoid a cycle. In star apple, the v-structure $PET \rightarrow PEW \leftarrow FRL$, recovered by IAMB, would not be recovered if the path PET–FRL (detected by Jackknife) was considered. Indeed, PET–FRL appears in the Tabu search output, which only showed the v-structure

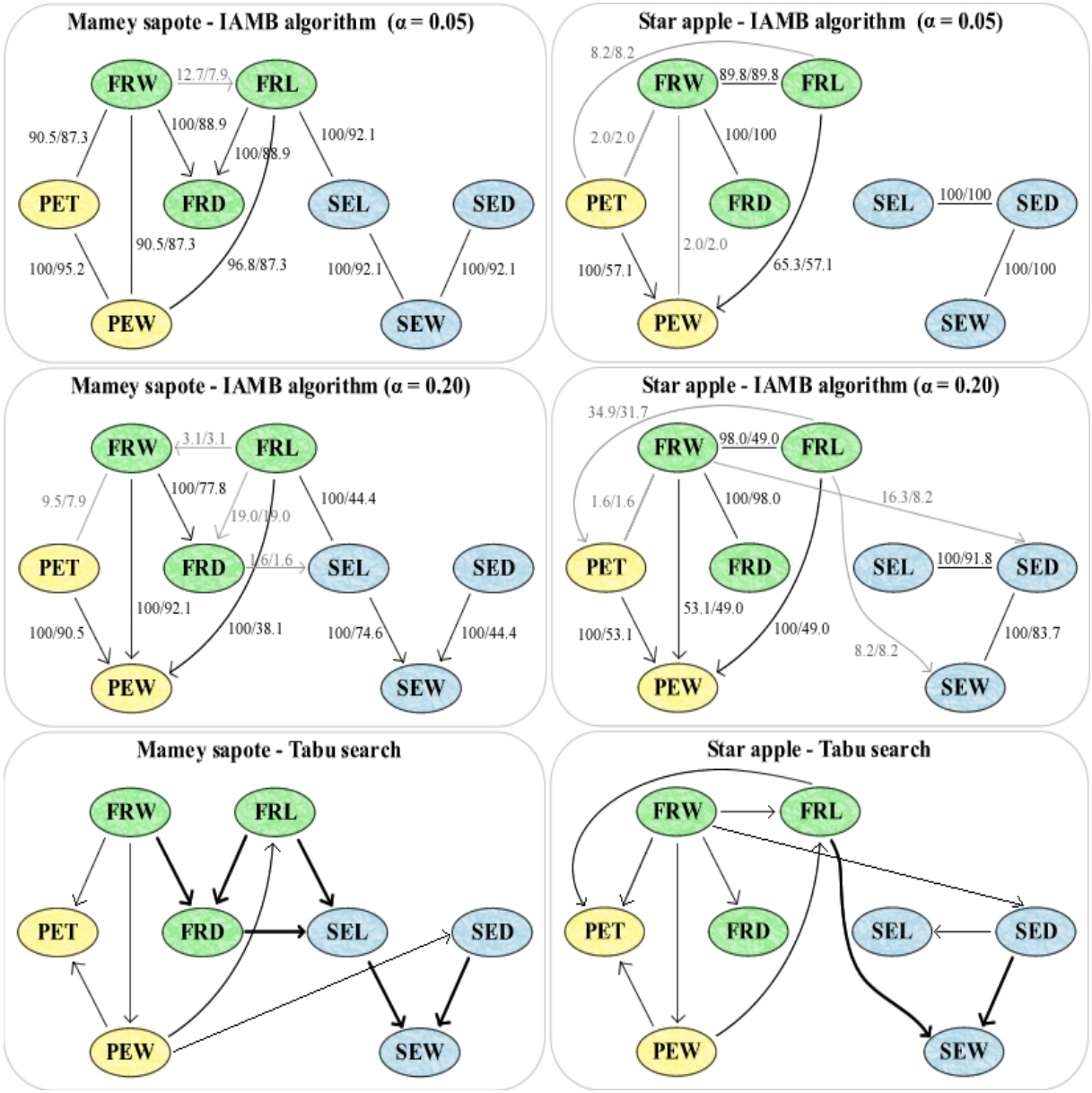

Figure 2.1. Structure learning by IAMB algorithm ($\alpha = 0.05$ and $\alpha = 0.20$) and Tabu search for both fruits. Numerical values represent the results of Jackknife resampling (percentage of occurrence the connection/ percentage with same orientation). Black edges depict the original output provided by the algorithms, while gray edges represent pathways detected during Jackknife process. Bold edges highlights fixed edges in the equivalence class of the structures provided by Tabu search.

$FRL \rightarrow SEW \leftarrow SED$, which was not detected by the IAMB algorithm.

Comparing the results of the two structure learning algorithms and taking into account the pathways detected via Jackknife resampling, one can see similarities regarding the connections recovered. All edges detected by IAMB with $\alpha = 0.20$ were also detected by Tabu search (except the path FRW–FRL), though not always with the same directions. Both methods showed clusters of traits responsible for similar roles in the fruits analyzed: seed traits (blue), peel traits (yellow) and fruit traits (green). However, the pattern of connection among traits differ with respect to the fruit.

Regarding the phenotypic architecture of both fruits, one can observe remarkable similarities. Table 2.4 presents counts of undirected and directed edges for the learning algorithms, and also the number of common connections and directed edges for both fruits. From six connections recovered for star apple

using IAMB with $\alpha = 0.05$, four were also recovered for mamey sapote: FRW–FRD, PET–PEW, FRL–PEW and SED–SEW. One more common connection (FRW–PEW) was detected using IAMB with $\alpha = 0.20$. Tabu search recovered 11 directed edges for both fruits, six of them were in common: $FRW \rightarrow PET$, $FRW \rightarrow PEW$, $FRW \rightarrow FRD$, $PEW \rightarrow FRL$, $PEW \rightarrow PET$ and $SED \rightarrow SEW$.

Table 2.4. Number of undirected and directed edges for mamey sapote and star apple along with the number of connections and directions in common.

Algorithms	**Mamey sapote**	**Star apple**	**Common edges**
	U (D)	**U (D)**	**C (C+D)**
IAMB ($\alpha = 0.05$)	7 (2)	4 (2)	4 (0)
IAMB ($\alpha = 0.20$)	1 (6)	4 (3)	5 (3)
Tabu search	0 (11)	0 (11)	6 (6)

[1]Note: U – undirected edges; D – directed edges; C – connections.

2.3.3 Networks selected from structure learning and prior knowledge

Following the structure learning study, we used the networks provided by Tabu search as a starting point for fitting a SEM. This decision was taken after evaluate the stability of structures and, principally, due to the sample size of our data set, as commented in Material and Methods Section. In addition, the constraint-based methods did not assign a direction to every edge due to their reliability on the d-separation criterion. For example, the seed triplet (SEL–SED–SEW) in the star apple structure (blue in Figure 2.1), recovered by IAMB, could be directed in three different ways, which correspond to the same joint distribution:

(i) $SEL \rightarrow SED \rightarrow SEW$:

$$P(SEL, SED, SEW) = P(SEW \mid SED)P(SED \mid SEL)P(SEL)$$

(ii) $SEL \leftarrow SED \leftarrow SEW$:

$$\begin{gathered} P(SEL, SED, SEW) = P(SEL \mid SED)P(SED \mid SEW)P(SEW) = \\ \frac{P(SEL,SED)}{P(SED)} \frac{P(SED,SEW)}{P(SEW)} P(SEW) = \\ [P(SED \mid SEL)\frac{P(SEL)}{P(SED)}][P(SEW \mid SED)\frac{P(SED)}{P(SEW)}]P(SEW) = \\ P(SEW \mid SED)P(SED \mid SEL)P(SEL) \end{gathered}$$

(iii) $SEL \leftarrow SED \rightarrow SEW$:

$$\begin{gathered} P(SEL, SED, SEW) = P(SEL \mid SED)P(SEW \mid SED)P(SED) = \\ \frac{P(SEL,SED)}{P(SED)} P(SEW \mid SED)P(SED) = \\ [P(SED \mid SEL)\frac{P(SEL)}{P(SED)}]P(SEW \mid SED)P(SED) = \\ P(SEW \mid SED)P(SED \mid SEL)P(SEL) \end{gathered}$$

In contrast, using Tabu search, $SED \rightarrow SEW$ was part of a v-structure ($FRL \rightarrow SEW \leftarrow SED$) not revealed by IAMB.

In this way, we recovered the equivalence class of the resulting DAGs provided by Tabu search and used prior knowledge to select the most likely architecture for the biological system within this class. Bold edges in Figure 2.1 represent fixed edges belonging to the equivalence classes of the resulted structures, i.e., all the possible solutions in the equivalence class are in accordance with their orientation. The final graphs are depicted in Figure 2.2.

2.3.4 Structural equation models

Both causal models depicted in Figure 2.2 consist of systems composed of eleven equations, each representing an autonomous mechanism governing one variable. The error variables are not shown explicitly in the graph, and by convention they are assumed to be mutually independent (Pearl, 2000). Standardized path coefficients were estimated through maximum likelihood and are shown in Figure 2.2. In star apple structure, the path coefficients of $FRL \rightarrow FRW (\lambda_{(FRW,FRL)} = 0.15; p = 0.109)$ and $FRL \rightarrow SEW (\lambda_{(SEW,FRL)} = 0.26; p = 0.056)$ were not significant by Wald test. However, we have kept these paths in the model due to their biological meaning.

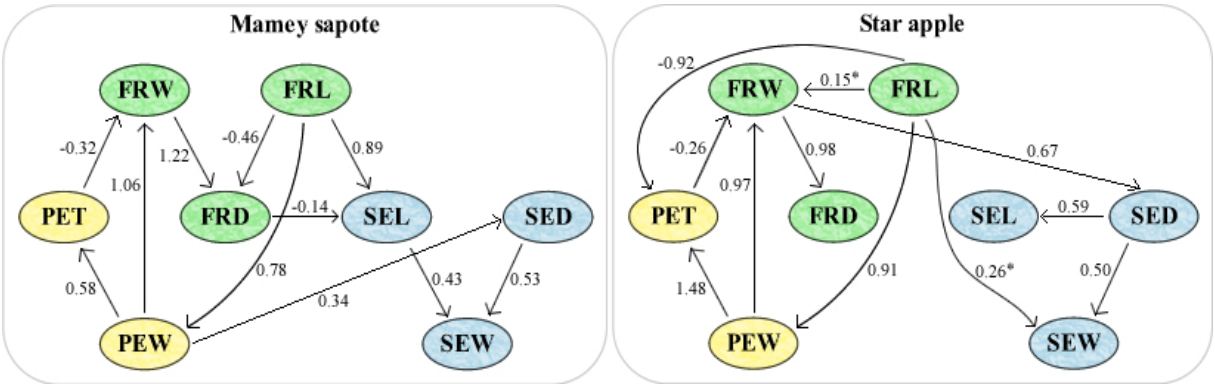

Figure 2.2. Structures resulting from the use of prior knowledge in choosing the most likely architecture for the biological system in the equivalence class provided by Tabu search. Numerical values represent standardized path coefficients estimated by the maximum likelihood method. All of them were significant (p<0.05), except those with an asterisk after the path coefficient.

Models' χ^2 tests indicated that the null hypothesis should not be rejected (Table 2.5), i.e., observed and predicted covariance matrices are statistically equivalent for both models. Goodness-of-fit evaluations (Table 2.5), using the thresholds proposed in the literature (Hooper, Coughlan and Mullen, 2003; Fox, 2002; Browne and Cudeck, 1992; Miles and Shevlin, 1998), indicated that the models fit the data well.

Table 2.5. Models' chi-square statistic and goodness-of-fit criteria

	Mamey sapote	**Star apple**
χ^2	$\chi^2_{17} = 12.81 (p = 0.749)$	$\chi^2_{17} = 13.29 (p = 0.716)$
GFI	0.95	0.94
AGFI	0.89	0.88
RMSEA	<0.001	<0.001

Although all path coefficients in mamey sapote have been significant by Wald test $p < 0.05$, the paths $FRD \rightarrow SEL$ and $PEW \rightarrow SED$, not detected with IAMB with $\alpha = 0.05$, were not significant considering $p < 0.01$. When removing these paths of the model, the goodness-of-fit measures ($\chi^2_{19} = 23.63 (p = 0.749); GFI = 0.92; AGFI = 0.85; RMSEA = 0.063$) indicated a reasonable fit to the data. In star apple, removing the paths FRL→FRW and FRL→SEW, which were not significant by Wald test $p < 0.05$, the goodness-of-fit measures ($\chi^2_{19} = 19.71 (p = 0.413); GFI = 0.91; AGFI = 0.84; RMSEA = 0.028$) indicated that model also fits well the data. Here, we decided to keep these paths in the models, however, further studies are required since they can be false positives.

Path coefficients (Figure 2.2) were standardized to account for the different units used for the different traits. In this way, the impact of one standard deviation difference in one variable can be compared to a standard deviation difference in another (Grace and Bollen, 2005). Two path coefficients

in mamey sapote ($PEW \rightarrow FRW$ and $FRW \rightarrow FRD$) and one in star apple ($PEW \rightarrow PET$) were estimated to be larger than one, such result is possible when there is multicollinearity in the data. The common misconception that such result is not possible probably stems from classical exploratory factor analysis where factors are standardized and orthogonal (Jöreskog, 1999). Here, however, the variables are correlated and the loadings among them are regression coefficients, which can be greater than one.

There are six common directed edges in both structures ($FRL \rightarrow PEW$, $PEW \rightarrow PET$, $PET \rightarrow FRW$, $PEW \rightarrow FRW$, $FRW \rightarrow FRD$ and $SED \rightarrow SEW$) and FRL affects directly or indirectly all the traits. In mamey sapote, FRL has a negative direct effect on $FRD(-0.46)$ and an indirect by the pathways $FRL \rightarrow PEW \rightarrow PET \rightarrow FRW \rightarrow FRD(-0.18)$ and $FRL \rightarrow PEW \rightarrow FRW \rightarrow FRD(1.01)$, which together have a total effect of 0.37 on FRD. In contrast, in star apple, FRL only affects FRD indirectly by the pathways $FRL \rightarrow FRW \rightarrow FRD(0.15)$, $FRL \rightarrow PEW \rightarrow PET \rightarrow FRW \rightarrow FRD(-0.34)$, $FRL \rightarrow PEW \rightarrow FRW \rightarrow FRD(0.86)$ and $FRL \rightarrow PET \rightarrow FRW \rightarrow FRD(0.23)$, which together have a total effect of 0.90 on FRD. Thus, in mamey sapote, a change of one standard deviation in FRL causes a change of 0.37 in FRD, revealing its ellipsoidal shape. In star apple, however, a change of one standard deviation in FRL causes a change of 0.90 standard deviation in FRD, displaying its spherical shape.

Regarding the effect of FRL on SEL in mamey sapote, FRL directly affects $SEL(0.89)$ and indirectly by the pathways $FRL \rightarrow FRD \rightarrow SEL(0.06)$, $FRL \rightarrow PEW \rightarrow PET \rightarrow FRW \rightarrow FRD \rightarrow SEL(0.02)$ and $FRL \rightarrow PEW \rightarrow FRW \rightarrow FRD \rightarrow SEL(-0.14)$, which together have a total effect of 0.83 on SEL. In star apple, FRL only affects SEL indirectly by the pathways $FRL \rightarrow PEW \rightarrow PET \rightarrow FRW \rightarrow SED \rightarrow SEL(-0.14)$, $FRL \rightarrow PEW \rightarrow FRW \rightarrow SED \rightarrow SEL(0.35)$, $FRL \rightarrow PET \rightarrow FRW \rightarrow SED \rightarrow SEL(0.09)$ and $FRL \rightarrow FRW \rightarrow SED \rightarrow SEL(0.06)$, which together have a total effect of 0.36 on SEL. Thus, a change of one standard deviation in FRL causes a change of 0.83 on SEL in mamey sapote, revealing that seeds length is influenced by the fruit length. In star apple, a change of one standard deviation in FRL causes a change of 0.36 on SEL, showing that FRL has not such a large effect on SEL.

The causal pathways bring us a lot of information about these fruits, however, their absence is even more important. For example, FRD does not affect any trait in star apple structure. So any intervention on FRD would not affect the seed traits for this fruit, i.e., industries interested in seeds as raw material cannot expect a positive effect for seed yield from changing FRD. On the other hand, a change will probably also have no negative effect so one does not need to take care for keeping FRD constant or improving it.

2.4 Discussion

It is known that the number of potential edges in a causal graph grows exponentially with increasing the number of variables in study. Prior knowledge helps to accelerate the enormous search task of the learning algorithms, since it puts some restrictions in the search space. Purely data-driven analyses allow to explore and compare unintuitive structures additionally and rank them according to goodness-of-fit indexes. Relying on prior knowledge alone to assess phenotypic structures, as in many studies (Iriondo, Albert and Escudero, 2003; Albert, Escudero and Iriondo, 2001; Del Cacho, Peñuelas and Lloret, 2013), might lead to missing interesting connections and to excessive distances between the observed and predicted covariance matrices. Therefore, combining structure learning algorithms with prior knowledge is important when recovering the underlying network of complex systems.

It is unfeasible to indicate which structure learning algorithm provided an output closer to the true biological network, since it is unknown. Each study requires a reflection on which algorithm best supports the data, e.g. based on goodness-of-fit or predictive ability. Constrain-based methods are

more conservatives than score-based ones because they only identify the direction of an edge if there is good support from the data. They can therefore be more efficient in studies with large sample sizes (Peñagaricano et al., 2015; Morota et al., 2012; Valente et al., 2010). However, the detection of conditional independencies is susceptible to failure in hypothesis tests, and they also may not assign a direction to every edge due to their reliability on the d-separation criterion, which is unsatisfying and problematic when it comes to translating the structure into a SEM. Thus, score-based approaches are generally preferred in studies with small sample size and noisy data (Cheng, Bell and Liu, 1997; Su et al., 2013).

The downside of score-based algorithms is on the other hand that their directing of edges is not as reliable as it is in constraint-based approaches. One therefore need to interpret the connections carefully with regard to their causal nature. It is helpful here to recover the equivalence class of the resulting DAG and use prior knowledge to select the most likely architecture for the biological system within this class, as shown in this study.

It should be stressed the potential presence of genetic confounders that could change the resulting Bayesian network configuration. There are many structure learning algorithms that take into account genetic data. Valente et al. (2010) proposed a constraint-based methodology that considers the total polygenic effect acting over the traits. They recovered the causal structure conditioned on the genetic effects. Neto et al. (2010) proposed a hybrid algorithm that considers QTLs information as instrumental variables to orient the pathways of the phenotypic network. Peñagaricano et al. (2015) described a methodology for assessing causal networks involving latent variables and genetic confounders. Therefore, analyzing the phenotypic networks provided by this study considering genetic effects or other confounders warrants further investigation.

Here, the use of SEM enabled the comparison of competing models and provided a causal interpretation of the biological system studied. In contrast to other approaches, such as multiple-trait analysis, SEM do not only allow significance tests of individual path coefficients, but also provide some goodness-of-fit of indices of the whole models. The models studied here could theoretically be analyzed with fewer assumptions, which would make the model more complex (e.g. the assumption of independent error terms), and also be extended to include additional variables such as marker/QTL genotypes and even latent variables.

2.5 Conclusions

Using structure learning algorithms of BN along with biological knowledge and reasoning, it was possible to understand the underlying causal phenotypic architecture of two fruit species of the Sapotaceae family. This procedure explores thoroughly all important and potentially unknown relationships among phenotypic traits. Despite being species from different genus, recovered structures were highly similar. It is therefore likely that the presented findings are transferable to other species of the Sapotaceae family, but further investigations relating other fruits are required. Specific knowledge of the networks can be of great value in agriculture production for the development of optimal management interventions and selective breeding.

References

ALBERT, M. J.; ESCUDERO, A.; IRIONDO, J. M. Female reproductive success of narrow endemic Erodium paularense in contrasting microhabitats. *Ecology*, v. 82(6), p. 1734-1747, 2001.

ALIA-TEJACAL, I. et al. Postharvest physiology and technology of sapote mamey fruit (*Pouteria sapota (Jacq.) HE Moore & Stearn*). *Postharvest Biology and Technology*, v. 45(3), p. 285-297, 2007.

ANDERBERG, A. A.; SWENSON, U. Evolutionary Lineages in Sapotaceae (Ericales): A Cladistic Analysis Based on ndh F Sequence Data. *International Journal of Plant Sciences*, v. 164, p. 763-773, 2003.

ARIAS, R. S. et al. Development of a large set of microsatellite markers in zapote mamey (*Pouteria sapota (Jacq.) HE Moore & Stearn*) and their potential use in the study of the species. *Molecules*, v. 20(6), p. 11400-11417, 2015.

BENTLER, P. M.; BONETT, D. G. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, v. 88(3), p. 588, 1980.

BENTLER, P. M. Comparative fit indexes in structural models. *Psychological Bulletin*, v. 107(2), p. 238, 1990.

BOUCKAERT, R. R. Bayesian belief networks: from construction to inference. Tese de Doutorado, 2001.

BROWNE, M. W.; CUDECK, R. Alternative ways of assessing model fit. *Sociological Methods & Research*, v. 21(2), p. 230-258, 1992.

CHENG, J.; BELL, D. A.; LIU, W. Learning belief networks from data: An information theory based approach. *Proceedings of the sixth international conference on Information and knowledge management*. ACM, p. 325-331, 1997.

DEL CACHO, M.; PEÑUELAS, J.; LLORET, F. Reproductive output in Mediterranean shrubs under climate change experimentally induced by drought and warming. *Perspectives in Plant Ecology, Evolution and Systematics*, v. 15(6), p. 319-327, 2013.

FOX, J. An R and S-Plus companion to applied regression. Sage, 2002.

FOX, J.; NIE, Z.; BYRNES, J. sem: structural equation models. R package version 3.0-0. 2012.

GAZEL FILHO, A. B. Caracterization sistemática de la collectión de Sapotáceas (*Pouteria sapota (Jacq.) HE Moore & Stearn*; *Manilkara zapota (L.) P. van Royen* y *Chrysophyllum cainito L*) del CATIE. MS Thesis, Centro Agronómico Tropical de Investigación y Ensenanza, CATIE, Turrialba, Costa Rica, 1995.

GAZEL FILHO, A. B. et al. Diversidad genética de la colección de zapote (*Pouteria sapota (Jacquin) HE Moore & Stearn*) del CATIE. Genetic diversity in CATIE's collection of sapodilla (Pouteria sapota (Jacquin) HE Moore & Stearn). *Plant Genetic Resources Newsletter*, n. 117, p. 37-42, 1999.

GEIGER, D.; HECKERMAN, D. Learning gaussian networks. *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence.*, Morgan Kaufmann Publishers Inc., p. 235-243, 1994.

GÓMEZ-JAIMES, R. Manejo postcosecha de zapote mamey (*Pouteria sapota (Jacq.) HE Moore and Stearn*) y su impacto en la calidad de la fruta. *Revista Chapingo*, Serie horticultura, v. 18(2), p. 253-262, 2012.

GOVAERTS, R.; FRODIN, D. G.; PENNINGTON, T. D. World checklist and bibliography of Sapotaceae. *Royal Botanic Gardens*, 2001.

GRACE, J. B.; BOLLEN, K. A. Interpreting the results from multiple regression and structural equation models. 2005.

HECKERMAN, D.; GEIGER, D.; CHICKERING, D. M. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, v. 20(3), p. 197-243, 1995.

HOOPER, D.; COUGHLAN, J.; MULLEN, M. Structural equation modelling: Guidelines for determining model fit. *Articles*, p. 2, 2008.

IRIONDO, J. M.; ALBERT, M. J.; ESCUDERO, A. Structural equation modelling: an alternative for assessing causal relationships in threatened plant populations. *Biological Conservation*, v. 113(3), p. 367-377, 2003.

JÖRESKOG, K. G. How large can a standardized coefficient be. Unpublished Technical Report, 1999.

KENNY, D. A.; MCCOACH, D. B. Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, v. 10(3), p. 333-351, 2003.

KOLLER, D.; SAHAMI, M. Toward optimal feature selection. *Stanford InfoLab*, 1996.

LI, L. et al. Poly-phenolic fraction of Chrysophyllum cainito extract induces cell death in osteosarcoma cells. *Bangladesh Journal of Pharmacology*, v. 15(4), p. 972-979, 2015.

LUO, X.; BASILE, M. J.; KENNELLY, E. J. Polyphenolic antioxidants from the fruits of Chrysophyllum cainito L.(star apple). *Journal of Agricultural and Food Chemistry*, v. 50(6), p. 1379-1382, 2002.

MARGARITIS, D.; THRUN, S. Bayesian network induction via local neighborhoods. *Advances in neural information processing systems*, p. 505-511, 2000.

MOROTA, G. et al. An assessment of linkage disequilibrium in Holstein cattle using a Bayesian network. *Journal of Animal Breeding and Genetics*, v. 129(6), p. 474-487, 2012.

MARTÍNEZ, M.; MARTÍNEZ, M. Plantas útiles de la flora mexicana. 1959.

MILES, J. N. V.; SHEVLIN, M. E. Multiple software review: Drawing path diagrams. *Structural Equation Modeling: A Multidisciplinary Journal*, v. 5(1), p. 95-103, 1998.

MORTON, J. F. et al. Fruits of warm climates. 1987.

NAGARAJAN, R.; SCUTARI, M.; LÈBRE, S. Bayesian networks in R. *Springer*, v. 122, p. 125-127, 2013.

NETO, E. C. et al. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics*, v. 4(1), p. 320, 2010.

PARKER, I. M. et al. Domestication syndrome in Caimito (*Chrysophyllum cainito L.*): fruit and seed characteristics. *Economic Botany*, v. 64(2), p. 161-175, 2010.

PEARL, J. Causality: models, reasoning, and inference, Cambridge Univ. 2000.

PEÑAGARICANO, F. et al. Exploring causal networks underlying fat deposition and muscularity in pigs through the integration of phenotypic, genotypic and transcriptomic data. *BMC Systems Biology*, v. 9(1), p. 58, 2015.

PENNINGTON, T. D. The genera of Sapotaceae. *Royal Botanic Gardens: Bronx, New York.: Kew & New York Botanical Garden*, p. 1-13, 1991.

PINO, J.; MARBOT, R.; ROSADO, A. Volatile constituents of star apple (*Chrysophyllum cainito L.*) from Cuba. *Flavour and Fragrance Journal*, v. 17(5), p. 401-403, 2002.

PINTO, R. M. et al. Genotype selection of *Pouteria sapota (Jacq.) HE Moore & Stearn*, under a multivariate framework. *Acta Agronómica*, v. 65(3), p. 312-317, 2016.

POPENOE, W. Manual of tropical and subtropical fruits: Excluding the banana, coconut, pineapple, citrus fruits, olive, and fig. *Macmillan*, 1920.

ROSA, G. J. M. et al. Inferring causal phenotype networks using structural equation models. *Genetics Selection Evolution*, v. 43(1), p. 6, 2011.

ROYSTON, J. P. Some techniques for assessing multivarate normality based on the Shapiro-Wilk W. *Applied Statistics*, p. 121-133, 1983.

SCHERMELLEH-ENGEL, K.; MOOSBRUGGER, H.; MÜLLER, H. Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of psychological research online*, v. 8(2), p. 23-74, 2003.

SCUTARI, M. Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:0908.3817, 2009.

SHIPLEY B. Cause and Correlation in Biology A User's Guide to Path Analysis, Structural Equations and Causal Inference. Cambridge University Press; 2002.

SÖRBOM, D. Model modification. *Psychometrika*, v. 54(3), p. 371-384, 1989.

SPIRTES, P.; GLYMOUR, C. N.; SCHEINES, R. Causation, prediction, and search. MIT press, 2000.

SU, C. et al. Using Bayesian networks to discover relations between genes, environment, and disease. *BioData Mining*, v. 6(1), p. 6, 2013.

SWENSON, U.; ANDERBERG, A. A. Phylogeny, character evolution, and classification of Sapotaceae (Ericales). *Cladistics*, v. 21, p. 101-130, 2005.

The R Core Team. R: A Language and Environment for Statistical Computing. Available from: http://www.r-project.org/, Vienna, Austria, 2013.

TÖPNER, K. et al. Bayesian Networks Illustrate Genomic and Residual Trait Connections in Maize (*Zea mays L.*). *G3: Genes, Genomes, Genetics*, v. 7(8), p. 2779-2789, 2017.

TSAMARDINOS, I. et al. Algorithms for Large Scale Markov Blanket Discovery. *FLAIRS Conference.*, p. 376-380, 2003.

VALENTE, B. D. et al. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, v. 185(2), p. 633-644, 2010.

VERMA, T. S.; PEARL, J. Equivalence and synthesis of causal models. *Proc. Sixth Conf. Uncertain. Artif. Intell.*, p. 220-227, 1990.

3 GENOME-WIDE ASSOCIATION STUDIES USING THREE DIFFERENT METHODS FOR QTL DETECTION IN AN F2 DUROC X PIETRAIN RESOURCE POPULATION

Abstract: The use of high-density single nucleotide polymorphism (SNP) panels has increased significantly the quantitative trait loci (QTL) mapping resolution and its applications have extended to outbred populations. The main purpose of this study was to compare results obtained from three different methodologies for genome-wide association studies (GWAS), using 35 traits measured in an F2 Duroc x Pietrain pig population, a single-marker regression, a ridge regression BLUP and a Bayes Cπ. Results showed that these methods were equally efficient in the detection of QTL regions, however we suggest the use of more than one for GWAS. In addition, important genomic regions on chromosomes 6 and 15 were detected associated with the expression of fat deposition and meat quality traits. QTLs situated in this regions can be used to assist on the process of learning phenotypic causal networks in the F2 pig population considered here, such that their absence in the modeling can bias the structures searches.

Keywords: Single-marker regression; Ridge regression BLUP; Bayes Cπ; Pleiotropic genomic regions

3.1 Introduction

Duroc and Pietrain are breeds utilized worldwide that have experienced intensive selection. These breeds have dominated the global pig industry. Duroc pigs grows faster and have more backfat than Pietrain (Edwards, Tempelman and Bates, 2006; Edwards et al., 2008a; Choi et al., 2010; Qiao et al., 2015). Several pig populations have been genotyped using microsatellite marker panels for quantitative trait loci (QTL) identification, including the F2 Duroc x Pietrain (Edwards et al., 2008a,b). However, fine mapping of these QTL is limited due to low mapping resolution. Thus, with the development of the high-density single nucleotide polymorphisms (SNPs) panels for pig genotyping (Ramos et al., 2009), that contains more than 60K SNPs (Porcine SNP60 Beadchip), QTL identification efforts have been intensified followed by changes in the types of population structure used for research studies, from intercrosses to a broad range of outbred and admixed populations (Ernst and Steibel, 2013).

Genome-wide association studies (GWAS) allow identifying genes that contribute to define the expression of economically important traits in pig breeds. Compared to QTL mapping that use microsatellite markers (e.g., Ai et al. (2012); Edwards et al. (2008a,b)), GWAS using high-density SNPs is more capable of capture enough linkage disequilibrium (LD) to identify strongest causal variants. The top SNPs are usually close to causal mutations, which allows pinpointing out the most likely candidate genes (Qiao et al., 2015).

GWAS have been successfully applied in pig populations (Sahana et al., 2013; Okumura et al., 2013; Qiao et al., 2015; Duarte et al., 2016; Casiró et al., 2017) and have been confirmed several QTLs previously reported through QTL mapping using microsatellites. For example, Qiao et al. (2015) performed a single-marker GWAS to analyze traits related to growth and fatness in two experimental populations: White Duroc x Erhualian F2 intercross and a Chinese Sutai half-sib. Duarte et al. (2016) implemented GWAS in an F2 Duroc x Pietrain population to identify genomic regions associated with traits related to growth and fat deposition, they calculate SNP effects by linear transformation of the genomic estimated breeding values (EBV) and, afterwards, selected and tested genomic segments of 2Mb that were built considering the SNPs with smallest p-values.

In this context, this study was carried out to identify genomic regions, containing major genes (genes that have markedly larger effects than the others), associated with the expression of multiple traits (pleiotropy) in an F2 Duroc x Pietrain pig population. 35 traits related to growth, carcass and meat quality were measured and submitted to analysis. Unlike the authors mentioned above, here we use three

different methodologies for GWAS: a single-marker regression (SMR), a ridge regression BLUP (RR) and a Bayes Cπ (BC). These methods were then compared regarding the SNP peaks detected in each of them.

3.2 Material and Methods

3.2.1 Data set

Animal protocols were approved by the All University Committee on Animal Use and Care at Michigan State University (Animal use form number 09/03-114-00). A 3rd-generation population from the Michigan State University Swine Teaching and Research Farm, East Lansing, MI, was used in this study (Edwards et al., 2008a,b). The initial generation (F_0) were 4 unrelated Duroc boars mated to 15 Pietrain sows by artificial insemination. From the F_1 progenies, 50 females and 6 males (sons of 3 F_0 sires) were selected, avoiding full or half sibling matings, to produce the 1,259 F_2 piglets born alive in 142 litters across 11 farrowing groups. Phenotypic data for growth, carcass and meat quality traits (Table 3.1) were collected for approximately 950 F2 pigs (details about animal management procedures and phenotyping can be consulted in Edwards et al. (2008a,b)).

Genotyping was performed using two SNP marker panels. 411 animals (including animals F_0, F_1 and 336 F_2) were genotyped using the Illumina PorcineSNP60 BeadChip (Ramos et al., 2009), and the remaining F_2 animals were genotyped using the GeneSeek Genomic Profiler for Porcine LD 9k SNP panel (GGP-Porcine, GeneSeek a Neogen Company, Lincoln, NE), which were imputed to the Illumina PorcineSNP60 Bead Chip (Duarte et al., 2013). The editing procedures performed, excluding SNPs with minor allele frequency below 0.05 and also removing animals with more than 10% of SNP missing, resulted in a data set with records from 940 pigs (F_0, F_1 and F_2) having 42,234 SNP per animal.

3.2.2 Models and statistical analysis

In this section, we detail the methods considered in this study: single-marker regression, ridge regression BLUP and Bayes Cπ. In all of them is possible to include the genomic and/or pedigree relationship matrix to control for polygenic background effects. The main difference is that in SMR the markers are independent of each other and fitted as a linear covariate; in RR, all SNPs were jointly considered in the genomic relationship matrix; and, in BC, a prior distribution was assigned to the proportion of markers included in the model.

3.2.2.1 Single-marker GWAS

Among the statistical methods for correcting confounders in GWAS are the linear mixed models (LMM), that can capture confounding by population structure, family structure and hidden relatedness simultaneously, without knowledge of which are present. LMMs use measures of genetic similarity to capture the probabilities that pairs of individuals have causative alleles in common (Lippert et al., 2011). This approach is based on a series of single-marker association analyses, such that the following model was used for each SNP j ($j = 1, 2, ..., M$):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{m_j}g_j + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{3.1}$$

where $\mathbf{y}$ is the vector of phenotypes; $\boldsymbol{\beta}$, is a vector of fixed effects of sex and litter; $\mathbf{m_j}$ is a vector of genotypes for SNP j ($j = 1, 2, ..., M$), coded as -1, 0 and 1 for AA, AB and BB, respectively; g_j is the SNP effect, assumed fixed; $\mathbf{u}$ is a vector of polygenic effect, assumed $\mathbf{u} \sim N(0, \mathbf{G}\sigma_u^2)$, where $\mathbf{G}$ is the genomic relationship matrix calculated from the molecular markers as proposed by VanRaden (2008); $\mathbf{e}$ is a vector of residuals, assumed $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$, where σ_e^2 represents non-genetic variance assumed to be acting independently on individuals, and $\mathbf{I}$ is an identity matrix; $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices.

It should be noted that the polygenic effect **u** is included in the model to account for population structure and hence reduce false positive results. A likelihood ratio test can be used for assessing the significance of each SNP at a time.

These analyses were implemented using the factored spectrally transformed linear mixed models (FaST-LMM - Lippert et al. (2011)). The FaST-LMM implementation reparametrizes the maximum likelihood (ML), or the restricted maximum likelihood (REML), as a function of only a single parameter $\delta = \sigma_e^2/\sigma_u^2$ (the ratio of the residual variance to the genetic variance) and, thus, the identification of the ML (or REML) parameters becomes an optimization over δ. In addition, FaST-LMM requires only a single spectral decomposition to test all SNPs and, consequently, provides a decrease in computational time (Lippert et al., 2011). Genome-wide significance threshold was defined according to Bonferroni method as $0.05/N$, where N is the number of informative SNPs.

3.2.2.2 Ridge Regression BLUP

According to Endelman (2011), the ridge regression was one of the first methods proposed for genomic selection, which is equivalent to best linear unbiased prediction (BLUP) in the context of mixed models. The RR-BLUP simultaneously estimates all marker effects, assuming they are random effects with homogenous variance and normal distribution (Whittaker, Thompson and Denham, 2000; Meuwissen, Hayes and Goddard, 2001).

The following model was used:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{M}\mathbf{g} + \mathbf{e} \tag{3.2}$$

where $\mathbf{y}$ is the vector of phenotypes; $\boldsymbol{\beta}$is a vector of fixed effects of sex and litter, and $\mathbf{X}$ its incidence matrix; $\mathbf{g} \sim N(0, \mathbf{I}\sigma_g^2)$ is a vector of marker effects and $\mathbf{M}$ its genotype matrix, coded as -1, 0 and 1 for AA, AB and BB, respectively; and, $\mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$ is a vector of residuals, where σ_e^2 represents non-genetic variance and $\mathbf{I}$ an identity matrix.

Particular cases (e.g. Bayesian Ridge Regression, Bayesian Lasso, Bayes A, Bayes B, and Bayes Cπ) can be specified from the general model above by assuming different distributions for the SNP effects (Gianola et al., 2009).

The analysis was implemented using the "rrBLUP" package (Endelman, 2011). Variance components were estimated by REML using the spectral decomposition algorithm of (Kang et al., 2008).

3.2.2.3 Bayes Cπ

Bayesian models allow that a huge number of markers can be analyzed simultaneously, such that the data are modeled at two levels: at the level of the data and at the level of the variance of genetic markers. Meuwissen, Hayes and Goddard (2001) presented two hierarchical Bayesian models, BayesA and BayesB, and Habier et al. (2011) proposed the BayesCπ.

The statistical model for BayesCπ is similar to that considered in RR (model 3.2), however here it is assumed that only a few genetic markers contribute with genetic variance and others have null effect. The prior distribution of marker effects is:

$$p(g_j \mid \sigma_{g_j}^2, \pi) = \begin{cases} 0 & \text{with probability} \quad \pi \\ \sim N(0, \sigma_{g_j}^2) & \text{with probability} \quad (1-\pi) \end{cases}$$

where π is the proportion of markers with null genetic effetcs (treated as unknown) and it is assigned a Beta prior $\pi \sim Beta(p_0, \pi_0)$ with $p_0 > 0$ and $\pi_0 \in [0, 1]$. The priors of all SNP effects have a common variance, $\sigma_{g_j}^2 = \sigma_g^2$, which has a scaled inverse chi-square prior with parameters v_g and S_g,

where S_g is derived from the expected value of a scale inverse chi-square distributed random variable. Thus, the effect of a SNP fitted with probability (1-π) comes from a mixture of multivariate student's t-distributions, $t(0, v, \mathbf{I}S)$ (Habier et al., 2011).

The analysis was performed using the BGLR package (de los Campos and Pérez-Rodriguez, 2013). It was assigned flat priors for fixed effects, specifically a Gaussian prior with mean zero and variance 10^{10}. By default, the following hyper-parameters values was adopted $v_g = 5$, $p_0 = 10$ and $\pi_0 = 0.5$. The scale parameter S_g is solved for to match the desired variance partition (Pérez and de los Campos, 2014). Inference of model parameters was done through Markov Chain Monte Carlo (MCMC) algorithm, which consists of Gibbs sampler steps. In this study, we considered a burn-in of 20,000, thin of 20 cycles out of 60,000.

3.3 Results and Discussion

Firstly, the 35 phenotypic traits measured in the F2 pig population were submitted to descriptive analysis. Table 3.1 shows each trait and its label, followed by its mean, standard error and coefficient of variation.

Using the three models described in the last section to search for pleiotropic genomic regions of great influence on the 35 phenotypic traits, important genomic regions were detected in the porcine chromosomes (SSC) 6 and 15, which appeared associated to a great number of traits. In this way, we specifically explored these two chromosomes. However, in the Appendix B (Table B.2. and B.3.) it is presented the three largest SNP peaks, and its position in the chromosome, for each of the 35 traits considered in this study, as well as the Manhattan plots, resulted from the use of the three models adopted: SMR, RR and BC.

Figure 3.1 shows one of the analyzes considering the three models for one of the phenotypic trait related to fat deposition: 10-th rib backfat in the sixteenth week. One can observe the great similarity among the models. In the first line of the Figure 3.1 are presented two Manhattan plots obtained from the application of SMR, the first chart considering the p-value of each SNP, together with the Bonferroni threshold, and the second, considering SNPs effects. In the second line of the figure, are the Manhattan plots, considering SNPs effects, resulting from the application of RR and BC, respectively.

According to the Figure 3.1, a great association between a region located on SSC6 and the analyzed trait was detected. Thus, we decided to explore the association of SSC6 with other phenotypic traits using the three models. Table 3.2 shows the result, in which each trait is followed by the corresponding SNP peak on SSC6 and its position in Megabase (Mb). It should be noted that in the SMR method a Bonferroni threshold was considered and, thus, effects followed by an asterisk indicate that the p-value of the association between trait and SNP was smaller than the critical Bonferroni threshold.

Considering the most recent studies (Casiró et al., 2017; Duarte et al., 2016), in which the same dataset was used, we highlight here the detection of association between a greater number of traits with genomic regions on SSC6. Casiró et al. (2017) identified association of three traits (tenth rib backfat thickness, last-lumbar vertebrae backfat thickness and loin weight) with SNP genotypes on SSC6. The 95% confidence interval for those QTL peaks overlapped each other and defined a large QTL region (between 127.6 and 140.8 Mb) on SSC6. They reported the previous identification of QTLs associated with backfat and loin weight traits in low-resolution linkage analyses studies (Edwards et al., 2008a; Choi et al., 2011), but emphasized that QTLs associated with last vertebrae lumbar backfat thickness has not been reported before. Edwards et al. (2008a) and Choi et al. (2010) performed a QTL mapping using microsatellite and showed putative QTLs on SSC 6 for fat deposition ranging from 134 to 143 cM and 164 to 174 cM, respectively.

Table 3.1. Descriptive statistics (mean, standard error and coefficient of variation) of the phenotypic traits measured in the F2 Duroc x Pietrain resource population

Label	Trait	Mean (SE)	CV (%)
BF10	10wk 10th-rib backfat (mm)	7.964 (0.058)	22.26
LRF10	10wk last-rib backfat (mm)	6.109 (0.035)	17.45
BF13	13wk 10th-rib backfat (mm)	9.725 (0.087)	27.46
LRF13	13wk last-rib backfat (mm)	7.131 (0.045)	19.39
BF16	16wk 10th-rib backfat (mm)	12.345 (0.112)	27.90
LRF16	16wk last-rib backfat (mm)	9.565 (0.075)	23.90
BF19	19wk 10th-rib backfat (mm)	15.926 (0.164)	31.58
LRF19	19wk last-rib backfat (mm)	11.794 (0.108)	27.99
BF22	22wk 10th-rib backfat (mm)	19.912 (0.209)	32.15
LRF22	22wk last-rib backfat (mm)	14.377 (0.136)	29.00
TFAT	22wk total body fat tissue (kg)	11.332 (0.103)	27.89
EBP	22wk empty body protein (kg)	9.968 (0.063)	19.27
DP	dressing percent (%)	73.000 (0.069)	2.90
CY	cook yield (%)	77.268 (0.093)	3.68
WBS	Warner-Bratzler shear force (kg)	3.208 (0.023)	21.41
JC	juiciness (1 to 8)	5.231 (0.019)	11.25
TD	tenderness (1 to 8)	5.552 (0.020)	11.08
OTD	overall tenderness (1 to 8)	5.627 (0.018)	9.84
MB	marbling (1 to 10)	2.824 (0.028)	30.00
FM	firmness (1 to 5)	2.855 (0.026)	27.63
DL	drip loss (%)	1.831 (0.038)	64.07
CT45	45min carcass temperature (ºC)	39.421 (0.071)	5.49
CT24	24h carcass temperature (ºC)	2.898 (0.039)	41.04
PH24	24h pH	5.513 (0.005)	2.53
CFBF	carcass first-rib backfat (mm)	40.619 (0.243)	17.37
CLBF	carcass last-rib backfat (mm)	28.656 (0.211)	22.48
CLLBF	carcass last-lumbar vert. backfat (mm)	22.233 (0.205)	28.10
CBF10	carcass 10th-rib backfat (mm)	24.135 (0.240)	30.33
HW	ham weight (kg)	9.633 (0.025)	8.02
LW	loin weight (kg)	8.288 (0.027)	10.08
BSW	boston shoulder weight (kg)	3.900 (0.018)	14.46
PSW	picnic shoulder weight (kg)	3.720 (0.019)	15.43
BW	belly weight (kg)	5.025 (0.022)	13.44
SW	spareribs weight (kg)	1.527 (0.007)	13.02
PT	protein (%)	23.440 (0.037)	4.84

Duarte et al. (2016) showed that a long segment of 6 Mb on SSC 6 (between 131.9 and 137.9 Mb), that included markers positioned 2 Mb up- and downstream from the extreme SNPs to cover the linkage disequilibrium (LD) of the flanking markers from the region, was associated with fat deposition traits (tenth and last rib backfat thickness from 10 to 22 weeks). The SNP peaks were M1GA0008917 (133.8855 Mb), ASGA0029651 (133.9292 Mb), ALGA0122657 (136.078566 Mb) and ALGA0104402 (136.0844 Mb). They also reported that, despite not being adjacent, SNP pairs M1GA0008917/ASGA0029651 and ALGA0122657/ALGA0104402 had substantial LD.

In this study, we found some SNP peaks on chromosome 6 associated with 27 traits (Table 3.2). The pair ALGA0122657 (136.078566 Mb) and ALGA0104402 (136.084448 Mb), consecutively located and with LD of 1, it is associated with backfat traits, total body fat tissue and empty body protein. The SNP peaks ALGA0036944 (128.386175 Mb) or ASGA0029597 (128.458999 Mb), despite not being adjacent, they had substantial LD and were associated with carcass temperature, marbling, dressing percent and primal cut weights. It can be seen that the first pair of SNPs, consecutively located, was also found by Duarte et al. (2016), although they have not reported association with some traits listed here (Table

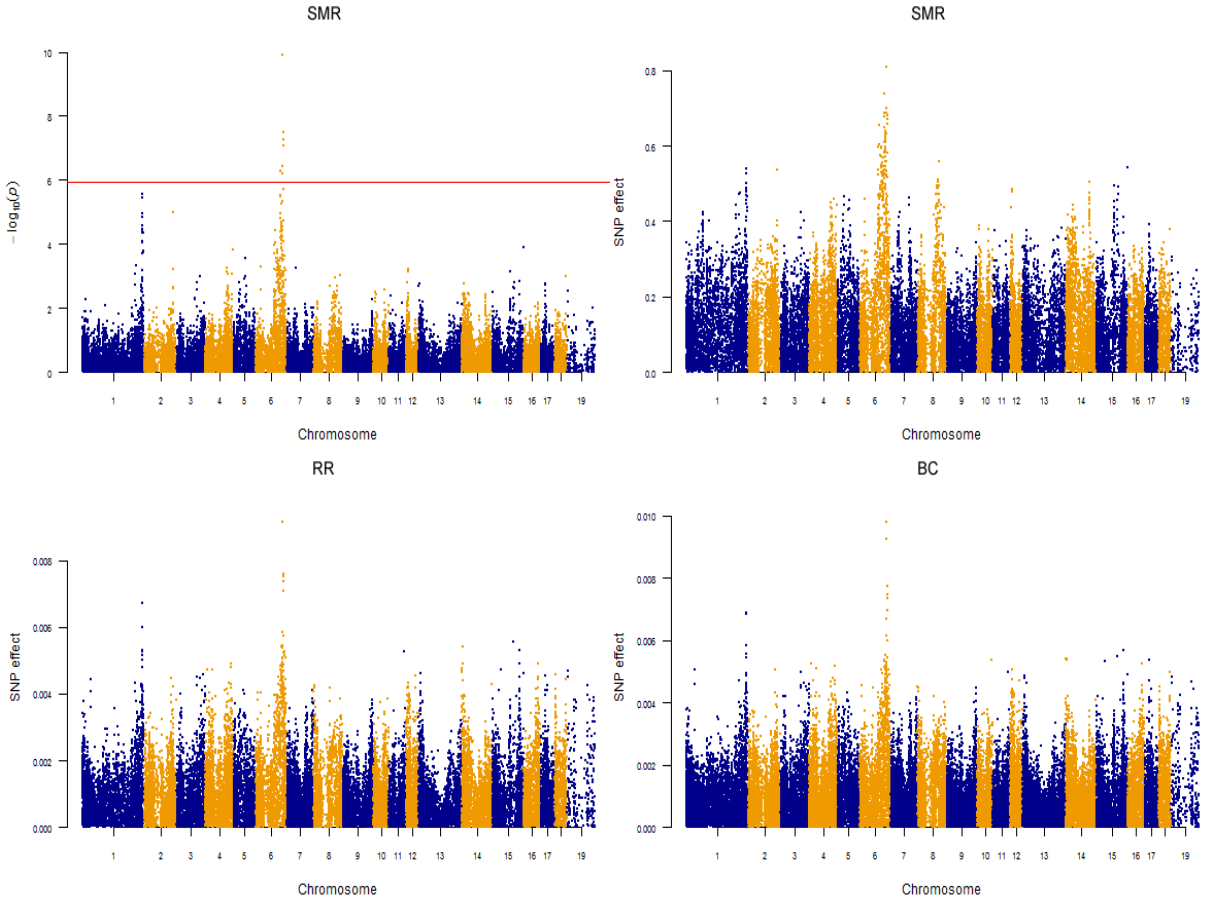

Figure 3.1. Manhattan plot for 16wk 10-th rib backfat using SMR, RR and BC

3.2), such as TFAT, EBP, CFBF and CLBF. This fact may have occurred because they used Bonferroni's level of significance, which is an extremely stringent cut off, however, here we considered three different methods for GWAS and a visual inspection to define regions of great influence.

Regarding the comparison of the three methods utilized for GWAS, one can observe that they were quite similar in the detection of association between QTLs and phenotypic traits. In all cases, at least two of the methods simultaneously found a SNP peak in the genomic region on SSC6 established by some authors as responsible for large phenotypic variation. In addition, the SNP peaks were the same in the three methods in more than 85% of the cases. The exceptions correspond to peaks included in the importance region that slightly differ in position from the highlighted SNP, for example, in SMR it was found the SNP peak ASGA0029651 (133.929215 Mb) associated to LRF10, which was also reported by Duarte et al. (2016) as a SNP peak. Other SNP peaks using SMR were: ALGA0037046 (132.322578 Mb) for CLBF, ALGA0036946 (128.444017 Mb) for DP and DL and ALGA0036046 (88.024202 Mb) for MB.

It should be reported that for some traits the RR detected two SNP peaks with similar absolute effects and p-values, for instance for all tenth-rib backfat traits the method detected markers ALGA0122657 and ALGA0104402 as SNP peaks with the same absolute effect values. However, BC method showed the SNP peak ALGA0122657 (136.078566 Mb) for all of them. In all other cases the RR and BC models appeared very similar, almost always detecting the same SNP peaks.

Next to the SNPs pair ALGA0122657 (136.078566 Mb) and ALGA0104402 (136.084448 Mb), reported in Table 3.2, it is the Leptin Receptor Overlapping Transcript (LEPROT on SSC6: 135.37 to 135.38 Mb). Leptin hormone has important effect in feed intake, growth and backfat traits, some studies reported that the serum concentrations of leptin are positively correlated with backfat thickness and

Table 3.2. Summary of the SNP peaks on SSC6 and their position (in Megabase) for the phenotypic traits listed below

Trait	SMR	RR	BC
BF10	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
BF13	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
BF16	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
BF19	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
BF22	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
LRF10	ASGA0029651* (133.9)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
LRF13	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
LRF16	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
LRF19	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
LRF22	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
TFAT	ALGA0122657 (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
EBP	ALGA0122657 (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0104402 (136.1)
CFBF	ALGA0037046 (132.3)	M1GA0008917 (133.9)	ALGA0104402 (136.1)
CLBF	ALGA0122657 (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0104402 (136.1)
CLLBF	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0122657 (136.1)
CBF10	ALGA0122657* (136.1)	ALGA0122657/ALGA0104402 (136.1)	ALGA0104402 (136.1)
DP	ALGA0036946* (128.4)	ASGA0029597 (128.5)	ASGA0029597 (128.5)
MB	ALGA0036046 (88.0)	ASGA0029597 (128.5)	ALGA0036944 (128.4)
DL	ALGA0036946 (128.4)	ALGA0036944 (128.4)	ASGA0029880 (142.3)
CT45	ALGA0036944* (128.4)	ALGA0036944 (128.4)	ALGA0036944 (128.4)
CT24	ASGA0029597* (128.5)	ASGA0029597 (128.5)	ALGA0036944 (128.4)
HW	ALGA0036944 (128.4)	ASGA0029597 (128.5)	MARC0085467 (13.2)
LW	ALGA0036944* (128.4)	ALGA0036944 (128.4)	ALGA0036944 (128.4)
BSW	ALGA0036944* (128.4)	ALGA0036944 (128.4)	ALGA0036944 (128.4)
PSW	ALGA0036944* (128.4)	ALGA0036944 (128.4)	ALGA0036944 (128.4)
BW	ALGA0036944 (128.4)	ALGA0036944 (128.4)	ALGA0036944 (128.4)
SW	ALGA0036944* (128.4)	ALGA0036944 (128.4)	ALGA0036944 (128.4)

[1]* P-values were smaller than the critical Bonferroni threshold in SMR

negatively correlated with carcass muscle content (Berg et al., 2012; Okumura et al., 2013; Casiró et al., 2017).

Regarding the region on SSC15, also associated with a great number of traits, Figure 3.2 shows one of the analyzes considering the three models for one of the phenotypic trait related to meat quality: tenderness. In the same way, one can observe the great similarity among the models, such that all of them detected large peaks on SSC2 and SSC15. In the first line of the Figure 3.2 are presented two Manhattan plots obtained from the application of SMR, the first one considering the p-value of each SNP, together with the Bonferroni threshold, and the second, considering SNPs effects. In the second line of the figure, are the Manhattan plots, considering SNPs effects, resulting from the application of RR and BC, respectively.

Exploring the association of SSC15 with other phenotypic traits using the three models, Table 3.3 shows the results, in which each trait is followed by the corresponding SNP peak on SSC15 and its position in Megabase (Mb). Bonferroni threshold was considered in the SMR and, thus, effects followed by an asterisk indicate that the p-value of the association between trait and SNP was smaller than the critical Bonferroni threshold.

As we can see in Table 3.3, regarding fat deposition and primal cut weights, the SNP peaks found on SSC15 was slightly different when comparing the SMR model with the RR and BC models. However, the results found using RR and BC were quite similar. In contrast, for meat quality traits all the three models detected similar regions.

Casiró et al. (2017) reported a QTL region on SSC15 that contains markers associated with

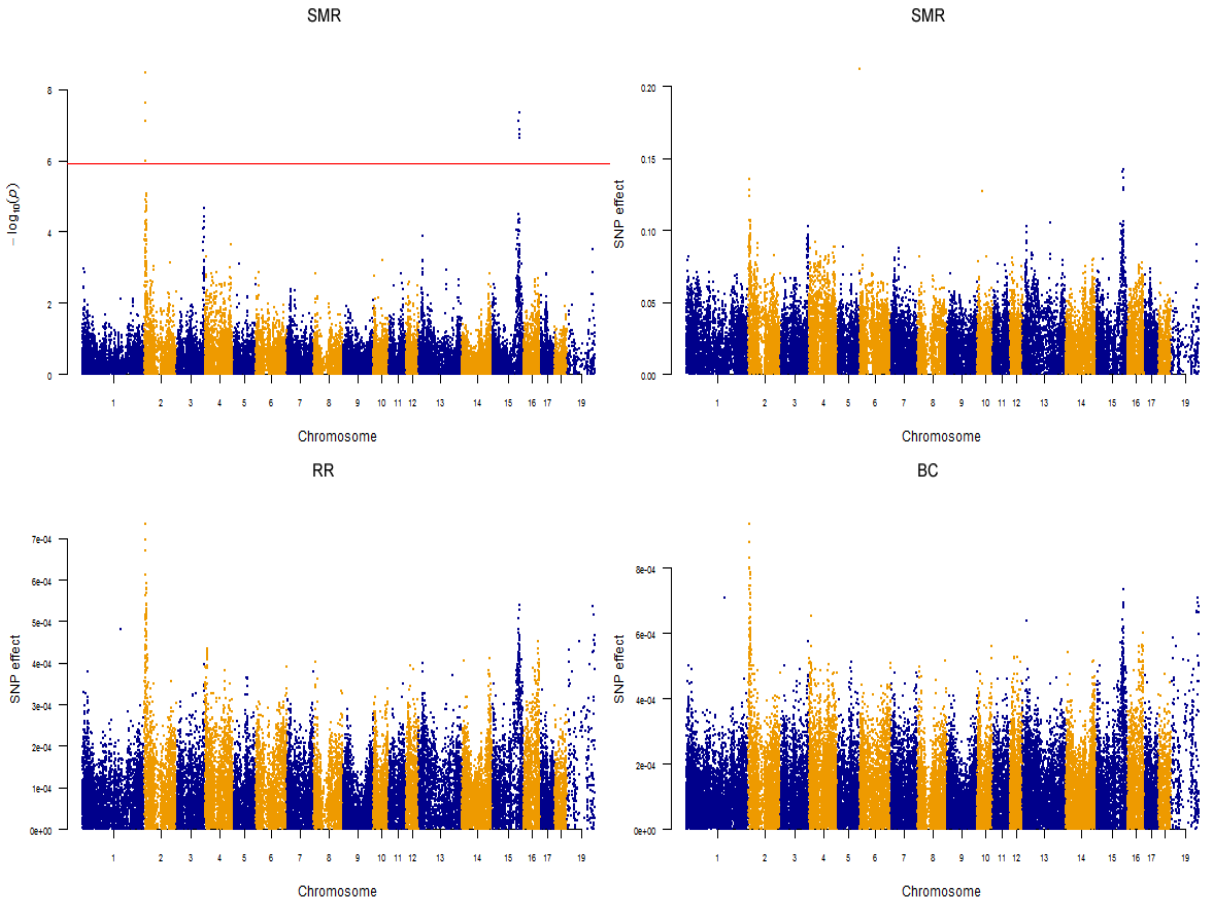

Figure 3.2. Manhattan plot for meat tenderness using SMR, RR and BC

7 traits: juiciness, tenderness/overtenderness, Warner Bratzler shear force, 24-h pH, drip loss, protein and cook yield. They reported that the SNP peaks varied across the 7 traits and, thus, they considered a single genomic region spanning from 133.4 to 137.1 Mb, because the 95% confidence interval of the QTL peaks overlapped each other. Some studies have proposed the Protein Kinase AMP-activated γ 3-subunit (PRKAG3; 133.8 Mb) as the likely candidate gene for this QTL (Choi et al., 2011; Nonneman et al., 2013; Bernal Rubio et al., 2015; Zhang et al., 2015). Here, using SMR model we found the SNP peak ALGA0087078 (133.108407 Mb) for all of these traits. RR and BC models found peaks ranging from 135.1 to 136.8 Mb, except for juiciness and drip loss that had SNP peak at position 105.595495 Mb, together with eight traits related to carcass temperature and primal cut weights.

The SSC15 has been studied in different swine populations due to its relation to meat quality traits (Thomsen et al., 2004; Rohrer et al., 2006; Edwards et al., 2008a; Li et al., 2010; Choi et al., 2011; Nonneman et al., 2013; Zhang et al., 2015), however, a very few authors reported association of this chromosome with fat deposition traits. Here, in SMR, we found a precise association between the SNP MARC0043543 (156.793131 Mb) with fat deposition traits (Table 3.3). Some backfat traits were also associated with SNPs in the range of 136 to 138.5 Mb using BC model, for example BF16 and BF19 had the SNP peak H3GA0045092 (136.981095 Mb), and LRF16 and LRF19 the SNP peaks DRGA0015530 (136.510681 Mb) and MARC0010057 (136.429511 Mb), respectively.

Finally, it should be noted that, although Sahana et al. (2010) have found best performance in a Bayesian method compared to other association mapping methods (single-marker test, haplotype-based analysis, and mixed model approach) using a simulation study with a complex pedigree structure, Legarra et al. (2015) performed the first comparison of three QTL mapping methods that correct for relatedness

Table 3.3. Summary of the SNP peaks on SSC15 and their position (in Megabase) for the phenotypic traits listed below

Trait	SMR	RR	BC
BF10	MARC0043543 (156.8)	MARC0043543 (156.8)	MARC0043543 (156.8)
BF13	MARC0043543* (156.8)	MARC0043543 (156.8)	MARC0043543 (156.8)
BF16	MARC0043543 (156.8)	ALGA0086432 (105.6)	H3GA0045092 (137.0)
BF19	MARC0043543 (156.8)	H3GA0045092 (137.0)	H3GA0045092 (137.0)
BF22	MARC0043543 (156.8)	H3GA0043939 (22.5)	H3GA0043939 (22.5)
LRF10	MARC0043543 (156.8)	ALGA0084571 (31.9)	ALGA0084571 (31.9)
LRF13	MARC0043543 (156.8)	ALGA0086091 (89.1)	MARC0043543 (156.8)
LRF16	MARC0043543 (156.8)	MARC0043543 (156.8)	DRGA0015530 (136.5)
LRF19	MARC0043543* (156.8)	ASGA0070822 (136.5)	MARC0010057 (136.4)
LRF22	MARC0043543 (156.8)	ASGA0070712 (138.5)	ASGA0070712 (138.5)
CLLBF	MARC0043543 (156.8)	MARC0043543 (156.8)	MARC0043543 (156.8)
CBF10	MARC0043543 (156.8)	MARC0043543 (156.8)	H3GA0045092 (137.0)
CY	ALGA0087078* (133.1)	ASGA0070822 (136.5)	ALGA0087317 (136.8)
WBS	ALGA0087078* (133.1)	DRGA0015526 (136.6)	MARC0047188 (135.2)
JC	ALGA0087078 (133.1)	ALGA0086432 (105.6)	ALGA0086432 (105.6)
TD	ALGA0087078* (133.1)	MARC0047188 (135.2)	H3GA0052416 (135.2)
OTD	ALGA0087078* (133.1)	MARC0047188 (135.2)	ASGA0070932 (135.1)
DL	ALGA0087078* (133.1)	ALGA0086432 (105.6)	ALGA0084571 (31.9)
PH24	ALGA0087078* (133.1)	MARC0027291 (135.2)	H3GA0052416 (135.2)
PT	ALGA0087078* (133.1)	ASGA0070822 (136.5)	ASGA0070822 (136.5)
FM	H3GA0045092 (137.0)	SIRI0000138 (136.2)	SIRI0000138 (136.2)
CT45	ALGA0086432 (105.6)	ALGA0086432 (105.6)	ALGA0086432 (105.6)
CT24	ALGA0084571* (31.9)	ALGA0086432 (105.6)	ALGA0086432 (105.6)
HW	ALGA0084571 (31.9)	ALGA0086432 (105.6)	ALGA0084571 (31.9)
LW	ALGA0084571 (31.9)	ALGA0086432 (105.6)	ALGA0086432 (105.6)
BSW	ALGA0084571 (31.9)	ALGA0086432 (105.6)	ALGA0086432 (105.6)
PSW	ALGA0084571* (31.9)	ALGA0086432 (105.6)	ALGA0086432 (105.6)
BW	ALGA0086538 (115.1)	ALGA0086432 (105.6)	ALGA0086432 (105.6)
SW	ALGA0086432* (105.6)	ALGA0086432 (105.6)	ALGA0086432 (105.6)

[1]* P-values were smaller than the critical Bonferroni threshold in SMR

in animal genetics (a linkage disequilibrium and linkage analysis, an efficient mixed-model association, and a Bayesian whole-genome regression) using real data and concluded that all the methods performed similarly. Here, in the same way, the three methods compared were also very similar and efficient.

3.4 Conclusion

Single-marker regression, ridge regression BLUP and Bayes Cπ were equally efficient in detecting genomic regions that contribute to economically important traits in the F2 Duroc x Pietrain resource population. In addition, it was detected association between genomic regions of great importance on SSC6 and SSC15 with some traits not reported in previous studies with the same population. For example, on SSC6 we found the SNP peaks ALGA0036944 (128.4 Mb) and ASGA0029597 (128.5 Mb) associated with marbling, dressing percent, drip loss, carcass temperature and primal cut weights; and, on SSC15, that is a chromosome widely reported due to its relation to meat quality traits, we found the SNP peak MARC0043543 (156.8 Mb) associated with fat deposition traits.

References

AI, H. et al. Detection of quantitative trait loci for growth-and fatness?related traits in a large-scale White Duroc x Erhualian intercross pig population.*Animal Genetics*, v. 43, n. 4, p. 383-391, 2012.

BERG, E. P. et al. Serum concentrations of leptin in six genetic lines of swine and relationship with growth and carcass characteristics. *Journal of Animal Science*, v. 81, n. 1, p. 167-171, 2003.

BERNAL RUBIO, Y. L. et al. Implementing meta-analysis from genome-wide association studies for pork quality traits. *Journal of Animal Science*, v. 93, n. 12, p. 5607-5617, 2015.

CASIRÓ, S. et al. Genome-wide association study in an F2 Duroc x Pietrain resource population for economically important meat quality and carcass traits. *Journal of Animal Science*, v. 95, n. 2, p. 545-558, 2017.

CHOI, I. et al. Application of alternative models to identify QTL for growth traits in an F 2 Duroc x Pietrain pig resource population. *BMC Genetics*, v. 11, n. 1, p. 97, 2010.

CHOI, I. et al. Identification of carcass and meat quality QTL in an F2 Duroc Pietrain pig resource population using different least-squares analysis models. *Frontiers in Genetics*, v. 2, 2011.

DE LOS CAMPOS, G.; P REZ-RODR GUEZ, P. BGLR: Bayesian generalized linear regression. R package version, v. 1, n. 3, 2013.

DUARTE, J. L. G. et al. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genetics*, v. 14, n. 1, p. 38, 2013.

DUARTE, J. L. et al. Refining genomewide association for growth and fat deposition traits in an F pig population. *Journal of Animal Science*, v. 94, n. 4, p. 1387-1397, 2016.

EDWARDS, D. B. et al. Quantitative trait loci mapping in an F Duroc Pietrain resource population: I. Growth traits. *Journal of Animal Science*, v. 86, n. 2, p. 241-253, 2008.

EDWARDS, D. B. et al. Quantitative trait locus mapping in an F Duroc Pietrain resource population: II. Carcass and meat quality traits. *Journal of Animal Science*, v. 86, n. 2, p. 254-266, 2008.

EDWARDS, D. B.; TEMPELMAN, R. J.; BATES, R. O. Evaluation of Duroc-vs. Pietrain-sired pigs for growth and composition. *Journal of Animal Science*, v. 84, n. 2, p. 266-275, 2006.

ENDELMAN, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome*, v. 4, n. 3, p. 250-255, 2011.

ERNST, C. W.; STEIBEL, J. P. Molecular advances in QTL discovery and application in pig breeding. *Trends in Genetics*, v. 29, n. 4, p. 215-224, 2013.

GIANOLA, D. et al. Additive genetic variability and the Bayesian alphabet. *Genetics*, v. 183, n. 1, p. 347-363, 2009.

HABIER, D. et al. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, v. 12, n. 1, p. 186, 2011.

KANG, H. M. et al. Efficient control of population structure in model organism association mapping. *Genetics*, v. 178, n. 3, p. 1709-1723, 2008.

LEGARRA, A. et al. A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. *Genetics Selection Evolution*, v. 47, n. 1, p. 6, 2015.

LI, H. D. et al. Quantitative trait loci analysis of swine meat quality traits. *Journal of Animal Science*, v. 88, n. 9, p. 2904-2912, 2010.

LIPPERT, Christoph et al. FaST linear mixed models for genome-wide association studies. *Nature Methods*, v. 8, n. 10, p. 833-835, 2011.

MEUWISSEN, T.H.E.; HAYES, B. J. and GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v. 157, n. 4, p. 1819-1829, 2001.

HAYES, B. J. et al. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v. 157, n. 4, p. 1819-1829, 2001.

NETO, E. C. et al. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics*, v. 4, n. 1, p. 320, 2010.

NONNEMAN, D. J. et al. Genome-wide association of meat quality traits and tenderness in swine. *Journal of Animal Science*, v. 91, n. 9, p. 4043-4050, 2013.

OKUMURA, N. et al. Genomic regions affecting backfat thickness and cannon bone circumference identified by genome?wide association study in a Duroc pig population. *Animal Genetics*, v. 44, n. 4, p. 454-457, 2013.

PÉREZ, P.; DE LOS CAMPOS, G. Genome-wide regression & prediction with the BGLR statistical package. *Genetics*, p. genetics. 114.164442, 2014.

QIAO, R. et al. Genome-wide association analyses reveal significant loci and strong candidate genes for growth and fatness traits in two pig populations. *Genetics Selection Evolution*, v. 47, n. 1, p. 17, 2015.

RAMOS, A. M. et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PloS One*, v. 4, n. 8, p. e6524, 2009.

ROHRER, G. A. et al. A genome scan for loci affecting pork quality in a Duroc?Landrace F2 population. *Animal Genetics*, v. 37, n. 1, p. 17-27, 2006.

SCUTARI, M. Learning Bayesian networks with the bnlearn R package. arXiv preprint arXiv:0908.3817, 2009.

SAHANA, G. et al. Comparison of association mapping methods in a complex pedigreed population. *Genetic Epidemiology*, v. 34, n. 5, p. 455-462, 2010.

SAHANA, G. et al. A genome-wide association scan in pig identifies novel regions associated with feed efficiency trait. *Journal of Animal Science*, v. 91, n. 3, p. 1041-1050, 2013.

THOMSEN, H. et al. Characterization of quantitative trait loci for growth and meat quality in a cross between commercial breeds of swine. *Journal of Animal Science*, v. 82, n. 8, p. 2213-2228, 2004.

TSAMARDINOS, I. et al. Algorithms for Large Scale Markov Blanket Discovery. *FLAIRS Conference*, p. 376-380, 2003.

VALENTE, B. D. et al. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, v. 185, n. 2, p. 633-644, 2010.

VANRADEN, P. M. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, v. 91, n. 11, p. 4414-4423, 2008.

WANG, C. S.; RUTLEDGE, J. J.; GIANOLA, D. Marginal inferences about variance components in a mixed linear model using Gibbs sampling. *Genetics Selection Evolution*, v. 25, n. 1, p. 41, 1993.

WHITTAKER, J. C.; THOMPSON, Robin; DENHAM, Mike C. Marker-assisted selection using ridge regression. *Genetics Research*, v. 75, n. 2, p. 249-252, 2000.

ZHANG, C. et al. Genome-wide association studies (GWAS) identify a QTL close to PRKAG3 affecting meat pH and colour in crossbred commercial pigs. *BMC Genetics*, v. 16, n. 1, p. 33, 2015.

4 SEARCHING FOR CAUSAL PHENOTYPIC NETWORKS UNDERLYING POLYGENIC TRAITS AFFECTED BY MAJOR GENES

Abstract: Graphical models such as Bayesian networks and structural equation models have been successfully used in many areas to investigate causal relationships between variables. In quantitative genetics, a constraint-based algorithm was proposed to search for recursive causal structures among phenotypes conditional to additive genetic effects. However, it does not take into account effects from major genes on traits, but only polygenic effects. Thus, a hybrid method, called PolyMaGNet (Polygenic traits with Major Genes Network analysis), is being proposed here, which consider polygenic effects based on pedigree information and those related to quantitative trait loci (QTL) that also contribute to assist on the determination of causal directions between phenotypic traits. A multiple-trait animal mixed model was fitted using a Bayesian approach considering major genes as covariates. Posterior samples of the residual covariance matrix were used as input for the Inductive Causation (IC) algorithm to search for putative causal structures, which were compared using the Akaike information criterion. Results from a simulated study considering a QTL mapping population showed that, in the presence of major genes, the PolyMaGNet recovered the true skeleton structure as well as the causal directions with a higher rate of true positives. Hence, we applied the PolyMaGNet method on real dataset of an F2 Duroc × Pietrain pig resource population to recover the causal structure underlying on carcass, meat quality and chemical composition traits.
Keywords: Bayesian networks; Phenotypic causal network; Graphical models

4.1 Introduction

Knowledge regarding causal networks underlying phenotypic traits is fundamental for the development of efficient management and breeding strategies in agricultural production (Rosa et al., 2011; Valente et al., 2013, 2015). In this context, graphical models such as Bayesian networks (BN) and structural equation models (SEM) have been successfully used in many areas to investigate causal relationships between variables, and to estimate the magnitude of such effects (Ribeiro et al., 2016; Rosa, Felipe and Pe agaricano, 2016; Sinoquet, 2014). Inferring the structure of a causal network however is not a simple task, given the large number of potential networks to be compared, even for a modest set of variables. With high-dimensional data, for example in genetics and genomics applications in which a huge number of variables are observed in each unit (animal or plant), things get even more complex (Sinoquet, 2014).

There are many structure learning algorithms for BN in the area of quantitative genetics (Logsdon and Mezey, 2010; Neto et al., 2008, 2010; Valente et al., 2010; Wang and Van Eeuwijk, 2014), which allow to explore the structure space compatible with the joint distribution function of the variables studied. Among those, Valente et al. (2010) proposed a constraint-based algorithm to search for recursive causal structures among phenotypes conditionally to unobservable polygenic effects, which act as confounders. Basically, a standard Bayesian multiple-trait model is fitted using a Markov chain Monte Carlo implementation to obtain samples from the posterior distribution of a residual covariance matrix, which are then used as input for the inductive causation (IC) algorithm (Verma and Pearl, 1990; Pearl, 2009). However, their method is based on an infinitesimal model and, as such, it does not take into account (or leverage on) the possibility of major genes affecting the traits. Often, their method produces a partially directed network that represents a class of possible equivalent solutions, and the use of prior biological knowledge is necessary to determine a final, fully directed graph.

Some alternative methods have been proposed to investigate networks in the context of gene-phenotype systems involving major genes or quantitative trait loci (QTL). For example, Schadt et al. (2005) used QTL data to infer relationships between RNA levels and complex traits. They developed

a likelihood-based causality model selection (LCMS) test that uses conditional correlation measures to decide which model (causal, reactive or independent) is best supported by the data. Basically, likelihoods associated with each of the models are constructed and maximized, and the model with the smallest Akaike Information Criterion (AIC) value is identified as the best one. Li et al. (2006) presented an extension of the work of Schadt et al. (2005) by investigating different possible causal relationships among the traits studied and, thus, providing a better characterization of the overall genetic architecture. Neto et al. (2008) proposed a hybrid algorithm, called QTL-directed dependency graph (QDG), which starts by building an undirected graph inferring associations among phenotypes using a skeleton derived from the PC (Peter-Clark) algorithm of Spirtes, Glymour and Scheines (2000). Next, a score-based step is performed including information on QTL for each phenotype to help determining causal directions in the phenotypic network using a LOD score conditional on genotypes at multiple QTL. Wang and Van Eeuwijk (2014) proposed an alternative algorithm, called QTL+ phenotype supervised orientation (QPSO), in which the main advantage is that it does not require assuming QTLs for each and every trait.

In this study, we propose a hybrid method called PolyMaGNet (Polygenic traits with Major Genes Network analysis) by combining ideas of the aforementioned approaches. The method, similarly to that proposed by Valente et al. (2010) implements a structure learning algorithm to search for recursive causal structures among complex phenotypic traits with polygenic inheritance, but allowing also the possibility of major genes affecting the traits. Such major genes are used also as instrumental variables in a final step of the algorithm to orient remaining edges not directed by the IC algorithm.

Briefly, a standard multiple-trait model is fitted using Bayesian methods considering major genes as covariates, in addition to an unobservable polygenic component. Next, posterior samples of the residual covariance matrix are used as input for the Inductive Causation (IC) algorithm to search for plausible causal network structures. In most cases, this step results in a partially oriented network. Finally, goodness of fit indexes, such as the AIC or Bayesian information criterion (BIC), are used to compare putative directed causal network within a class of structures provided by the IC algorithm. The algorithm is especially useful for the analysis of QTL mapping population data involving crosses of outbred populations. In the next few sections we provide a detailed description of the proposed method, and illustrate it with a simulation study as well as the analysis of a dataset from a F2 Duroc $\times$ Pietrain pig resource population to recover the causal structure underlying carcass and meat quality traits.

4.2 Material and Methods

4.2.1 PolyMaGNet method

The proposed PolyMaGNet method is composed by three main steps. In Step (1), a Bayesian multiple trait model (MTM) is fitted to obtain posterior samples of the residual covariance matrix. In Step (2), a partially directed acyclic graph is sought. Posterior samples of the residual covariance matrix obtained in Step (1) are used as input for the IC algorithm to obtain the initial putative Bayesian network. Lastly, in Step (3) a fully oriented causal structure is obtained. All possible networks belonging to the equivalence class of the partially directed graph resulted from the IC algorithm are scored and compared.

4.2.1.1 Multiple-trait model (MTM)

The first step of the proposed method consists of fitting a multiple-trait animal model (MTM) considering both polygenic and major gene effects:

$$\mathbf{y_i} = \mathbf{X_i}\boldsymbol{\beta} + \mathbf{u_i} + \mathbf{e_i}, \tag{4.1}$$

where $\mathbf{y_i}$ is a (tx1) vector of phenotypic records of the *ith* individual; $\mathbf{X_i}\boldsymbol{\beta}$ is a linear regression on exogenous covariates, in which the incidence matrix $\mathbf{X_i}$ contains the covariates and major genes, and $\boldsymbol{\beta}$ is a vector of fixed effects; $\mathbf{u_i}$ and $\mathbf{e_i}$ are (tx1) vectors of random additive genetic effects and residuals, respectively, assumed to be distributed as:

$$\begin{bmatrix} \mathbf{u_i} \\ \mathbf{e_i} \end{bmatrix} \sim N\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{G_0} & 0 \\ 0 & \Psi_0 \end{bmatrix} \right\}$$

where $\mathbf{G_0}$ and Ψ_0 are the additive genetic and residual covariance matrices, respectively.

Hence, the MTM model for n experimental units is described as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \tag{4.2}$$

and the joint distribution of vectors $\mathbf{u}$ and $\mathbf{e}$ is:

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

where $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices relating the vectors $\boldsymbol{\beta}$ and $\mathbf{u}$ to $\mathbf{y}$, respectively; $\mathbf{A}$ is the additive genetic relationship matrix among all individuals; and $\mathbf{I_n}$ an identity matrix of order n.

The MTM is fitted using a Bayesian approach, with the following joint prior distribution assumed for the parameters of model [4.2]:

$$p(\beta, u, G_0, \Psi_0) = p(\beta)p(G_0)p(u|G_0)\prod_{j=1}^{t} p(\psi_j) \tag{4.3}$$

$$\propto IW(G_0|v_G, G_0^{\circ})MN(u|0, G_0 \otimes A)\prod_{j=1}^{t} Inv\chi^2(\psi_j|v_\psi, s^2) \tag{4.4}$$

where $IW(\mathbf{G_0}|v_G, \mathbf{G_0^{\circ}})$ is an inverse Wishart density with v_G degrees of freedom (d.f.) and scale matrix $\mathbf{G_0^{\circ}}$, $MN(\mathbf{u}|0, \mathbf{G_0} \otimes \mathbf{A})$ is a multivariate normal density with mean vector 0 and covariance matrix $G_0 \otimes A$, $Inv\chi^2(\psi_j|v_\psi, s^2)$ is a scaled inverse chi-square distribution with v_ψ d.f. and scale parameter s^2, and ψ_j is the variance of model residuals for trait j; a uniform distribution is assumed to β.

The joint posterior distribution is then:

$$p(\beta, u, G_0, \Psi_0|y) \propto p(y|\beta, u, \Psi_0)p(G_0)p(u|G_0)\prod_{j=1}^{t} p(\psi_j). \tag{4.5}$$

A Gibbs sampling algorithm (Geman and Geman, 1984) can be employed to draw samples of the posterior distribution using its fully conditional distributions - see Valente et al. (2010) for demonstrations.

4.2.1.2 Recovering a partially directed acyclic graph

MTMs can be expressed equivalently as a recursive causal structural model, since they generate the same distribution for the response variables (Valente et al., 2010). In this way, model [4.1] can be described as (Gianola and Sorensen, 2004):

$$\mathbf{y_i} = \boldsymbol{\Lambda}\mathbf{y_i} + \mathbf{X_i}\boldsymbol{\beta} + \mathbf{u_i} + \mathbf{e_i} \tag{4.6}$$

or for n experimental units as:

$$\mathbf{y} = (\boldsymbol{\Lambda} \otimes \mathbf{I_n})\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{4.7}$$

where $\mathbf{\Lambda}$ is a (txt) matrix with zeros in the diagonal and structural coefficients in the off-diagonal.

From [4.6], the reduced model is represented as:

$$\mathbf{y_i} = (\mathbf{I_t} - \mathbf{\Lambda})^{-1}\mathbf{X_i}\boldsymbol{\beta} + (\mathbf{I_t} - \mathbf{\Lambda})^{-1}\mathbf{u_i} + (\mathbf{I_t} - \mathbf{\Lambda})^{-1}\mathbf{e_i} \tag{4.8}$$

Which implies that:

$$Var(\mathbf{y_i}) = (\mathbf{I_t} - \mathbf{\Lambda})^{-1}\mathbf{G_0}(\mathbf{I_t} - \mathbf{\Lambda})'^{-1} + (\mathbf{I_t} - \mathbf{\Lambda})^{-1}\Psi_0(\mathbf{I_t} - \mathbf{\Lambda})'^{-1}. \tag{4.9}$$

One can note that $(\mathbf{I_t} - \mathbf{\Lambda})^{-1}\mathbf{G_0}(\mathbf{I_t} - \mathbf{\Lambda})'^{-1}$ and $(\mathbf{I_t} - \mathbf{\Lambda})^{-1}\Psi_0(\mathbf{I_t} - \mathbf{\Lambda})'^{-1}$ are covariance matrices of additive genetic effects ($\mathbf{G_0^*}$) and residuals ($\mathbf{R_0^*}$) obtained from an MTM that does not account for causal relationships among phenotypes (Gianola and Sorensen, 2004; Varona, Sorensen and Thompson, 2007). The covariance matrix between traits conditionally on the additive genetic effects, considering major genes as covariates in the model, can be expressed as:

$$Var(\mathbf{y_i}|\mathbf{u_i}) = (\mathbf{I_t} - \mathbf{\Lambda})^{-1}\Psi_0(\mathbf{I_t} - \mathbf{\Lambda})'^{-1} = \mathbf{R_0^*}. \tag{4.10}$$

In this way, samples from the posterior distribution of the residual covariance matrix $\mathbf{R_0^*}$ obtained from a MTM can be used to search for a causal structure through the IC algorithm using Bayesian methods, as proposed by Valente et al. (2010).

The IC algorithm (Verma and Pearl, 1990; Pearl, 2009) has been widely used to recover an underlying acyclic structure (or a class of equivalent structures) from observed associations between traits. The search is done based on conditional independencies between variables, assuming that such independencies reflect d-separations in the underlying causal graph. Considering a set V of random variables, the IC algorithm can be described by three main steps:

1. For each pair of variable $\{A, B\}$ in V, search for a set of variables $S - \{A, B\}$ that makes $\{A\}$ independent of $\{B\}$ given $S - \{A, B\}$. If one cannot find any such set, connect both variables with an undirected edge;

2. For each connected triple $\{A, B, C\}$, such that $\{A\}$ and $\{C\}$ are nonadjacent, search for a set $S - \{A, C\}$ that contains $\{B\}$ and makes $\{A\}$ and $\{C\}$ independent given $S - \{A, C\}$. If one cannot find any such set, add arrowheads pointing to $\{B\}$: $A \rightarrow B \leftarrow C$;

3. In the resulting partially oriented graph, orient as many undirected edges as possible, without generating new colliders or any cycles.

In the context of linear mixed models, using a Bayesian framework, the following queries can be used to decide about the independence between variables A and B giving a set of variables $S - \{A, B\}$ (Valente et al., 2010):

1. Compute the posterior distribution of residual partial correlation $\rho_{(A,B|S-\{A,B\})}$, which are functions of $\mathbf{R_0^*}$.

2. Obtain the 95% highest posterior density (HPD) interval for the posterior distribution of $\rho_{(A,B|S-\{A,B\})}$.

3. If the HPD interval contains 0, declare $\rho_{(A,B|S-\{A,B\})}$ as null. Otherwise, declare $\{A\}$ and $\{B\}$ as conditionally dependent.

This process provides a partially directed acyclic graph (PDAG), that only assign directions to edges whose d-separations are supported. PDAGs represent classes of statistically equivalent BN structures (same joint probability distributions), with no cycles, containing directed edges only for nodes participating in a v-structure (Verma and Pearl, 1990).

4.2.1.3 Selecting a fully oriented causal structure

After determining the PDAG, all possible fully oriented causal structures belonging to its equivalence class are scored and compared using a model comparison criterion, such as the Akaike Information Criterion (AIC), as in Schadt et al. (2005).

Let $\Upsilon_q (q = 1, \ldots, m)$ be the fully oriented causal structures belonging to the equivalence class of the resulted PDAG of some data D, and t_q be the number of estimated parameters of Υ_q. Let $\hat{L_q}$ be the maximum value of the likelihood function for Υ_q; i.e. $\hat{L_q} = P(D \mid \hat{\theta_q}, \Upsilon_q)$, where $\hat{\theta_q}$ are the parameter values that maximize the likelihood function of Υ_q. Then the AIC value for model q is:

$$AIC_q = 2t_q - 2ln(\hat{L_q}) \tag{4.11}$$

The causal structure with the smallest AIC score is then selected as the best representative model. On the basis of the chosen causal structure retrieved, appropriate entries of Λ are treated as unknown for fitting a SEM as in Models [4.6] and [4.7], using a Bayesian approach similar to that presented in the first step.

4.2.2 Simulated data

Records from 1,600 individuals were generated by mimicking 50 families of full sibs with non-related parents. Polygenic (infinitesimal) components were simulated for five correlated traits according to the acyclic causal structure proposed by Valente et al. (2010) assuming independent residuals (Figure 4.1). In addition, seven QTLs, representing major genes, were simulated such that each trait had a single QTL and the remaining two QTLs affected three traits simultaneously. The causal model from which the data were generated can be graphically expressed as in Figure 4.1 or mathematically as a SEM:

$$\begin{cases} y_{i1k} = \mu_1 + u_{i1k} + \alpha_{11}Q_1 + \epsilon_{i1k} \\ y_{i2k} = \mu_2 + u_{i2k} + \lambda_{21}y_{i1k} + \alpha_{22}Q_2 + \alpha_{26}Q_6 + \alpha_{27}Q_7 + \epsilon_{i2k} \\ y_{i3k} = \mu_3 + u_{i3k} + \lambda_{32}y_{i2k} + \alpha_{33}Q_3 + \alpha_{36}Q_6 + \alpha_{37}Q_7 + \epsilon_{i3k} \\ y_{i4k} = \mu_4 + u_{i4k} + \lambda_{42}y_{i2k} + \alpha_{44}Q_4 + \alpha_{47}Q_7 + \epsilon_{i4k} \\ y_{i5k} = \mu_5 + u_{i5k} + \lambda_{53}y_{i3k} + \lambda_{54}y_{i4k} + \alpha_{55}Q_5 + \alpha_{56}Q_6 + \epsilon_{i5k} \end{cases} \tag{4.12}$$

where y_{ijk} and ϵ_{ijk} are the phenotype and residual effects for trait $j (j = 1, ..., 5)$ on the ith individual belonging to the kth family; μ_j is the overall mean of trait j; u_{ijk} is the additive genetic effect of the ith animal in the kth family for trait j; Q_l is the lth QTL $(l = 1, ..., 7)$; $\lambda_{jj'}$ and α_{jl} are the intensity of the effect of trait j' and QTL l on trait j, respectively. This system of equations can be expressed by model [4.13].

$$\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{y} + \Phi\mathbf{Q} + \mathbf{u} + \boldsymbol{\epsilon} \tag{4.13}$$

where,

$$\Lambda = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ \lambda_{21} & 0 & 0 & 0 & 0 \\ 0 & \lambda_{32} & 0 & 0 & 0 \\ 0 & \lambda_{42} & 0 & 0 & 0 \\ 0 & 0 & \lambda_{53} & \lambda_{54} & 0 \end{bmatrix},$$

$$\Phi = \begin{bmatrix} \alpha_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \alpha_{22} & 0 & 0 & 0 & \alpha_{26} & \alpha_{27} \\ 0 & 0 & \alpha_{33} & 0 & 0 & \alpha_{36} & \alpha_{37} \\ 0 & 0 & 0 & \alpha_{44} & 0 & 0 & \alpha_{47} \\ 0 & 0 & 0 & 0 & \alpha_{55} & \alpha_{56} & 0 \end{bmatrix}$$

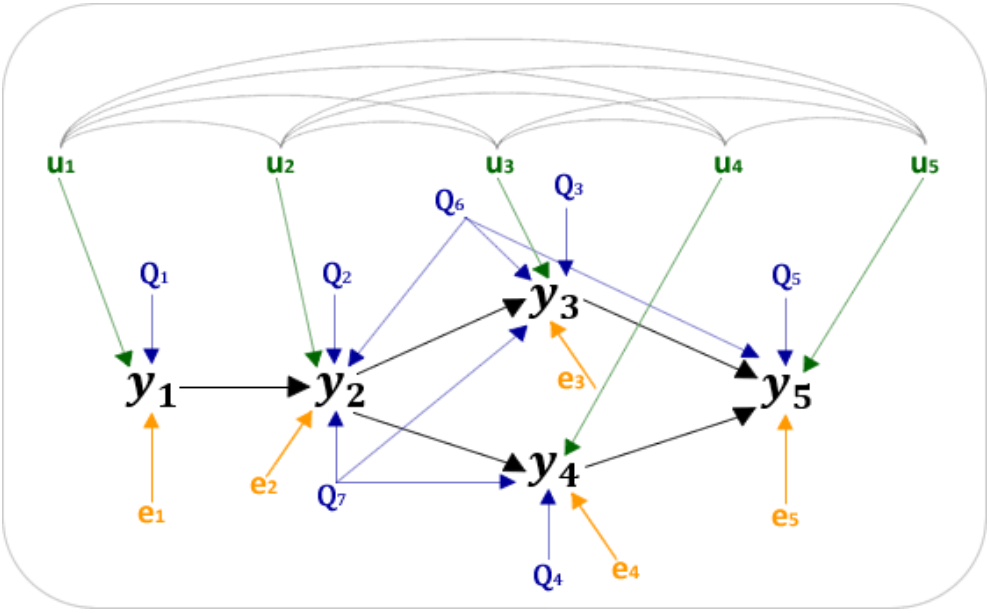

Figure 4.1. Causal graph from which simulated data were obtained; **y**'s, **u**'s, **Q**'s and **e**'s are phenotypic scores, additive genetic effects, major genes and residuals, respectively. Arcs connecting **u**'s represent genetic correlations.

The genetic relationship matrix (**A**) was a block diagonal matrix in which the first block consisted of a 100×100 identity square matrix, where the off-diagonal entries were 0 and diagonal entries were 1 (representing the relationships among parents), and the remaining 50 blocks consisted of a 30×30 square matrices where the off-diagonal entries were 0.5 (additive relationship between full sibs) and diagonal entries were 1. Vectors of additive genetic effects and residuals were kept as in Valente et al. (2010), where they were sampled from $\mathbf{u} \sim N(0, \mathbf{G_0} \otimes \mathbf{A})$ and $\mathbf{e} \sim N(0, \psi_0 \otimes \mathbf{I_n})$, respectively, in which,

$$G_0 = \begin{bmatrix} 100.00 & 47.373 & 20.283 & -38.839 & 9.773 \\ & 100.00 & 31.993 & -46.357 & -49.791 \\ & & 100.00 & 60.625 & -14.557 \\ & sym & & 100.00 & 6.490 \\ & & & & 100.00 \end{bmatrix} \text{ and}$$

$$\psi_0 = \begin{bmatrix} 200 & 0 & 0 & 0 & 0 \\ & 200 & 0 & 0 & 0 \\ & & 200 & 0 & 0 \\ & sym & & 200 & 0 \\ & & & & 200 \end{bmatrix}.$$

The simulation was repeated 100 times with different values assigned for path coefficients. Major gene effects, α's, were obtained by multiplying sampled values of an uniform distribution $U(0.1, 0.2)$ by the square root of the additive genetic variance of each trait, and λ's, representing the path coefficients between phenotypes, were sampled from $U(0.5, 1.0)$.

The MTM of step 1 was implemented using BLUPF90 family programs (Misztal et al., 2014). A single chain of 120,000 iterations was considered, discarding 20,000 as burn-in and using a thinning interval of 10 to reduce serial correlation; the remaining 10,000 samples were used to approximate features of the posterior distribution of the parameters. The remaining analyses of Step 2 and 3 were carried out using R (R Team, 2014).

4.2.3 Real data set

A 3rd-generation population from the Michigan State University Swine Teaching and Research Farm, East Lansing, MI, was used in this study (Edwards et al., 2008a,b). The initial generation (F_0) were 4 unrelated Duroc boars mated to 15 Pietrain sows by artificial insemination. From the F_1 progenies, 50 females and 6 males (sons of 3 F_0 sires) were selected, avoiding full or half sibling matings, to produce the 1,259 F_2 piglets born alive in 142 litters across 11 farrowing groups. Phenotypic data for growth, carcass and meat quality traits were collected for approximately 950 F_2 pigs (details about animal management procedures and phenotyping can be obtained in Edwards et al. (2008a,b)).

Genotyping was performed using two SNP marker panels. 411 animals (including animals F_0, F_1 and 336 F_2) were genotyped using the Illumina PorcineSNP60 BeadChip (Ramos et al., 2009), and the remaining F_2 animals were genotyped using the GeneSeek Genomic Profiler for Porcine LD 9k SNP panel (GGP-Porcine, GeneSeek a Neogen Company, Lincoln, NE), which were imputed to the Illumina PorcineSNP60 Bead Chip (Duarte et al., 2013). The editing procedures performed, excluding SNPs with minor allele frequency below 0.05, and also removing animals with more than 10% of SNP missing, resulted in a data set with records from 940 pigs (F_0, F_1 and F_2) having 42,234 SNPs per animal.

In this study, we selected two groups of phenotypic traits to recover the underlying causal structures: (i) longitudinal back fat traits; and (ii) traits related to meat quality, fat and chemical composition. Table 4.1 presents these two groups, followed by two single nucleotide polymorphism (SNP) selected from the results provided by using three different models for genome-wide association studies: a single-marker regression, a ridge regression BLUP and a Bayes Cπ.

Table 4.1. Major SNP peaks for some selected traits and their respective chromosome and position in Megabase

Label	Trait		SNP	SSC	position (Mb)
BF10wk	10wk 10th-rib backfat (mm)	1.	ALGA0122657	6	136.078566
		2.	MARC0025122	3	135.854270
BF13wk	13wk 10th-rib backfat (mm)	1.	ALGA0122657	6	136.078566
		2.	ALGA0082172	14	139.293190
BF16wk	16wk 10th-rib backfat (mm)	1.	ALGA0122657	6	136.078566
		2.	H3GA0005044	1	302.398792
BF19wk	19wk 10th-rib backfat (mm)	1.	ALGA0122657	6	136.078566
		2.	H3GA0045092	15	136.981095
BF22wk	22wk 10th-rib backfat (mm)	1.	ALGA0122657	6	136.078566
		2.	ALGA0022075	4	26.59522
WBS	Warner-Bratzler shear force (kg)	1.	M1GA0002229	2	2.921459
		2.	MARC0047188	15	135.199210
Marb	marbling (1-10)	1.	MARC0022716	10	2.581084
		2.	ALGA0043983	7	104.352654
pH45	45-min pH	1.	H3GA0055161	1	304.843284
		2.	ALGA0020318	3	104.064335
BF10	carcass 10th-rib backfat (mm)	1.	ALGA0104402	6	136.084448
		2.	H3GA0005023	1	301.969614
LMA	LM area (cm^2)	1.	ASGA0029653	6	134.141272
		2.	ASGA0081175	19	48.722031

The MTM in the first step of the PolyMaGNet method was fitted including "litter" and "sex" as covariates, along with the effects of the major genes of Table 4.1 and the kinship matrix of the animals. A chain of 3,000,000 iterations was considered in the Bayesian inference process, discarding 300,000 as burn-in and using thinning of 30 to reduce autocorrelation. Convergence was checked by visual inspection.

4.3 Results and Discussion

4.3.1 Simulation study

The PolyMaGNet method was applied to the simulated data and results were compared to those obtained using the Valente's algorithm (Valente et al., 2010). After applying the IC algorithm to the posterior samples of the residual covariance matrices, three possible causal structures were obtained (Figure 4.2), whose occurrence rates are shown in Table 4.2.

Table 4.2. Occurrence rates of the causal networks of Figure 4.2, after applying the IC algorithm to samples from the posterior distribution of the residual covariance matrix in 100 simulations

	Valente's method	**PolyMaGNet**
Fig 4.2.A	12/100 (12%)	15/100 (15%)
Fig 4.2.B	78/100 (78%)	83/100 (83%)
Fig 4.2.C	10/100 (10%)	02/100 (02%)
Skeleton	90/100 (90%)	98/100 (98%)

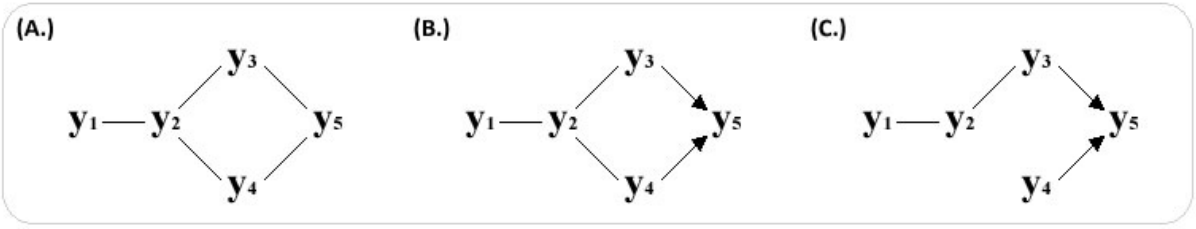

Figure 4.2. Resulted causal networks obtained by the IC algorithm to samples from the posterior distribution of R_0^*.

In the presence of major genes (Table 4.2), the PolyMaGNet method recovered the correct skeleton structure with a higher rate compared to Valente et al. (2010) approach. This may have happened because major genes act as confounders and when they are not considered in the model their effects are (at least partially) captured by the model residual, which is used as input for the IC algorithm to obtain the skeleton structure in both methods. In our simulation study, Valente's method did not recover an important edge in 10% of the simulated models, declaring Y_2 and Y_4 as independent (Figure 4.2C), which is a strong assumption in the context of causal models.

Valente's approach quite often produces a PDAG after the application of the IC algorithm. A PDAG represents a class of equivalent structures, where some edges are directed and some are undirected. The directed edges represent arrows that are common to every member in the equivalence class, while the undirected edges represent ambivalence (Pearl, 2009). Valente et al. (2010) recommended the use of prior knowledge to orient the undirected edges, respecting the possible solutions within the equivalent class. Here, however, we have measured distinct QTLs for different phenotypes and, thus, the possible solutions in the equivalence class are not likelihood equivalent anymore, because the predictive densities disagree (Neto et al., 2008). The proposed PolyMaGNet method scores all possible oriented graphs constructed from the IC output using the AIC, as in Schadt et al. (2005).

Figure 4.3 shows all possible putative graphs of the output depicted in Figure 4.2B. This model was chosen to proceed the analysis for two main reasons: (i) the occurrence rate was 83% (Table 4.2), and (ii) only 8 acyclic graphs can be constructed using this graph as starting point (Figure 4.3), which contributes to the visual inspection of all possible putative causal network. Table 4.3 shows the joint probability distribution of the 8 graphs (Figure 4.3), decomposed into a multiplication of conditional probabilities, followed by the percentage that the respective model had the smallest AIC in the 83 simulations. Hence, the PolyMaGNet method recovered the correct fully oriented causal network in 94% of the cases, without the use of any prior knowledge.

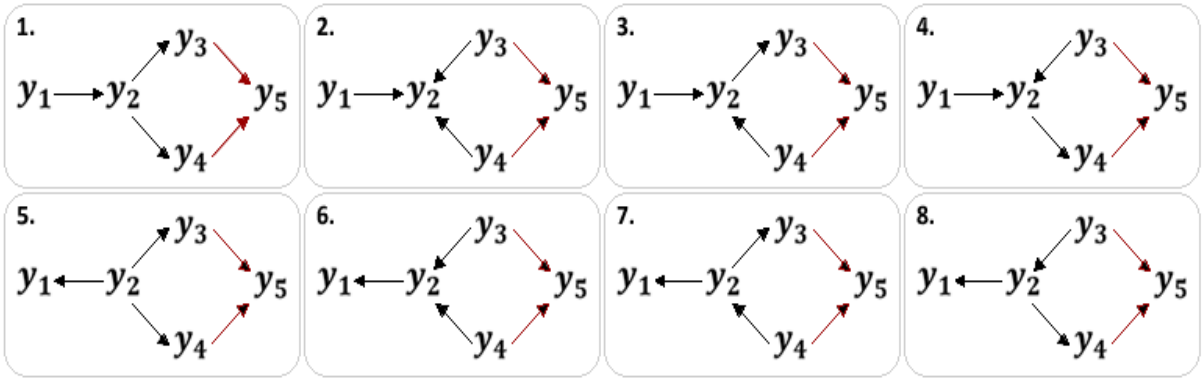

Figure 4.3. Possible graphs from the equivalent class produced by the IC output, depicetdin Figure 4.2B.

Table 4.3. Joint probability distribution of the selected partially causal graph and mean of the AIC score followed by its standard error from 83 simulations

Graph	$\mathbf{P(Y_1, Y_2, Y_3, Y_4, Y_5)}$	**Rate (%)**
1.	$P(Y_1) \times P(Y_2\|Y_1) \times P(Y_3\|Y_2) \times P(Y_4\|Y_2) \times P(Y_5\|Y_3, Y_4)$	78/83 (94)
2.	$P(Y_1) \times P(Y_3) \times P(Y_4) \times P(Y_2\|Y_1, Y_3, Y_4) \times P(Y_5\|Y_3, Y_4)$	00/83 (0)
3.	$P(Y_1) \times P(Y_4) \times P(Y_2\|Y_1, Y_4) \times P(Y_3\|Y_2) \times P(Y_5\|Y_3, Y_4)$	00/83 (0)
4.	$P(Y_1) \times P(Y_3) \times P(Y_2\|Y_1, Y_3) \times P(Y_4\|Y_2) \times P(Y_5\|Y_3, Y_4)$	00/83 (0)
5.	$P(Y_2) \times P(Y_1\|Y_2) \times P(Y_3\|Y_2) \times P(Y_4\|Y_2) \times P(Y_5\|Y_3, Y_4)$	05/83 (6)
6.	$P(Y_3) \times P(Y_4) \times P(Y_2\|Y_3, Y_4) \times P(Y_1\|Y_2) \times P(Y_5\|Y_3, Y_4)$	00/83 (0)
7.	$P(Y_4) \times P(Y_2\|Y_4) \times P(Y_3\|Y_2) \times P(Y_1\|Y_2) \times P(Y_5\|Y_3, Y_4)$	00/83 (0)
8.	$P(Y_3) \times P(Y_2\|Y_3) \times P(Y_4\|Y_2) \times P(Y_1\|Y_2) \times P(Y_5\|Y_3, Y_4)$	00/83 (0)

Figure 4.4 shows the behavior of the AIC values on the 83 simulations of the 8 possible putative causal networks (Figure 4.3). Model 1 (light blue) presented the lowest AIC values, with a very low standard error compared to other models. Model 5 (dark green) shows the lowest AIC values in only 6% of the cases, such that the AIC values for model 1 were very close in these cases. In this way, to avoid possible mistakes in studies with real datasets, we recommend to observe models with AIC values close to the lowest one and select that with reasonable biological meaning.

The same analysis was also performed with the 15 skeleton structures (Figure 4.2A) provided by the IC algorithm with PolyMaGNet method. For this purpose, the 28 possible acyclic models constructed from the skeleton were scored. Results showed that model 1 (Figure 4.3.1) had the lowest AIC value for 14 of 15 models scored (93.3%), and model 2 for only one of them (6.67%). In this way, even if the unshielded collider ($Y_3 \rightarrow Y_5 \leftarrow Y4$) was not recovered by the IC algorithm, the second step of the PolyMaGNet method was able to retrieve it.

Valente et al. (2010) also reported a simulation study with 50% reduced values for all structural coefficients and their results showed that, although the skeleton structure was still retrieved, their algorithm failed to recognize the unshielded collider $Y_3 \rightarrow Y_5 \leftarrow Y_4$. Here, 10 simulations were performed by

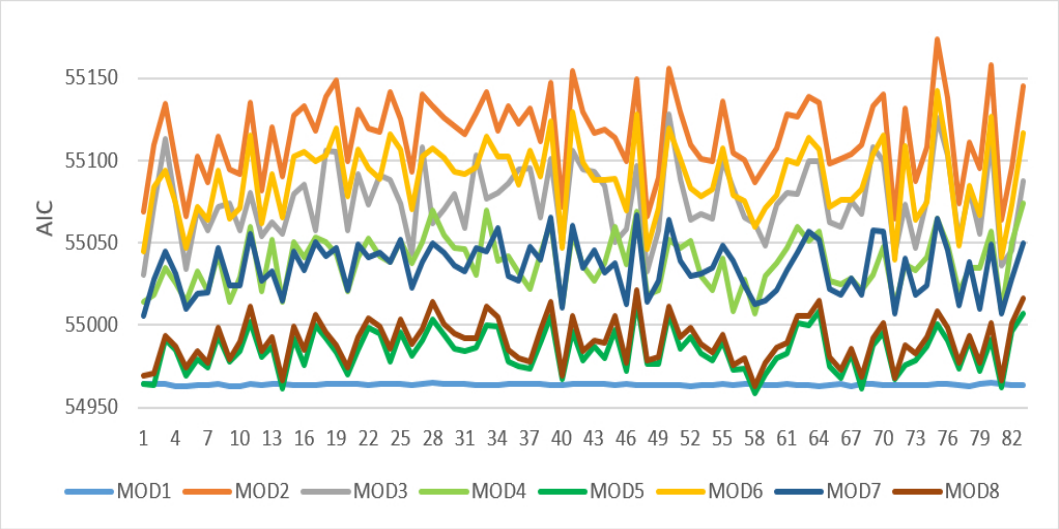

Figure 4.4. AIC values for the putative causal networks of Figure 4.3. Model 1 (light blue) presented the smallest AIC values besides a very low standard error compared to other models.

reducing all structural coefficients by 50% and both models were compared. The PolyMaGNet method recovered the correct skeleton in 100% of cases, finding the unshielded collider in 50% of them. Valente's method recovered the correct skeleton in only 30% of cases, claiming Y_1 and Y_2 as independent in the remaining 70%, probably due to the major gene effects on Y_2 not being considered in it.

The analyses were performed in a 64-bit Operating System with processor Intel(R) Core(TM) i5-2410M CPU @ 2.30GHz and 6.00 GB memory (RAM), running on Windows 10 Enterprise. It took 40min to run the IC algorithm for each model using Valente's method (67h for 100 simulations) and 45min using PolyMaGNet method (75h for 100 simulations). The second step of PolyMaGNet method, in which each possible solution of the output provided were scored by the IC algorithm, took 2min to score each of the 8 possible models totalizing 22h to obtain the results for the 83 simulations shown in the Figure 4.3. However, in the case that only the skeleton structure was recovered by the IC algorithm (Figure 4.2A), it took 3h to score the 28 possible acyclic solutions for each of the 15 models (totalizing 45h), such that the total run time varied considerably from one model to another.

4.3.2 Application to real data

In the following two subsections, the PolyMaGNet method was applied to two sets of variables related to growth traits and meat quality from a F2 pig population. In the first case, it was studied the causal structure underlying five back fat traits longitudinally measured. The purpose of this application was to validate the PolyMaGNet method using real data. Since the timeline in which the variables were collected, there is prior information on possible causal path that the system may have, as well as those that cannot happen (i.e. causal effcets backwards in time). In the second case, the PolyMaGNet method was applied to investigate the causal networks underlying five traits related to meat quality: marbling, tenderness, back fat, longissimus muscle area and pH 45min.

4.3.2.1 Growth traits

Many regions of the pig genome contribute to fat tissue phenotypes at many ages of development (Edwards et al., 2008a). Regarding back fat (BF) traits, Edwards et al. (2008a) performed a QTL-map analysis and showed a major gene on the porcine chromosome (SSC) 6 that affects back fat in all time points (10, 13, 16, 19, and 22 weeks of age), with all of them significant at the 1% of genome-wise level.

Here, a genome-wide association study was previously performed, which identified the pleiotropic SNP for all of these related traits, as well as other peaks SNPs that separately affect each one (Table 4.4).

Table 4.4. Growth traits followed by their mean, standard error and the major genes that affect them, the marker positions are in base pairs (bp)

Trait	Mean (se)	SNP	SSC	Position (bp)
BF10wk	7.96 (0.06)	MARC0025122	3	135854270
BF13wk	9.72 (0.09)	ALGA0082172	14	139293190
BF16wk	12.35 (0.11)	H3GA0005044	1	302398792
BF19wk	15.93 (0.16)	H3GA0045092	15	136981095
BF22wk	19.91 (0.21)	ALGA0022075	4	2659522
-	-	ALGA0122657	6[1]	136078566

[1]SSC6*: affect all traits simultaneously

Measures of skewness and kurtosis indicated that the BF phenotypes (Table 4.4) moderately follow the normal distribution.

By subjecting the posterior samples of the residual covariance matrix to the IC algorithm, with 95% HPD, the phenotypic structure composed of black arrows in Figure 4.5 was obtained, i.e. the unshielded collider $BF13 \rightarrow BF16 \leftarrow$ BF10 and the undirected path $BF10 - BF19$. Scoring the two possible orientation of the undirected path $BF10 - BF19$, it was obtained $BF10 \rightarrow BF19$ (AIC=12641.31) and $BF19 \rightarrow BF10$ (AIC=12642.67). Following the PolyMaGNet method guidelines, the one producing the lowest AIC, i.e. $BF10 \rightarrow BF19$, was deemed the best path represented by the data set.

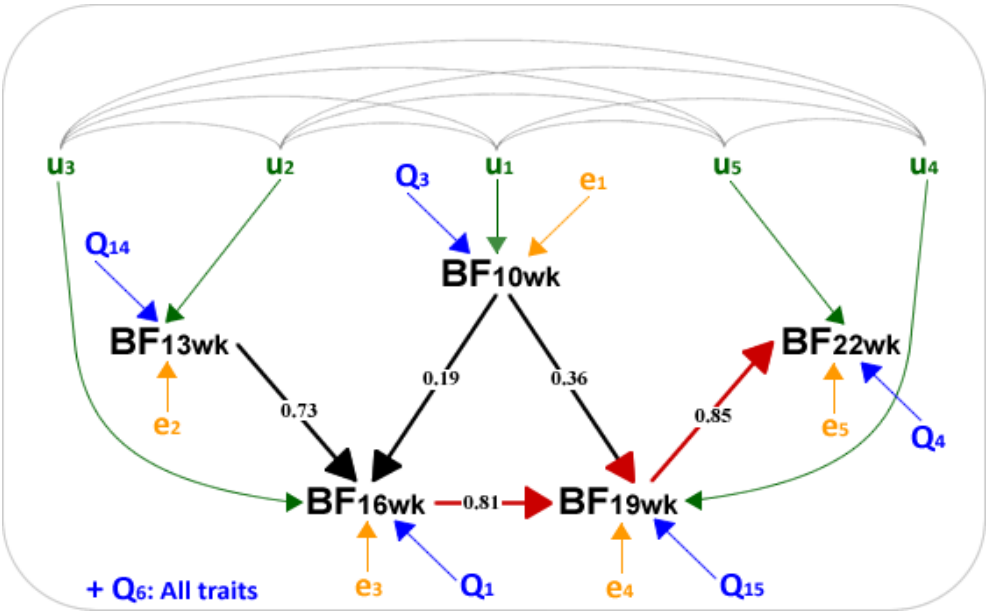

Figure 4.5. Output from PolyMaGNet method of 5 back fat (BF) traits measured at five different weeks (10^a, 13^a, 16^a, 19^a, 22^a week). u's are additive genetic effect, Q's are QTLs (major genes) and e's are residuals. Arcs connecting u's represents genetic correlations. Black edges among traits were obtained with 95% HPD and red edges with 55% HPD.

Relaxing the HPD interval to 65%, the shielded collider $BF16 \rightarrow BF19 \leftarrow BF10$ was obtained, and with 55% HPD the directed path $BF19 \rightarrow BF22$ that are represented in red in Figure 4.5. The fact previously known that these variables were measured in five consecutive times reinforces the efficiency of the PolyMaGNet algorithm, since it provided a causal structure that makes biological sense regarding the sequence of causality over time.

Applying the Valente's algorithm in the same dataset without considering the major gene effects, two undirected paths were obtained using 95% HPD: $BF10 - BF22$ and $BF16 - BF19$. Relaxing the HPD to 65% the algorithm oriented the path $BF19 \rightarrow BF16$, which does not make biological sense, since the trait BF16 was collected before BF19. This bias may have occurred because the pleiotropic major gene on SSC 6 simultaneously affects all measured traits acting as confounder in the system.

Estimates for all path coefficients are presented in Table B.1 in the Appendix C. According to the final putative causal network (Figure 4.5), BF10 does not have a direct effect on BF22, however it is possible to calculate its total effect on BF22 by adding all indirect effects of BF10 on BF22 that were mediated by BF19. Thus, the total effect of BF10 on BF22 is: $(0.19 * 0.81 * 0.85) + (0.36 * 0.85) = 0.437$. Such value can be understood as an expectation that BF22 will increase by 0.437 mm as BF10 increases by one.

The resulting causal model (Figure 4.5) provides important information that could be used in animal breeding strategies. For instance, the knowledge of the total effect of each trait on the last back fat measure (BF22) could assist deciding the best week to perform an intervention. Here, in addition, this application involving back fat traits measured longitudinally helped us to test the effectiveness of PolyMaGNet method, since as result we didn't have edges with no temporal sense, i.e., the resulting causal network corroborated our expectations.

4.3.2.2 Meat quality traits

The second analysis using real data considered five meat quality traits (Table 4.5) of a F2 population: marbling (Marb), Warner-Bratzler shear force (WBS), pH 45min (pH45), loin muscle area (LMA) and tenth-rib back fat (BF10). Marbling is a measure that expresses the amount of intramuscular fat, however, in the strict sense refers only to the fat that appears visible on cut meat surfaces (Blumer, 1963). WBS is the most widely used measure of meat tenderness, such that is the only method used for raw meat and is suitable for commercial application (Culioli, 1995; Choe et al., 2016). Meat pH is an indicator of eating quality and it is determinant for beef tenderness (Van Laack, Stevens and Stalder, 2001).

Table 4.5. Growth traits followed by their mean, standard error and the major genes that affect them, the marker positions are in base pairs (bp)

Trait	Mean (se)	SNP	SSC	Position (bp)
Marb	2.82 (0.03)	MARC0022716	10	2581084
		ALGA0043983	7	104352654
WBS	3.21 (0.02)	M1GA0002229	2	2921459
		MARC0047188	15	135199210
pH45	6.37 (0.01)	H3GA0055161	1	304843284
		ALGA0020318	3	104064335
LMA	40.61 (0.16)	ASGA0029653	6	134141272
		ASGA0081175	19	48722031
BF10	24.14 (0.24)	ALGA0104402	6	136084448
		H3GA0005023	6	301969614

Considering 95% HPD, the phenotypic structure composed of black arrows in Figure 4.6 was obtained, i.e. the unshielded collider $Marb \rightarrow WBS \leftarrow BF10$. Relaxing the HPD interval to 70%, the directed path $LMA \rightarrow WBS$ was obtained, and with 50% HPD the directed path $pH45 \rightarrow WBS$ that are represented in red in Figure 4.6. Estimates for all path coefficients are presented in Table B.2 in the Appendix C.

Wheeler et al. (1994) reported the effect of marbling degree on beef palatability in Bos Taurus and Bos Indicus cattle and concluded that the meat decreased in shear force as marbling increased

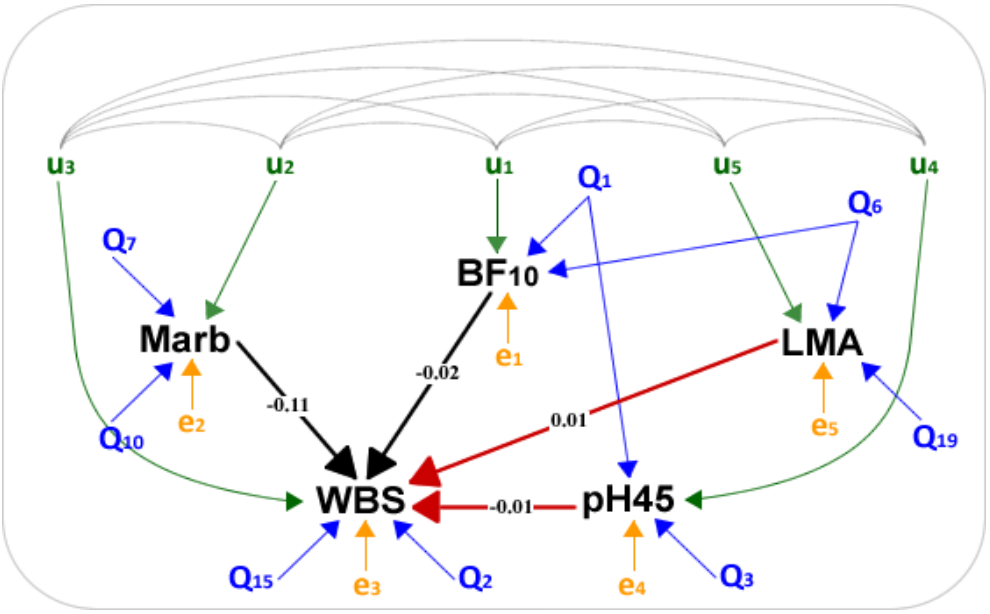

Figure 4.6. Output from PolyMaGNet method of 5 meat quality traits: marbling (Marb), Warner-Bratzler shear force (WBS), pH 45 min (pH45), longissimus muscle area (LMA) and back fat carcass (BF10). u's are additive genetic effect, Q's are QTLs (major genes) and e's are residuals. Arcs connecting u's represents genetic correlations. Black edges among traits were obtained with 95% HPD and red edges with 50% HPD.

from trace to small. However, approximately 10% of the variation in tenderness can be accounted for by marbling degree (Blumer, 1993; Wheeler et al. 1994). Here, according to the final causal network obtained, WBS will decrease in about 11% with the increase of marbling by one unit.

Regarding the relationship between pH and WBS, Laack et al. (2001) reported that, in meat from Duroc pigs, WBS decreased as pH increased, however, the same was not found with meat from Hampshire, in which WBS increased linearly as pH increased. Here, a weak causal relationship between pH45 and WBS was observed, indicating that WBS will decrease in about 1% as pH45 increases by one unit. However, great care must be taken in the interpretation of this causal path, since its standard error (Table B.2 - Appendix C) may allow the occurrence of the two scenarios discussed. The results provided in this application allow to state that when selection procedures favors production traits such as marbling, back fat, loin muscle area and pH45, it also will favor meat tenderness.

Finally, it is worth mentioning that, although we have improved the causal structure recovering method and reduced the assumptions, some of those are still required, as the causal sufficiency, which states that there are no hidden confounders and the error are jointly independent; causal Markov condition, implying that the causal structures satisfies the Markov condition; causal faithfulness, all conditional independence relations in the graph are consequences of the Markov condition applied to the true causal structure; and, major genes previously detected, since we assumed at least one distinct QTL for each phenotype, which came from earlier gene mapping of phenotypes, to ensure that the last step of the algorithm is able to differentiate all possible structures belonging to the equivalence class retrieved.

4.4 Conclusions

The hybrid method proposed, called PolyMaGNet, allows inferring Bayesian networks underlying phenotypic traits conditional to major genes and unobservable additive (polygenic) genetic effects. Such networks describe how phenotypic traits are related to each other, information of which might aid the establishment of efficient management and breeding strategies in agriculture.

Results of a simulated study considering a QTL mapping population showed that, in the presence of major genes, the PolyMaGNet was effective in recovering the correct skeleton structure and causal direction with a higher rate of true positive. PolyMaGNet was also applied to a real dataset of a F2 Duroc × Pietrain pig resource population to recover the causal structure underlying: (i) longitudinal back fat traits; and (ii) traits related to meat quality, fat and chemical composition. Regarding the first application, the final causal network was compatible with the longitudinal biological profile. In the second, the resulted causal network provided an interesting scenario showing the causal effect of marbling, back fat, longissimus muscle area and ph45 to the meat tenderness index.

Thus, if high density molecular marker data is available along with the phenotypic traits under study, more reliable causal networks can be obtained through more efficient genetic learning structures approaches, such as the PolyMaGNet. Causal methods outperform MTMs, which only describe the probabilistic relationship among traits, since they allow us to predict the effect of external interventions and, consequently, the improvement of economic important traits.

References

BLUMER, T. N. Relationship of marbling to the palatability of beef. *Journal of Animal Science*, v. 22, n. 3, p. 771-778, 1963.

CHOE, J. et al. Estimation of sensory pork loin tenderness using warner-bratzler shear force and texture profile analysis measurements. *Asian-Australasian Journal of Animal Sciences*, v. 29, n. 7, p. 1029, 2016.

CULIOLI, J. Meat tenderness: Mechanical assessment. *Expression of tissue proteinases and regulation of protein degradation as related to meat quality*, p. 239-263, 1995.

DUARTE, J. L. G. et al. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. *BMC Genetics*, v. 14, n. 1, p. 38, 2013.

DUARTE, J. L. et al. Refining genomewide association for growth and fat deposition traits in an F pig population. *Journal of Animal Science*, v. 94, n. 4, p. 1387-1397, 2016.

EDWARDS, D. B. et al. Quantitative trait loci mapping in an F Duroc Pietrain resource population: I. Growth traits. *Journal of Animal Science*, v. 86, n. 2, p. 241-253, 2008.

EDWARDS, D. B. et al. Quantitative trait locus mapping in an F Duroc Pietrain resource population: II. Carcass and meat quality traits. *Journal of Animal Science*, v. 86, n. 2, p. 254-266, 2008.

GEMAN, S.; GEMAN, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, n. 6, p. 721-741, 1984.

GIANOLA, D.; SORENSEN, D. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics*, v. 167, n. 3, p. 1407-1424, 2004.

LI, R. et al. Structural model analysis of multiple quantitative traits. *PLoS Genetics*, v. 2, n. 7, p. e114, 2006.

LOGSDON, B. A.; MEZEY, J. Gene expression network reconstruction by convex feature selection when incorporating genetic perturbations. *PLoS Computational Biology*, v. 6, n. 12, p. e1001014, 2010.

MISZTAL, Ignacy et al. Manual for BLUPF90 family of programs. Athens: University of Georgia, 2014.

NETO, E. C. et al. Inferring causal phenotype networks from segregating populations. *Genetics*, v. 179, n. 2, p. 1089-1100, 2008.

NETO, E. C. et al. Causal graphical models in systems genetics: a unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics*, v. 4, n. 1, p. 320, 2010.

PEARL, J. Causality. Cambridge university press, 2009.

RAMOS, A. M. et al. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. *PloS One*, v. 4, n. 8, p. e6524, 2009.

RIBEIRO, A. H. et al. Causal Inference and Structure Learning of Genotype?Phenotype Networks Using Genetic Variation. *Big Data Analytics in Genomics. Springer International Publishing*, p. 89-143, 2016.

ROSA, G. J. M.; FELIPE, V. P. S.; PE AGARICANO, F. Applications of Graphical Models in Quantitative Genetics and Genomics. *Systems Biology in Animal Production and Health*, Springer International Publishing, v. 1. p. 95-116, 2016.

ROSA, G. J. M. et al. Inferring causal phenotype networks using structural equation models. *Genetics Selection Evolution*, v. 43, n. 1, p. 6, 2011.

SCHADT, Eric E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, v. 37, n. 7, p. 710-717, 2005.

SINOQUET, C. Probabilistic graphical models for genetics, genomics, and postgenomics. OUP Oxford, 2014.

SPIRTES, P.; GLYMOUR, C. N.; SCHEINES, R. Causation, prediction, and search. MIT press, 2000.

Team, R. C. (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014.

VALENTE, B. D. et al. Searching for recursive causal structures in multivariate quantitative genetics mixed models. *Genetics*, v. 185, n. 2, p. 633-644, 2010.

VALENTE, B. D. et al. Is structural equation modeling advantageous for the genetic improvement of multiple traits?. *Genetics*, v. 194, n. 3, p. 561-572, 2013.

VALENTE, B. D. et al. The causal meaning of genomic predictors and how it affects construction and comparison of genome-enabled selection models. *Genetics*, v. 200, n. 2, p. 483-494, 2015.

VAN LAACK, R. L.; STEVENS, S. G.; STALDER, K. J. The influence of ultimate pH and intramuscular fat content on pork tenderness and tenderization. *Journal of Animal Science*, v. 79, n. 2, p. 392-397, 2001.

VARONA, Luis; SORENSEN, Daniel; THOMPSON, Robin. Analysis of litter size and average litter weight in pigs using a recursive model. *Genetics*, v. 177, n. 3, p. 1791-1799, 2007.

VERMA, T. S.; PEARL, J. Equivalence and synthesis of causal models [Technical report R-150]. Department of Computer Science, University of California, Los Angeles, 1990.

WANG, H.; VAN EEUWIJK, F. A. A new method to infer causal phenotype networks using QTL and phenotypic information. *PloS One*, v. 9, n. 8, p. e103997, 2014.

WHEELER, T. L.; CUNDIFF, L. V.; KOCH, R. M. Effect of marbling degree on beef palatability in Bos taurus and Bos indicus cattle. *Journal of Animal Science*, v. 72, n. 12, p. 3145-3151, 1994.

5 CONCLUSION

Graphical models, such as Bayesian networks (BN) and structural equation models, are useful tools to express causality among phenotypic traits in complex systems biology. However, the task of fitting a causal model requires that the relationships of cause-effect among variables be pre-established. When these relationships are not known, the proposed structure using prior knowledge may not express the actual biological network, fact that result in erroneous inference of causal parameters and, consequently, biased interpretation of cause intensity between variables.

In this thesis, we proposed some ways of learning causal networks using search algorithms that require accepting specific assumptions, from which the causal sufficiency seems to be the strongest one. In Chapter 2, we explored constraint- and score-based algorithms of BN to recover the underlying phenotypic networks of two fruit species of the Sapotaceae family and concluded that these fruits have highly similar biological mechanisms. However, further studies using genetic data in the search for the causal structures of these fruits could improve the results and perhaps provide different paths among traits that would allow the fusion of the two proposed causal network into only one.

In Chapter 3, we used a real data set (F2 Duroc $\times$ Pietrain pig population) to compare three different methodologies for genome-wide association studies: a single-marker regression, a ridge regression BLUP and a Bayes Cπ. Methods were equally efficient in the detection of QTL regions, however we suggest the use of more than one method for GWAS. This study also allowed us to identify genomic regions, not reported in previous studies with the same pig population, associated with the expression of fat deposition and meat quality traits.

Finally, in Chapter 4 the main goal of this research was reached, in which we proposed a hybrid algorithm, called PolyMaGNet (Polygenic traits with Major Genes Network analysis), which allows inferring BN underlying phenotypic traits conditional to major genes and unobservable additive (polygenic) genetic effects. Results of a simulated study considering a QTL mapping population showed that, in the presence of major genes, the PolyMaGNet was effective in recovering the correct skeleton structure and causal direction with a higher rate of true positive. PolyMaGNet was also applied to a real dataset of a F2 Duroc $\times$ Pietrain pig resource population to recover the causal structure underlying longitudinal back fat traits and traits related to meat quality, fat and chemical composition. Regarding the first application, the final causal network was compatible with the longitudinal biological profile. In the second, the resulted causal network provided an interesting scenario showing the causal effect of marbling, back fat, longissimus muscle area and ph45 to the meat tenderness index.

In this way, if high density molecular marker data is available, more reliable causal networks can be obtained through more efficient genetic learning structures approaches, such as the PolyMaGNet method. It is noteworthy that the proposed method could be improved in many ways, such as allowing to deal with non-Gaussian traits, as well as to handle huge number of variables. Both suggestions require more complex methodologies and, consequently, computational time.

In summary, graphical models provide a flixible and insightful approach which allow the characterization of causal phenotypic networks and its genetic architectures in complex systems biology. Such information can be used then to predict the effect of external interventions and, consequently, the improvement of economically important traits. As such, it might promote the development of breeding programs and optimal decision-making strategies.

APPENDICES

Appendix A: Supplementary Material for Chapter 2

R Code of the analysis

```
setwd("C:/Users/Badger/Desktop/PAPER - FRUTAS")
# Packages required
require("bnlearn")
require("MVN")
require("sem")
require("psych")

# Datasets
### Star Apple dataset ###
dataC <- read.table("caimito.txt",h=T)
dataC <- data.frame(dataC[-c(1,7,11:14)])
head(dataC)

### Mamey Sapote dataset ###
dataS0 <- read.table("sapote.txt",h=T)
colnames(dataS0) <- c("ID","FRW","FRL","FRD","PUT","PET","PEW",
"PUP","NSE","SEL","SED","SEW","LEL","LEW","TRH","TTD","TCD",
"PRO","SAC","GLU","FRU")
dataS <- data.frame(dataS0[-c(1,5,8,9,13:21)])
head(dataC); head(dataS)
dim(dataC); dim(dataS)

### Descriptive analysis ###
descS <- describe(dataS)
descS <- data.frame("VARIABLES"=names(dataS),"MIN"=descS$min,
"MAX"=round(descS$max,1),"MEAN"=round(descS$mean,2),
"SE"=round(descS$se,2), "CV"=round((descS$sd/descS$mean)*100,2))
descC <- describe(dataC)
descC <- data.frame("VARIABLES"=names(dataC),"MIN"=descC$min,
"MAX"=round(descC$max,1),"MEAN"=round(descC$mean,2),
"SE"=round(descC$se,2), "CV"=round((descC$sd/descC$mean)*100,2))

# Normality tests
uniNorm(dataS, type = "SW", desc = TRUE)
uniNorm(dataC, type = "SW", desc = TRUE)
# Transformation of dataC
dataC <- data.frame("FRW"=log(dataC$FRW),"FRL"=dataC$FRL,"FRD"=log(dataC$FRD),
"PET"=log(dataC$PET),"PEW"=log(dataC$PEW),"SEL"=dataC$SEL,
"SED"=log(dataC$SED), "SEW"=log(dataC$SEW))

# Multivariate normality test
par(mfrow=c(1,2))
```

```
roystonTest(dataS, qqplot=T)
roystonTest(dataC, qqplot=T)

# Structure search
bnS <- iamb(dataS, test = "cor", alpha =0.05)
bnC <- iamb(dataC, test = "cor", alpha =0.05)
par(mfrow=c(1,2))
plot(bnS, main= expression(paste("Mamey sapote - IAMB algorithm (", alpha, "=0.2)")))
plot(bnC, main= expression(paste("Star apple - IAMB algorithm (", alpha, "=0.2)")))

bnS2 <- tabu(dataS, score="bge")
bnC2 <- tabu(dataC, score="bge")
plot(bnS2, main= "Mamey sapote - Tabu search")
plot(bnC2, main= "Star apple - Tabu search")

# Jackknife - Mamey sapote - IAMB
m <- matrix(0,8,8)
t <- matrix(0,8,8)
for(i in 1:nrow(dataS)){
f = dataS[-i,]
bnS <- iamb(f, test = "cor", alpha =0.2)
f1 <- amat(bnS)
m = m + f1
f2 <- f1-t(f1)
f3 <- f2>0
t = t + f3 }
print(m);print(t)

# Jackknife - Mamey sapote - Tabu
m <- matrix(0,8,8)
t <- matrix(0,8,8)
for(i in 1:nrow(dataS)){
f = dataS[-i,]
bnS <- tabu(dataS, score="bge")
f1 <- amat(bnS)
m = m + f1
f2 <- f1-t(f1)
f3 <- f2>0
t = t + f3 }
print(m);print(t)

# Jackknife - Star apple - IAMB
m <- matrix(0,8,8)
t <- matrix(0,8,8)
for(i in 1:nrow(dataC)){
f = dataC[-i,]
bnC <- iamb(f, test = "cor", alpha =0.2)
```

```
f1 <- amat(bnC)
m = m + f1
f2 <- f1-t(f1)
f3 <- f2>0
t = t + f3 }
print(m);print(t)

# Jackknife - Star Apple - Tabu
m <- matrix(0,8,8)
t <- matrix(0,8,8)
for(i in 1:nrow(dataC)){
f = dataC[-i,]
bnC <- tabu(dataC, score="bge")
f1 <- amat(bnC)
m = m + f1
f2 <- f1-t(f1)
f3 <- f2>0
t = t + f3 }
print(m);print(t)

# Model adjustment with SEM package
# Star Apple
modelC <- specifyModel()
FRW -> PET, lam1, NA
FRW -> PEW, lam2, NA
FRW -> FRD, lam3, NA
FRW -> FRL, lam4, NA
FRW -> SED, lam5, NA
PEW -> PET, lam6, NA
PEW -> FRL, lam7, NA
FRL -> PET, lam8, NA
FRL -> SEW, lam9, NA
SED -> SEL, lam10, NA
SED -> SEW, lam11, NA
SED <-> SED, alp1, NA
SEL <-> SEL, alp2, NA
SEW <-> SEW, alp3, NA
FRW <-> FRW, alp4, NA
FRL <-> FRL, alp5, NA
FRD <-> FRD, alp6, NA
PET <-> PET, alp7, NA
PEW <-> PEW, alp8, NA
sem2 <- sem(modelC, cor(dataC), N=nrow(dataC))
summary(sem2,fit.indices=c("GFI","AGFI","RMSEA"))
modIndices(sem2)

# Mamey Sapote
```

```
modelS <- specifyModel()
FRW -> PET, lam11, NA
FRW -> PEW, lam12, NA
FRW -> FRD, lam13, NA
FRL -> FRD, lam14, NA
FRL -> SEL, lam15, NA
FRD -> SEL, lam16, NA
SEL -> SEW, lam17, NA
SED -> SEW, lam18, NA
PEW -> PET, lam19, NA
PEW -> FRL, lam20, NA
PEW -> SED, lam21, NA
SED <-> SED, alp1, NA
SEL <-> SEL, alp2, NA
SEW <-> SEW, alp3, NA
FRW <-> FRW, alp4, NA
FRL <-> FRL, alp5, NA
FRD <-> FRD, alp6, NA
PET <-> PET, alp7, NA
PEW <-> PEW, alp8, NA
sem <- sem(modelS, cor(dataS), N=nrow(dataS))
summary(sem,fit.indices=c("GFI","AGFI","RMSEA"))
modIndices(sem)

# Diagrama de caminhos
pathDiagram(sem,style="ram",edge.labels="values",
edge.colors = c("black", "red"), output.type = c("html"),
rank.direction = c("TB"), edge.weight = c("proportional")
,standardize = T,ignore.self=T,error.nodes = TRUE,
main="Mamey Sapote")
pathDiagram(sem2,style="ram",edge.labels="values",
edge.colors = c("black", "red"), output.type = c("html"),
rank.direction = c("TB"), edge.weight = c("proportional")
,standardize = T,ignore.self=T,error.nodes = TRUE,
main="Star apple")

require("semPlot")
par(mfrow=c(2,1))
semPaths(sem, what="paths", whatLabels="est", residuals = F,
color = c("lightgreen", "lightgreen", "lightgreen", "khaki1", "khaki1",
"lightblue", "lightblue", "lightblue"),
style="ram", layout="tree2",
edge.color = c("dimgray"), edge.label.cex = 0.9)
title("Mamey sapote", line = 3)
semPaths(sem2, what="paths", whatLabels="est", residuals = F,
color = c("lightgreen", "lightgreen", "lightgreen", "khaki1", "khaki1",
"lightblue", "lightblue", "lightblue"),
```

```
style="ram", layout="tree2",
edge.color = c("dimgray"), edge.label.cex = 0.9)
title("Star apple", line = 3)

# Plots
par(mfrow=c(2,2))
plot(bnS, main= expression(paste("Mamey sapote - IAMB algorithm (", alpha, "=0.05)")))
plot(bnC, main= expression(paste("Star apple - IAMB algorithm (", alpha, "=0.05)")))
plot(bnS2, main= "Mamey sapote - Tabu search")
plot(bnC2, main= "Star apple - Tabu search")

# Equivalence classes
bnS2 <- tabu(dataS, score="bge")
bnC2 <- tabu(dataC, score="bge")
plot(bnS2, main= "Mamey sapote - Tabu search")
plot(bnC2, main= "Star apple - Tabu search")
test <- cpdag(bnS2, moral = TRUE, wlbl = FALSE, debug = FALSE)
test2 <- cpdag(bnC2, moral = TRUE, wlbl = FALSE, debug = FALSE)
plot(test)
plot(test2)
par(mfrow=c(2,2))

# After restabilish new directions in the CPDAG
# Model adjustment with SEM package
# Star Apple
modelC <- specifyModel()
FRW <- PET, lam1, NA
FRW <- PEW, lam2, NA
FRW -> FRD, lam3, NA
FRW <- FRL, lam4, NA
FRW -> SED, lam5, NA
PEW -> PET, lam6, NA
PEW <- FRL, lam7, NA
FRL -> PET, lam8, NA
FRL -> SEW, lam9, NA
SED -> SEL, lam10, NA
SED -> SEW, lam11, NA
SED <-> SED, alp1, NA
SEL <-> SEL, alp2, NA
SEW <-> SEW, alp3, NA
FRW <-> FRW, alp4, NA
FRL <-> FRL, alp5, NA
FRD <-> FRD, alp6, NA
PET <-> PET, alp7, NA
PEW <-> PEW, alp8, NA
sem2 <- sem(modelC, cor(dataC), N=nrow(dataC))
summary(sem2,fit.indices=c("GFI","AGFI","RMSEA"))
```

```
modIndices(sem2)

# Mamey Sapote
modelS <- specifyModel()
FRW <- PET, lam11, NA
FRW <- PEW, lam12, NA
FRW -> FRD, lam13, NA
FRL -> FRD, lam14, NA
FRL -> SEL, lam15, NA
FRD -> SEL, lam16, NA
SEL -> SEW, lam17, NA
SED -> SEW, lam18, NA
PEW -> PET, lam19, NA
PEW <- FRL, lam20, NA
PEW -> SED, lam21, NA
SED <-> SED, alp1, NA
SEL <-> SEL, alp2, NA
SEW <-> SEW, alp3, NA
FRW <-> FRW, alp4, NA
FRL <-> FRL, alp5, NA
FRD <-> FRD, alp6, NA
PET <-> PET, alp7, NA
PEW <-> PEW, alp8, NA
sem <- sem(modelS, cor(dataS), N=nrow(dataS))
summary(sem,fit.indices=c("GFI","AGFI","RMSEA"))
modIndices(sem)

# Diagrama de caminhos
require("semPlot")
par(mfrow=c(1,1))
semPaths(sem, what="paths", whatLabels="est", residuals = F,
color = c("lightgreen", "lightgreen", "lightgreen", "khaki1", "khaki1",
"lightblue", "lightblue", "lightblue"),
style="ram", layout= matrix, rotation = 1,
edge.color = c("dimgray"), edge.label.cex = 0.9)
title("Mamey sapote", line = 3)
semPaths(sem2, what="paths", whatLabels="est", residuals = F,
color = c("lightgreen", "lightgreen", "lightgreen", "khaki1", "khaki1",
"lightblue", "lightblue", "lightblue"),
style="ram", layout= matrix, rotation = 1,
edge.color = c("dimgray"), edge.label.cex = 0.9)
title("Star apple", line = 3)
```

Appendix B: Supplementary Material for Chapter 3

Supplementary tables and figures

Table B.1. Name of the traits analyzed followed by their respective abbreviations

Trait name	Trait abbreviation
10wk 10th-rib backfat (mm)	BF10
13wk 10th-rib backfat (mm)	BF13
16wk 10th-rib backfat (mm)	BF16
19wk 10th-rib backfat (mm)	BF19
22wk 10th-rib backfat (mm)	BF22
10wk last-rib backfat (mm)	LRF10
13wk last-rib backfat (mm)	LRF13
16wk last-rib backfat (mm)	LRF16
19wk last-rib backfat (mm)	LRF19
22wk last-rib backfat (mm)	LRF22
22wk total body fat tissue (kg)	TFAT
22wk empty body protein (kg)	EBP
carcass first-rib backfat (mm)	CFBF
carcass last-rib backfat (mm)	CLBF
carcass last-lumbar vert. backfat (mm)	CLLBF
carcass 10th-rib backfat (mm)	CBF10
dressing percent (%)	DP
cook yield (%)	CY
Warner-Bratzler shear force (kg)	WBS
juiciness (1 to 8)	JC
tenderness (1 to 8)	TD
overall tenderness (1 to 8)	OTD
marbling (1-10)	MB
firmness (1 to 5)	FM
drip loss (%)	DL
45min carcass temperature (^{o}C)	CT45
24h carcass temperature (^{o}C)	CT24
24h pH	PH24
ham weight (kg)	HW
loin weight (kg)	LW
boston shoulder weight (kg)	BSW
picnic shoulder weight (kg)	PSW
belly weight (kg)	BW
spareribs weight (kg)	SW
protein (%)	PT

Table B.2. The three largest SNP peaks detected in each of the listed backfat traits, followed by their chromosomes (SSC) and positions in Megabase (in parentheses)

Trait	SMR	RR	BC
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
BF10	MARC0043543 (15-156.8)	MARC0046321 (19-74.7)	MARC0025122 (3-135.8)
	ALGA0107397 (13-194.1)	MARC0043543 (15-156.8)	MARC0046321 (19-74.7)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
BF13	H3GA0010564* (3-119.3)	ASGA0022527 (4-129.8)	ALGA0082172 (14-139.3)
	MARC0043543* (15-156.8)	ALGA0082172 (14-139.3)	ASGA0018328 (4-12.2)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
BF16	H3GA0005192 (1-304.3)	H3GA0005044 (1-302.4)	H3GA0005044 (1-302.4)
	MARC0087200 (2-146.7)	ALGA0086432 (15-105.6)	H3GA0045092 (15-137.0)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
BF19	MARC0043543 (15-156.8)	H3GA0045092 (15-137.0)	H3GA0045092 (15-137.0)
	ASGA0080745 (19-8.6)	ASGA0080745 (19-8.6)	ASGA0018328 (4-12.2)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
BF22	ALGA0024802 (4-43.6)	ALGA0022075 (4-2.7)	ASGA0007789 (1-302.3)
	MARC0043543 (15-156.8)	ALGA0020170 (3-100.3)	ALGA0022075 (4-2.7)
	ASGA0029651* (6-133.9)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
LRF10	MARC0043543 (15-156.8)	ASGA0068060 (14-150.5)	ASGA0068060 (14-150.5)
	ASGA0054658 (12-43.4)	ASGA0051711 (11-77.0)	ALGA0114192 (11-26.5)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
LRF13	MARC0043543 (15-156.8)	ALGA0022075 (4-2.7)	ALGA0046186 (8-7.7)
	ASGA0074238 (16-77.7)	ALGA0046186 (8-7.7)	ASGA0018328 (4-12.2)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
LRF16	MARC0043543 (15-156.8)	ASGA0018328 (4-12.2)	ASGA0018328 (4-12.2)
	ALGA0031974 (5-56.5)	ALGA0074276 (14-1.8)	ALGA0074276 (14-1.8)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
LRF19	MARC0043543* (15-156.8)	ALGA0074276 (14-1.8)	ASGA0060319 (14-0.2)
	MARC0087200 (2-146.7)	H3GA0005192 (1-304.3)	ASGA0082996 (1-304.8)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
LRF22	MARC0043543 (15-156.8)	ALGA0074276 (14-1.8)	ASGA0070712 (15-138.5)
	MARC0043291 (4-109.4)	ALGA0022075 (4-2.7)	ALGA0022075 (4-2.7)
	DIAS0001383 (4-109.8)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
TFAT	ALGA0122657 (6-136.1)	ASGA0018328 (4-12.2)	ASGA0018328 (4-12.2)
	DRGA0000505 (1-36.5)	MARC0063610 (13-164.0)	MARC0063610 (13-164.0)
	ALGA0046300 (8-6.4)	ALGA0022075 (4-2.7)	ALGA0022075 (4-2.7)
EBP	ALGA0122657 (6-136.1)	ASGA0051711 (11-77.0)	ALGA0104402 (6-136.1)
	ALGA0027303 (4-109.6)	ALGA0122657 (6-136.1)	ASGA0051711 (11-77.0)
	ASGA0025575 (5-59.7)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
CFBF	ALGA0037046 (6-132.3)	ASGA0103989 (19-140.3)	ALGA0050238 (8-146.4)
	ALGA0024536 (4-36.8)	M1GA0008917 (6-133.9)	MARC0001310 (19-140.1)
	ALGA0122657 (6-136.1)	ALGA0122657/ALGA0104402	ALGA0104402 (6-136.1)
CLBF	MARC0032012 (5-67.3)	ALGA0031940 (5-53.9)	CASI0009949 (10-53.9)
	ALGA0019868 (3-83.6)	CASI0009949 (10-53.9)	H3GA0031331 (11-11.9)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0122657 (6-136.1)
CLLBF	ASGA0020769 (4-97.6)	ALGA0100124 (19-126.0)	ASGA0008038 (1-304.6)
	MARC0043543 (15-156.8)	H3GA0026371 (9-10.8)	ALGA0100124 (19-126.0)
	ALGA0122657* (6-136.1)	ALGA0122657/ALGA0104402	ALGA0104402 (6-136.1)
CBF10	MARC0043543 (15-156.8)	H3GA0005192 (1-304.3)	H3GA0005023 (1-302.0)
	H3GA0005192 (1-304.3)	MARC0043543 (15-156.8)	H3GA0045092 (15-137.0)
	ALGA0036046 (6-88.0)	MARC0022716 (10-2.6)	MARC0022716 (10-2.6)
MB	ALGA0092930 (17-7.3)	ALGA0108658 (7-104.5)	ALGA0043983 (7-104.3)
	ASGA0019822 (4-65.6)	ASGA0052010 (11-81.6)	ALGA0036944 (6-128.4)
	MARC0063610* (13-164.0)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
DP	ASGA0025539* (5-54.3)	MARC0063610 (13-164.0)	MARC0063610 (13-164.0)
	ALGA0036946* (6-128.4)	ALGA0086432 (15-105.6)	ALGA0024536 (4-36.8)

Table B.3. The three largest SNP peaks detected in each of the listed traits, followed by their chromosomes (SSC) and positions in Megabase (in parentheses)

Trait	SMR	RR	BC
	ALGA0087078* (15-133.1)	ASGA0070822 (15-136.5)	ALGA0087317 (15-136.8)
CY	ALGA0087273 (4-75.0)	ALGA0087273 (4-75.0)	ALGA0087273 (4-75.0)
	MARC0036560 (5-68.3)	MARC0036560 (5-68.3)	INRA0056638 (19-38.2)
	M1GA0025499* (2-5.5)	M1GA0025499 (2-5.5)	M1GA0002229 (2-2.9)
WBS	ALGA0087078* (15-133.1)	ALGA0036313 (6-101.3)	ALGA0036313 (6-101.3)
	MARC0050164 (10-43.5)	DRGA0015526 (15-136.6)	MARC0047188 (15-135.2)
	ALGA0087078 (15-133.1)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
JC	ASGA0010464 (2-62.1)	ALGA0024536 (4-36.8)	ALGA0050238 (8-146.4)
	ALGA0117175 (12-61.6)	ALGA0086432 (15-105.6)	MARC0063610 (13-164.0)
	H3GA0005676* (2-5.9)	H3GA0005676 (2-5.9)	H3GA0005676 (2-5.9)
TD	ALGA0087078* (15-133.1)	MARC0047188 (15-135.2)	H3GA0052416 (15-135.2)
	H3GA0011028 (3-136.7)	ASGA0081500 (19-129.9)	ALGA0100249 (19-137.0)
	H3GA0005676* (2-5.9)	H3GA0005676 (2-5.9)	H3GA0005676 (2-5.9)
OTD	ALGA0087078* (15-133.1)	MARC0047188 (15-135.2)	ALGA0007028 (1-193.3)
	H3GA0011028 (3-136.7)	ALGA0007028 (1-193.3)	ASGA0070932 (15-135.1)
	H3GA0045092 (15-137.0)	MARC0031918 (6-80.6)	SIRI0000138 (15-136.2)
FM	ALGA0045048 (7-120.7)	SIRI0000138 (15-136.2)	ASGA0081560 (19-138.5)
	MARC0031918 (6-80.6)	ASGA0081560 (19-138.5)	SIRI0001406 (14-134.7)
	ALGA0087078* (15-133.1)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
DL	ALGA0050238* (8-146.4)	ALGA0050238 (8-146.4)	ALGA0050238 (8-146.4)
	ALGA0058270* (10-34.8)	ALGA0024536 (4-36.8)	MARC0063610 (13-164.0)
	ASGA0051711* (11-77.0)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
CT45	MARC0063610* (13-164.0)	MARC0063610 (13-164.0)	MARC0063610 (13-164.0)
	ALGA0058270* (10-34.8)	ALGA0086432 (15-105.6)	ALGA0086432 (15-105.6)
	ASGA0051711* (11-77.0)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
CT24	ALGA0084571* (15-31.9)	ALGA0086432 (15-105.6)	MARC0063610 (13-164.0)
	MARC0063610* (13-164.0)	MARC0063610 (13-164.0)	ALGA0086432 (15-105.6)
	ALGA0087078* (15-133.1)	MARC0027291 (15-135.2)	H3GA0052416 (15-135.2)
PH24	MARC0036096 (16-19.7)	ALGA0099582 (19-38.2)	INRA0056800 (19-71.5)
	MARC0010481 (7-98.6)	H3GA0034274 (12-37.5)	ALGA0030427 (5-10.3)
	MARC0063610* (13-164.0)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
HW	ALGA0050238* (8-146.4)	MARC0063610 (13-164.0)	MARC0063610 (13-164.0)
	ASGA0051711* (11-77.0)	ALGA0024536 (4-36.8)	ALGA0024536 (4-36.8)
	ASGA0051711* (11-77.0)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
LW	ALGA0050238* (8-146.4)	MARC0063610 (13-164.0)	MARC0063610 (13-164.0)
	MARC0063610* (13-164.0)	ALGA0086432 (15-105.6)	ALGA0086432 (15-105.6)
	ASGA0051711* (11-77.0)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
BSW	MARC0063610* (13-164.0)	MARC0063610 (13-164.0)	MARC0063610 (13-164.0)
	ALGA0050238* (8-146.4)	ALGA0086432 (15-105.6)	ALGA0086432 (15-105.6)
	MARC0063610* (13-164.0)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
PSW	ASGA0051711* (11-77.0)	MARC0063610 (13-164.0)	MARC0063610 (13-164.0)
	ALGA0084571* (15-31.9)	ALGA0086432 (15-105.6)	ALGA0086432 (15-105.6)
	ASGA0029597 (6-128.5)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
BW	ALGA0086538 (15-115.1)	MARC0063610 (13-164.0)	MARC0063610 (13-164.0)
	MARC0063610 (13-164.0)	ALGA0050238 (8-146.4)	ALGA0024536 (4-36.8)
	MARC0063610* (13-164.0)	ASGA0051711 (11-77.0)	ASGA0051711 (11-77.0)
SW	ASGA0051711* (11-77.0)	MARC0063610 (13-164.0)	MARC0063610 (13-164.0)
	ALGA0086432* (15-105.6)	ALGA0086432 (15-105.6)	ALGA0086432 (15-105.6)
	ALGA0087078* (15-133.1)	ASGA0070822 (15-136.5)	ASGA0070822 (15-136.5)
PT	ALGA0087273 (4-75.0)	ALGA0087273 (4-75.0)	ALGA0087273 (4-75.0)
	MARC0070351 (5-71.1)	ALGA0056636 (10-7.9)	ALGA0116957 (9-148.6)

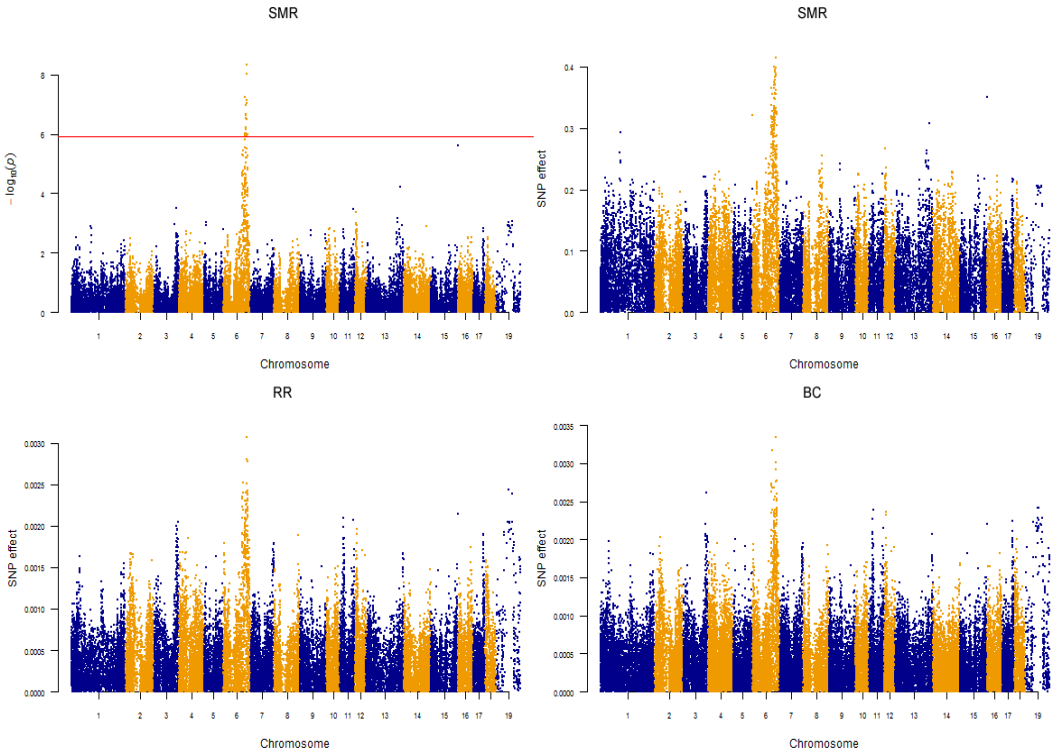

Figure B.1. Manhattan plots for 10wk 10^{th}-rib backfat

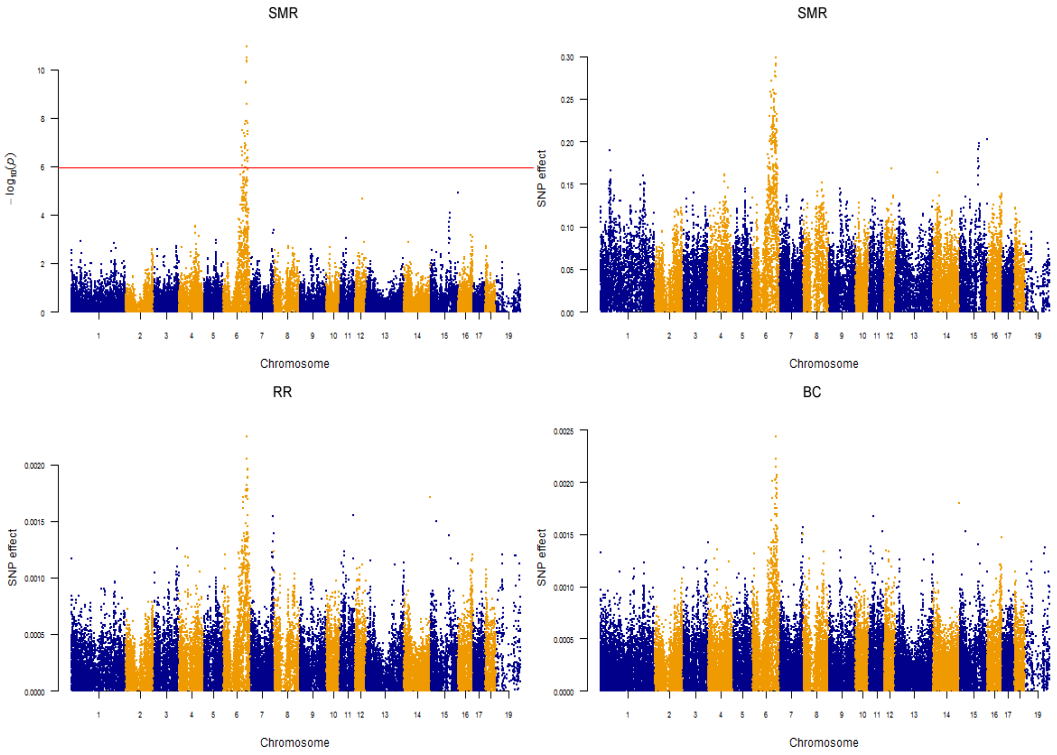

Figure B.2. Manhattan plots for 10wk last-rib backfat

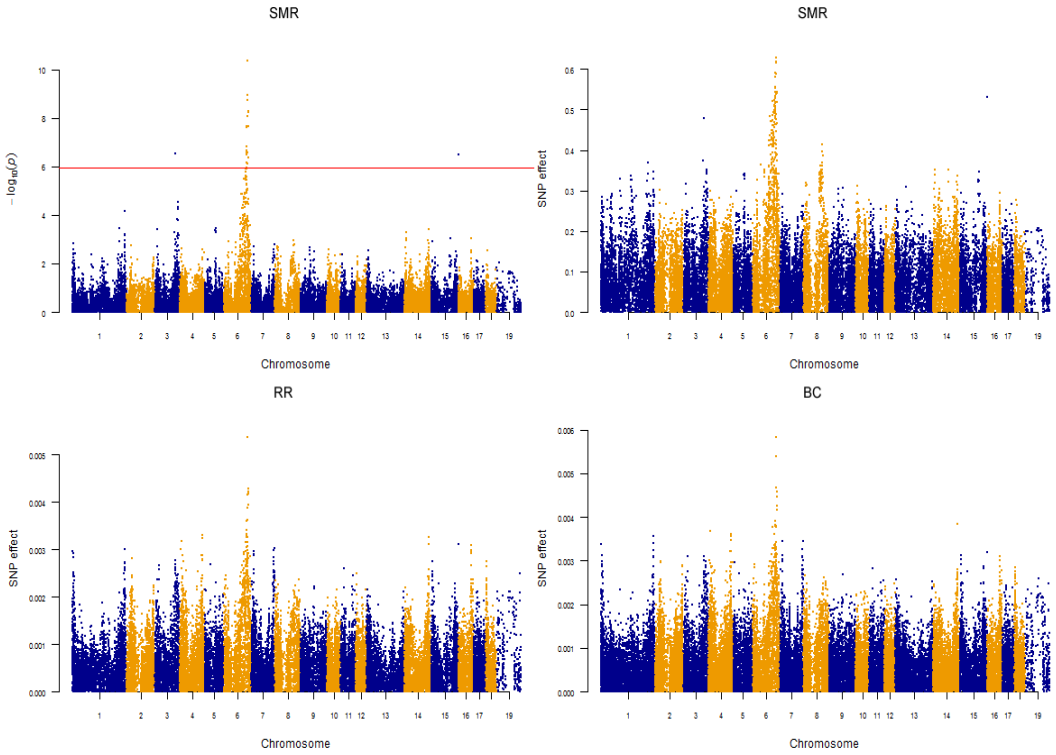

Figure B.3. Manhattan plots for 13wk 10^{th}-rib backfat

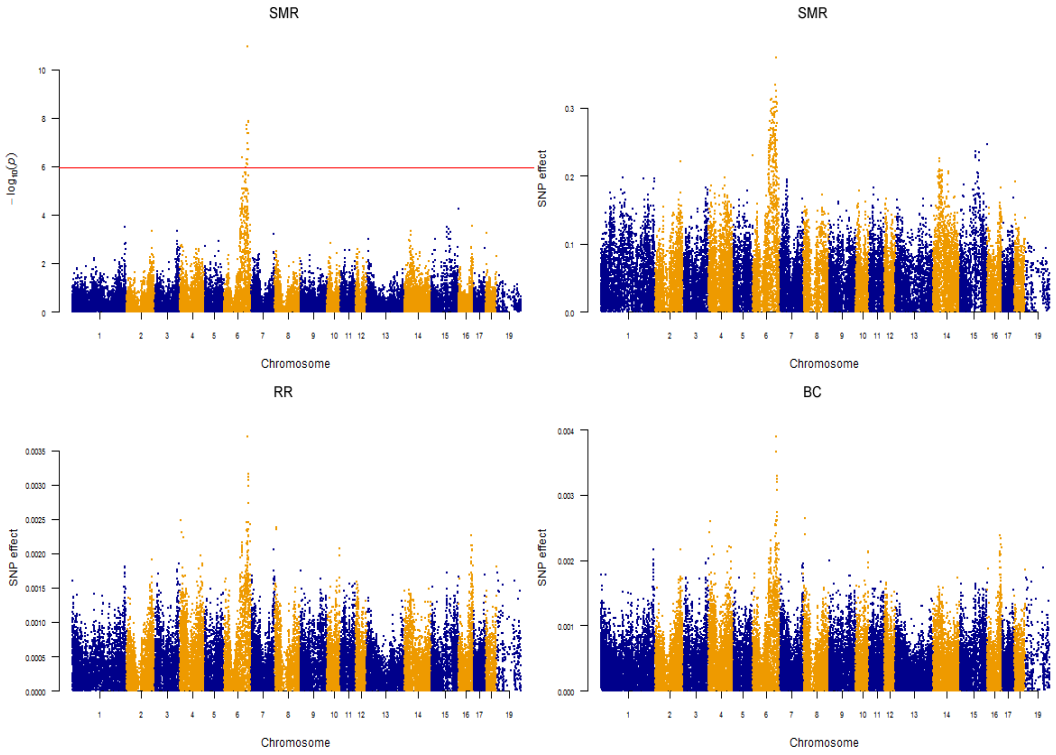

Figure B.4. Manhattan plots for 13wk last-rib backfat

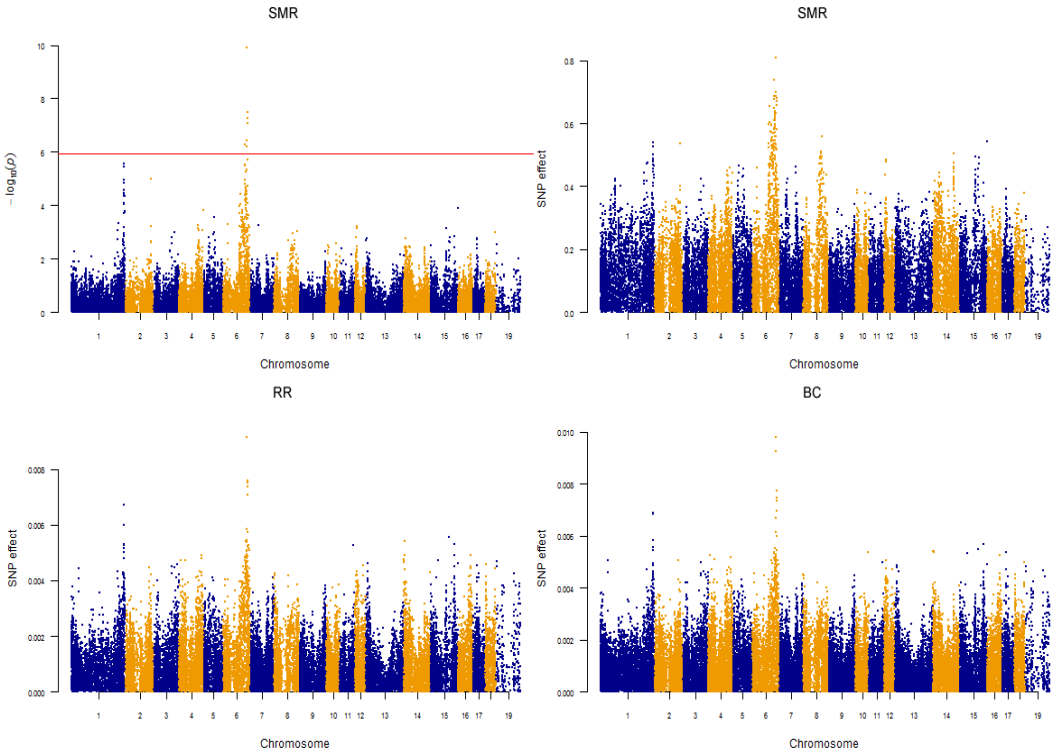

Figure B.5. Manhattan plots for 16wk 10^{th}-rib backfat

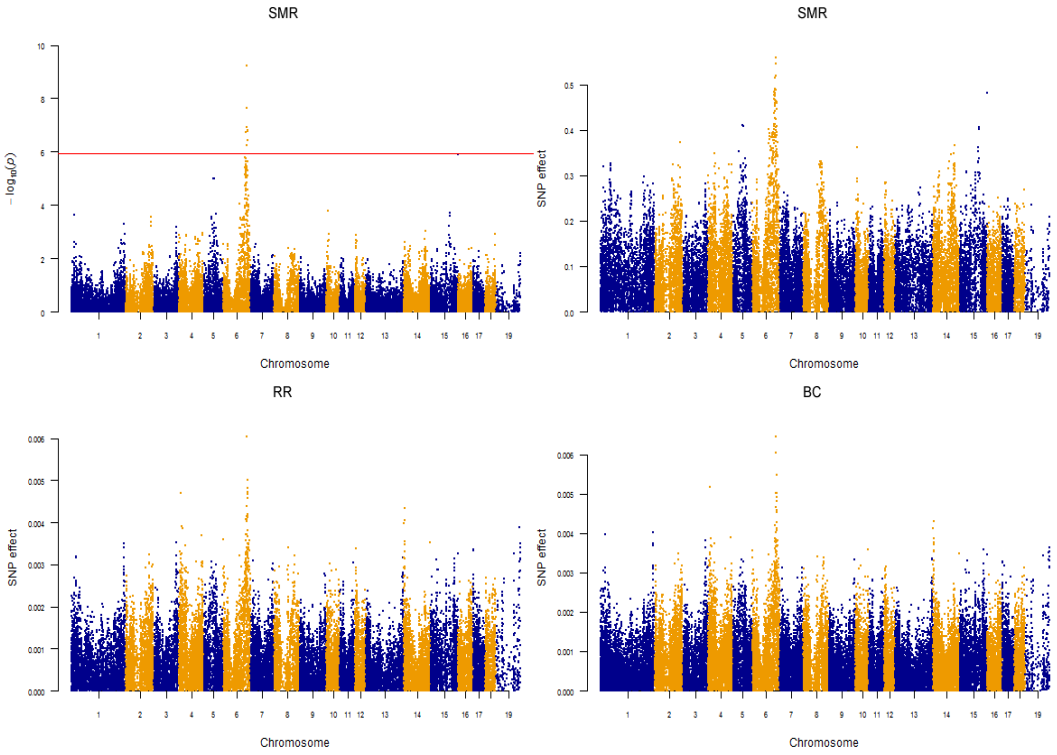

Figure B.6. Manhattan plots for 16wk last-rib backfat

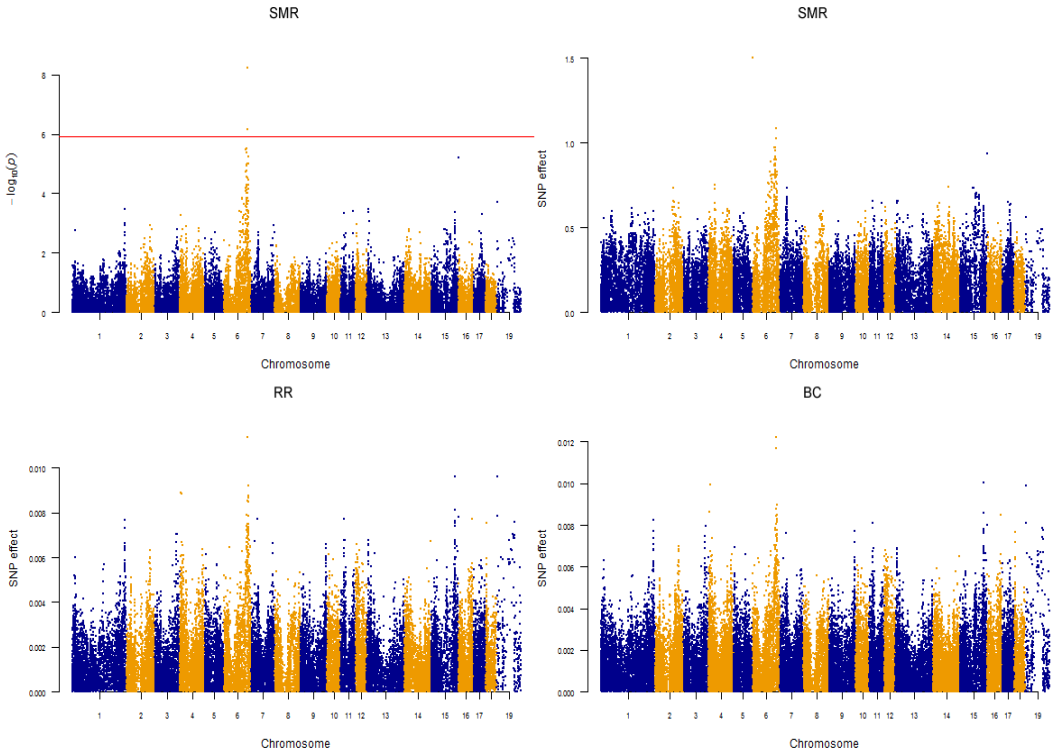

Figure B.7. Manhattan plots for 19wk 10^{th}-rib backfat

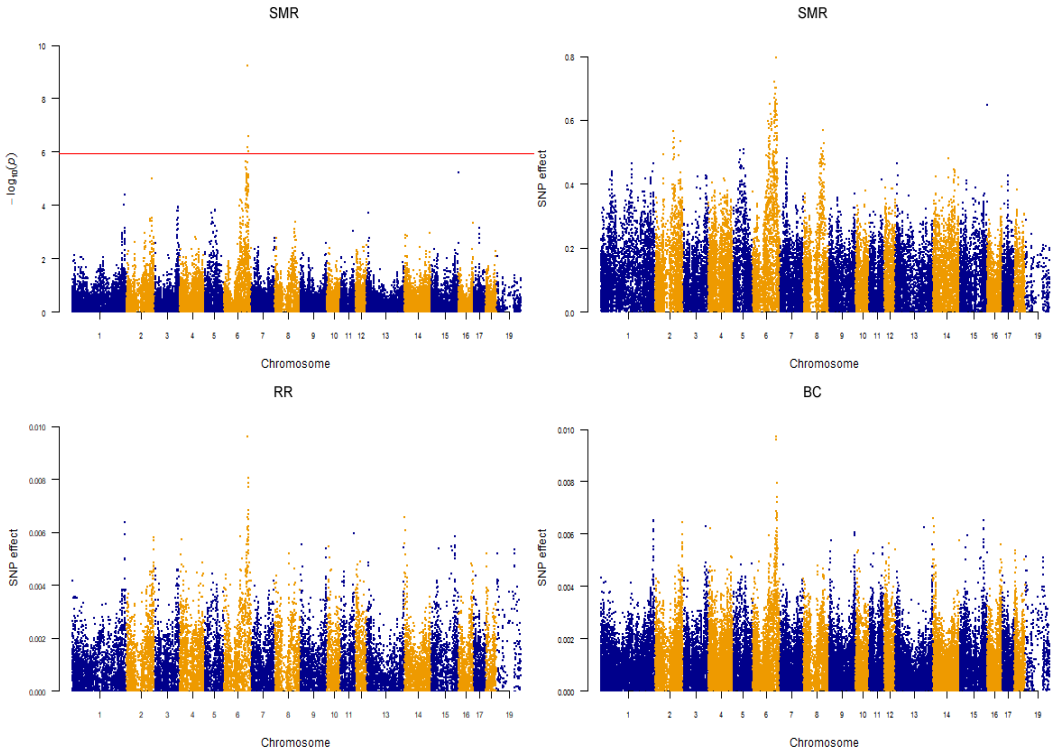

Figure B.8. Manhattan plots for 19wk last-rib backfat

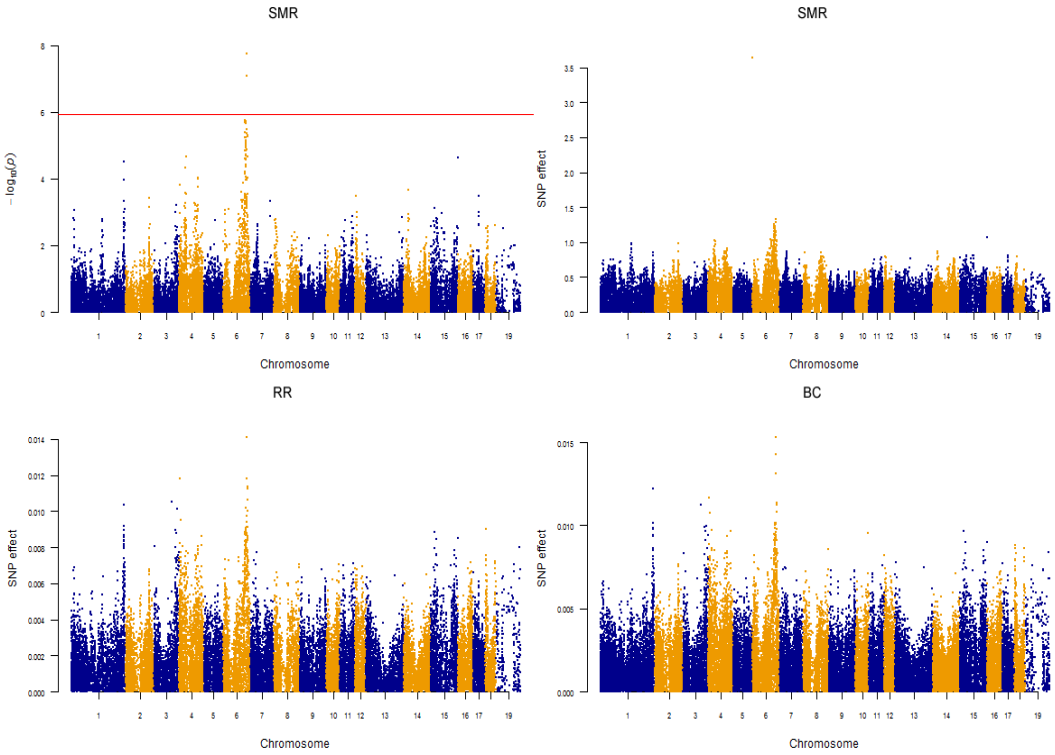

Figure B.9. Manhattan plots for 22wk 10^{th}-rib backfat

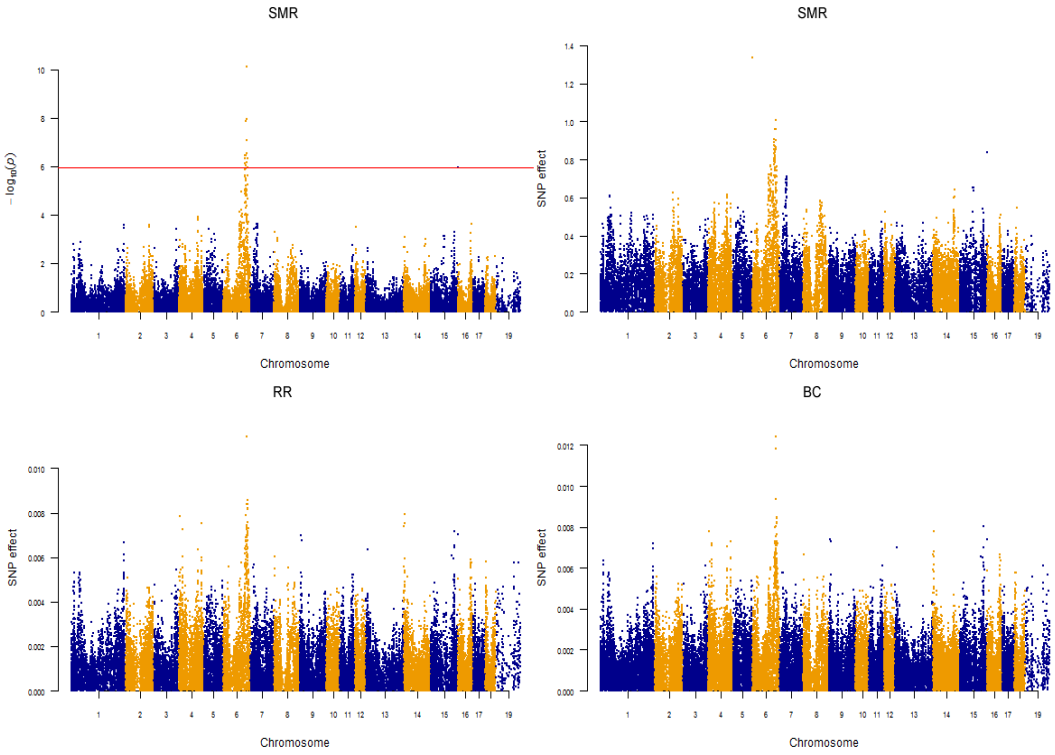

Figure B.10. Manhattan plots for 22wk last-rib backfat

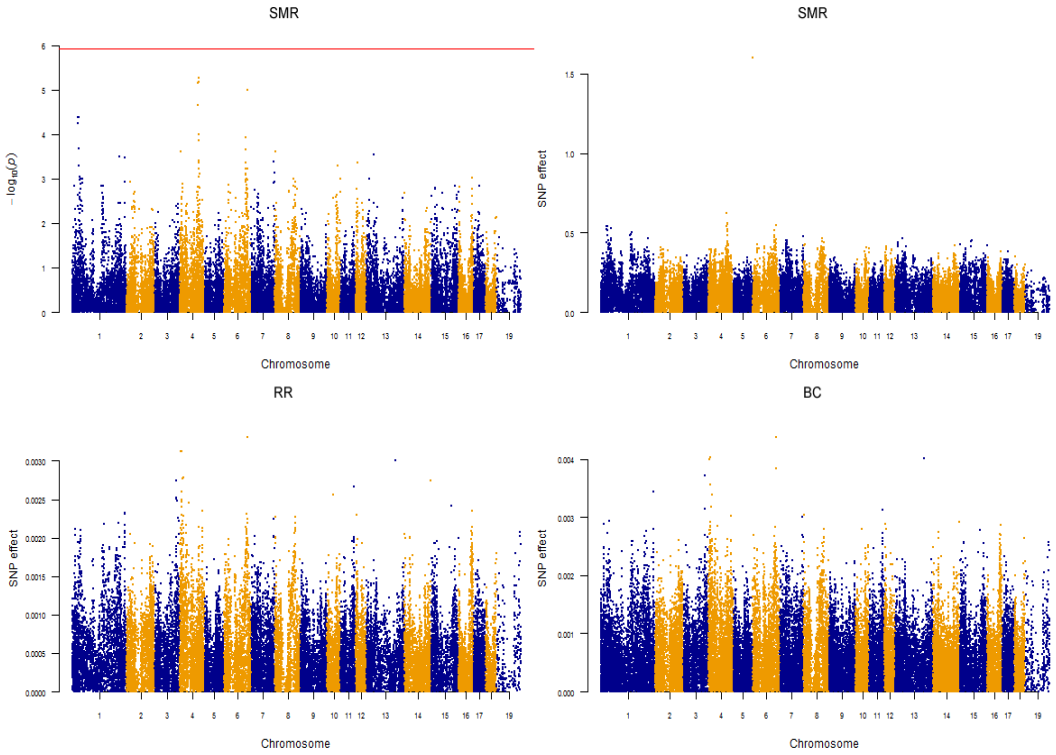

Figure B.11. Manhattan plots for 22wk total body fat tissue

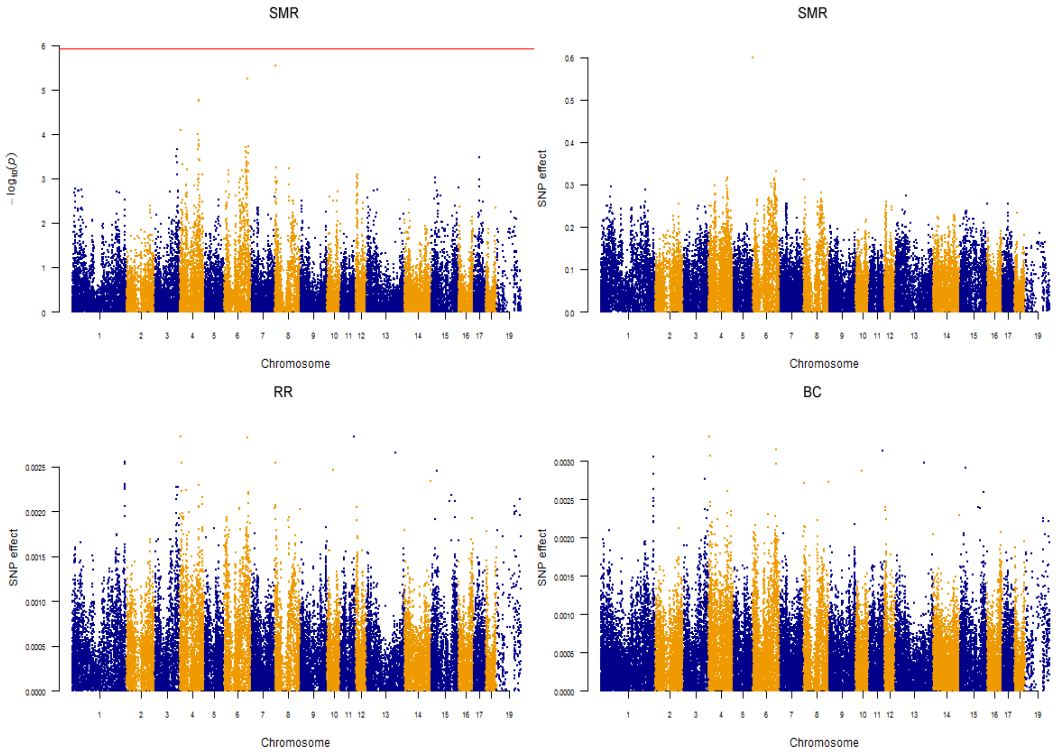

Figure B.12. Manhattan plots for 22wk empty body protein

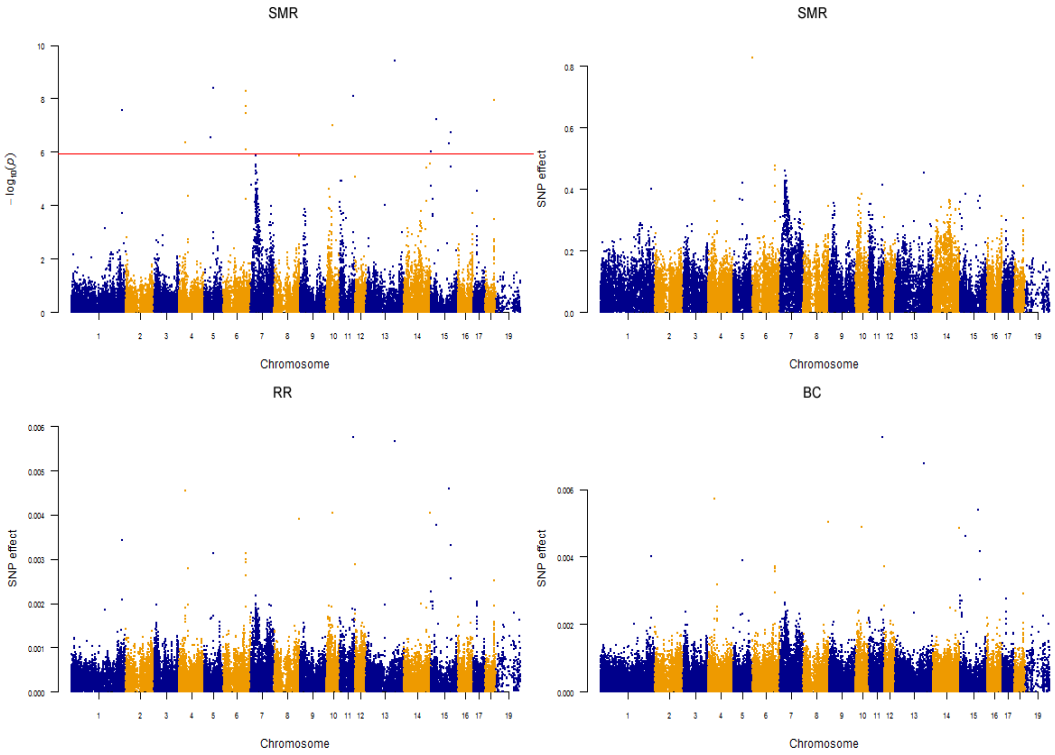

Figure B.13. Manhattan plots for dressing percent

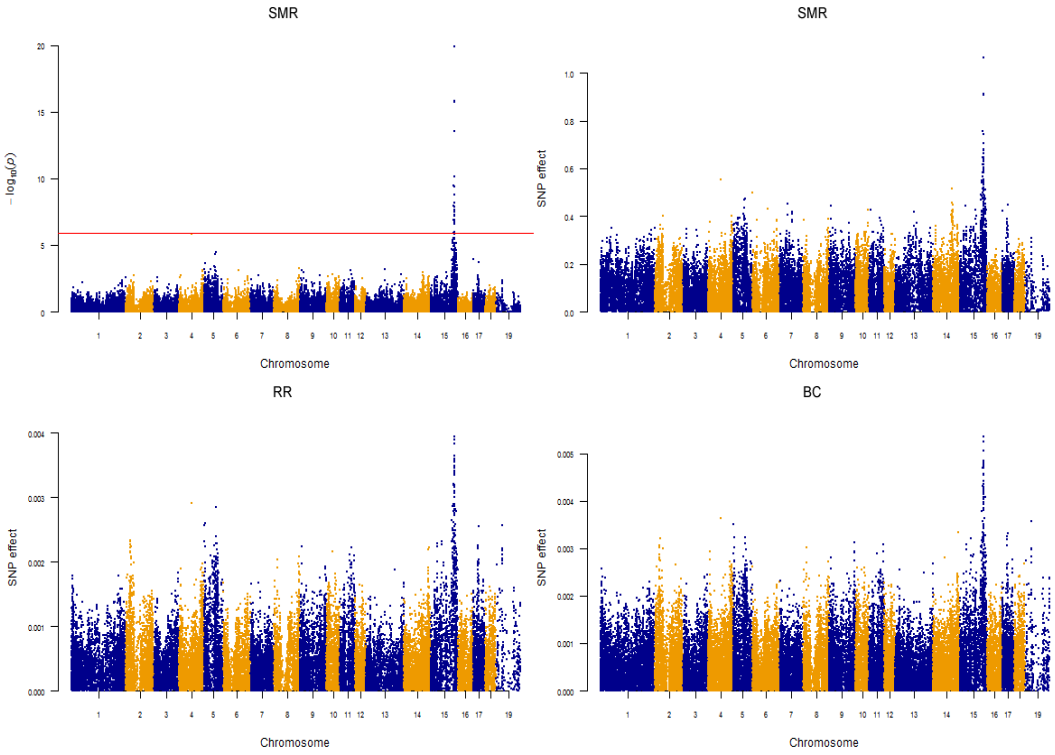

Figure B.14. Manhattan plots for cook yield

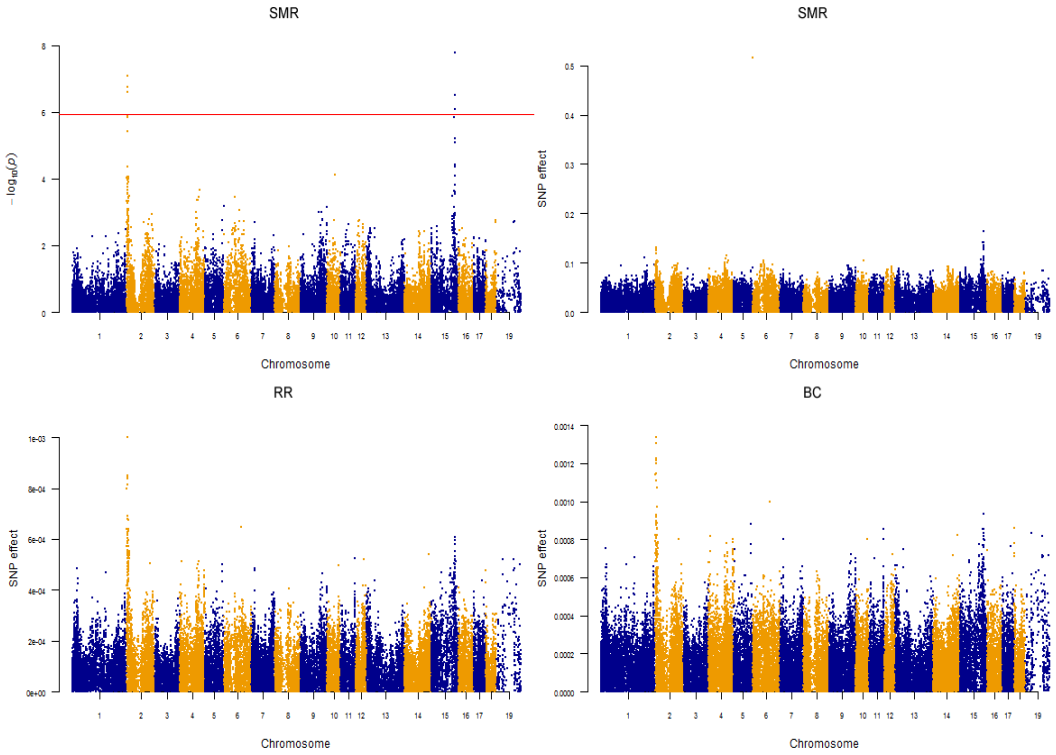

Figure B.15. Manhattan plots for Warner-Bratzler shear force

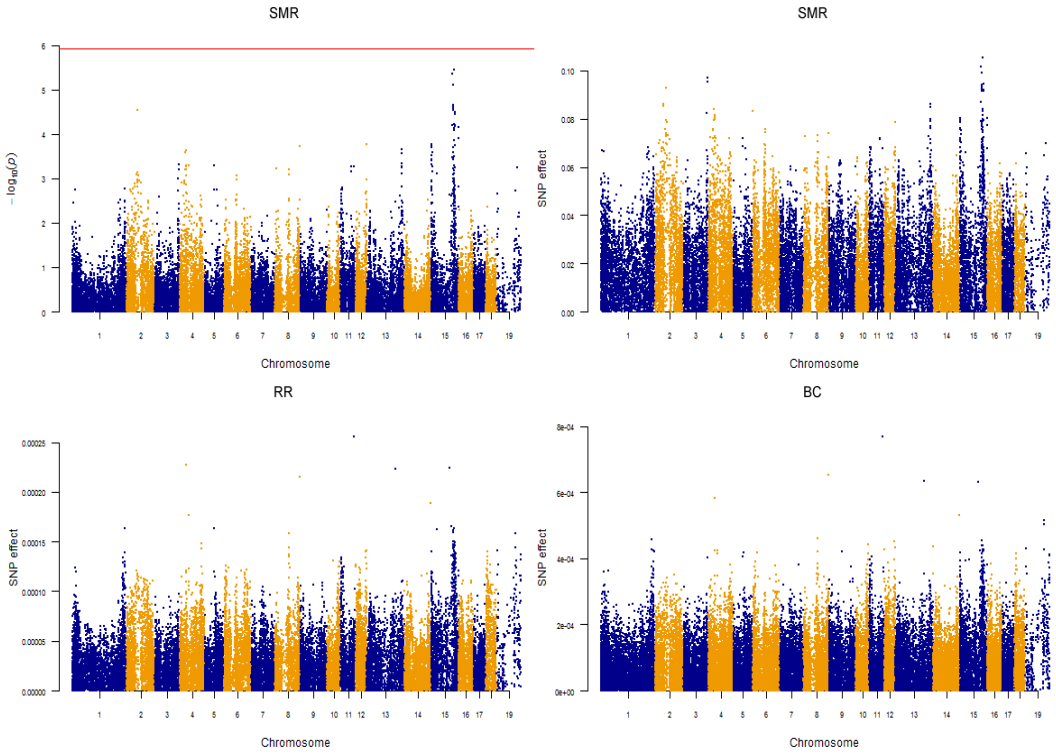

Figure B.16. Manhattan plots for juiciness

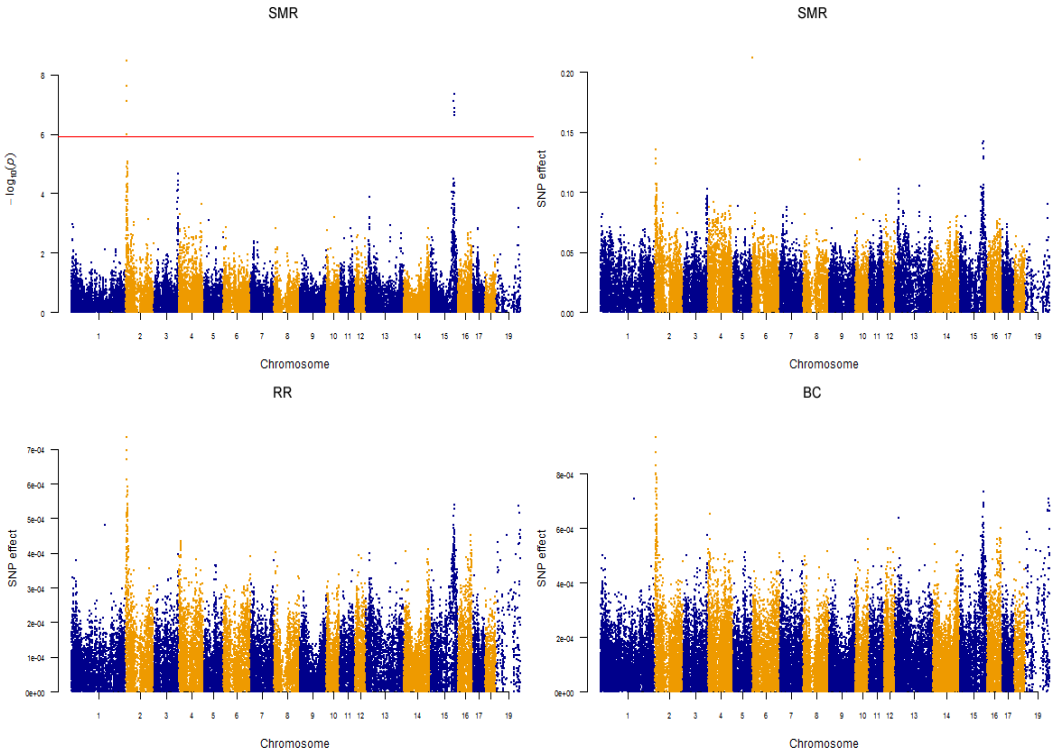

Figure B.17. Manhattan plots for tenderness

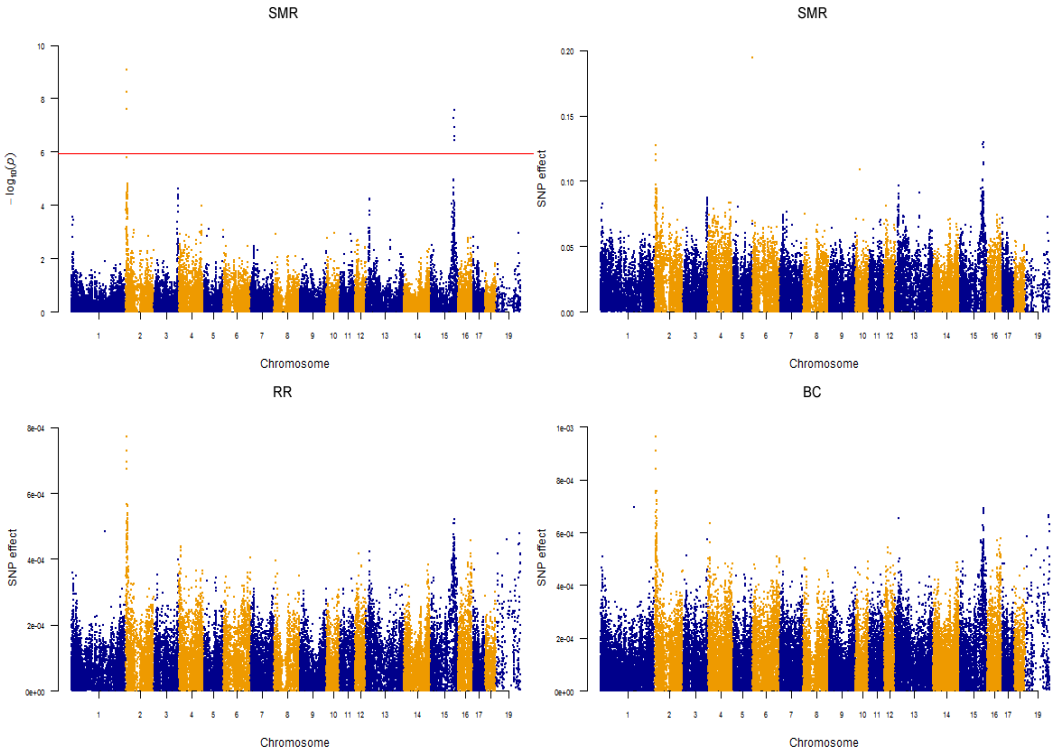

Figure B.18. Manhattan plots for overall tenderness

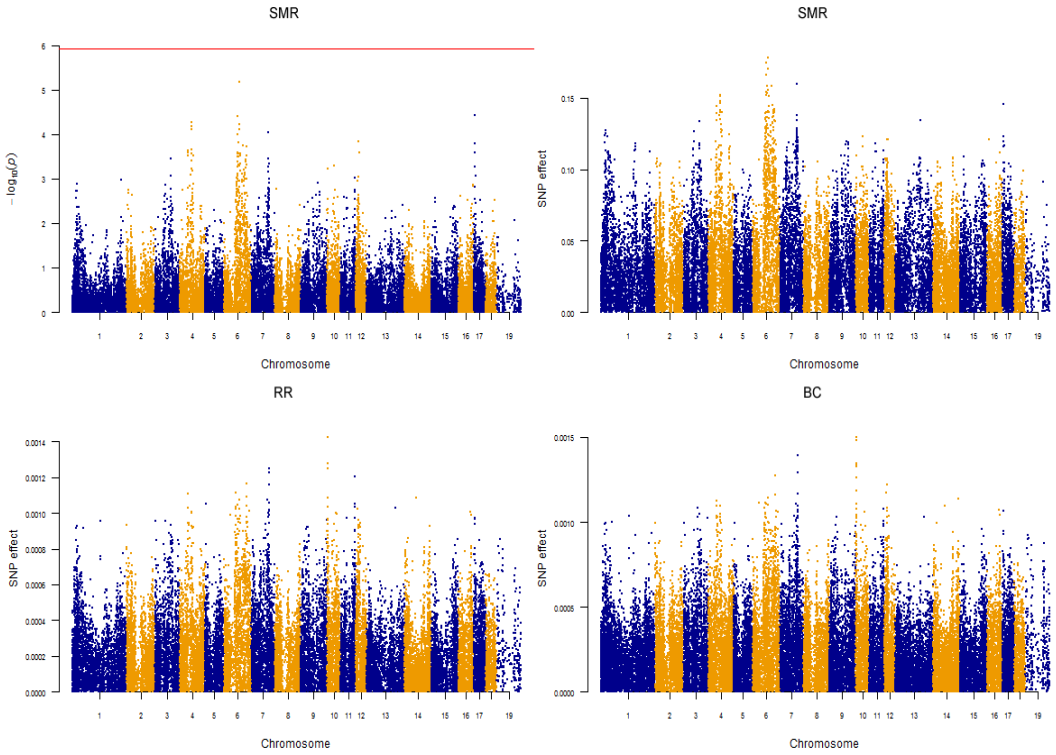

Figure B.19. Manhattan plots for marbling

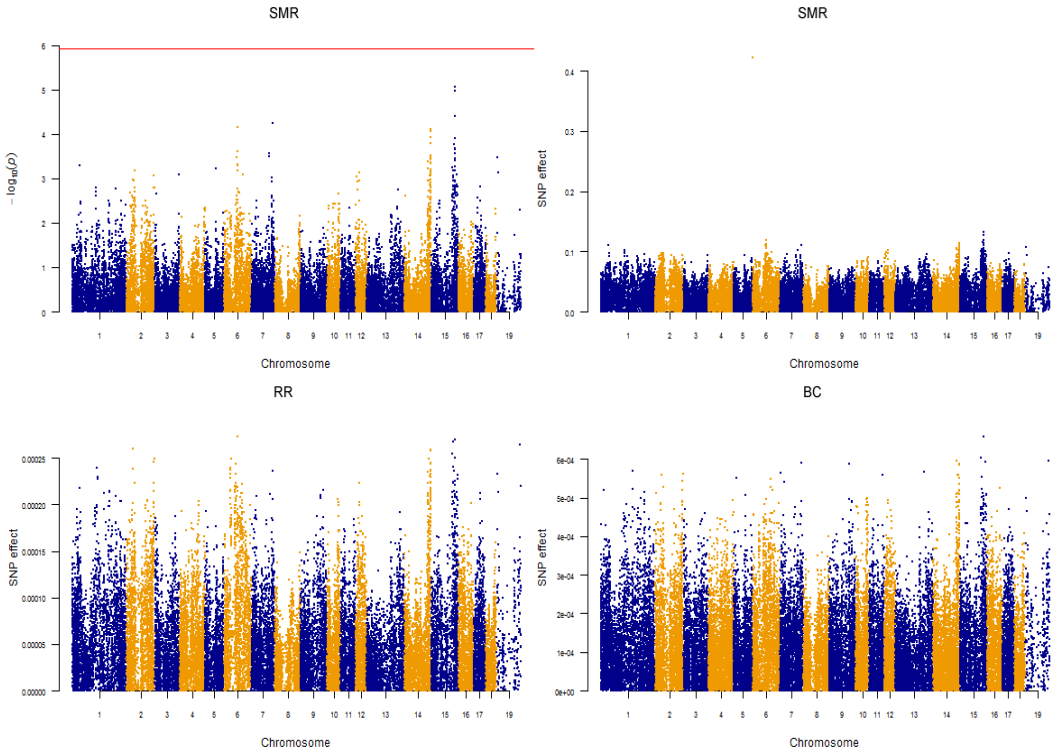

Figure B.20. Manhattan plots for firmness

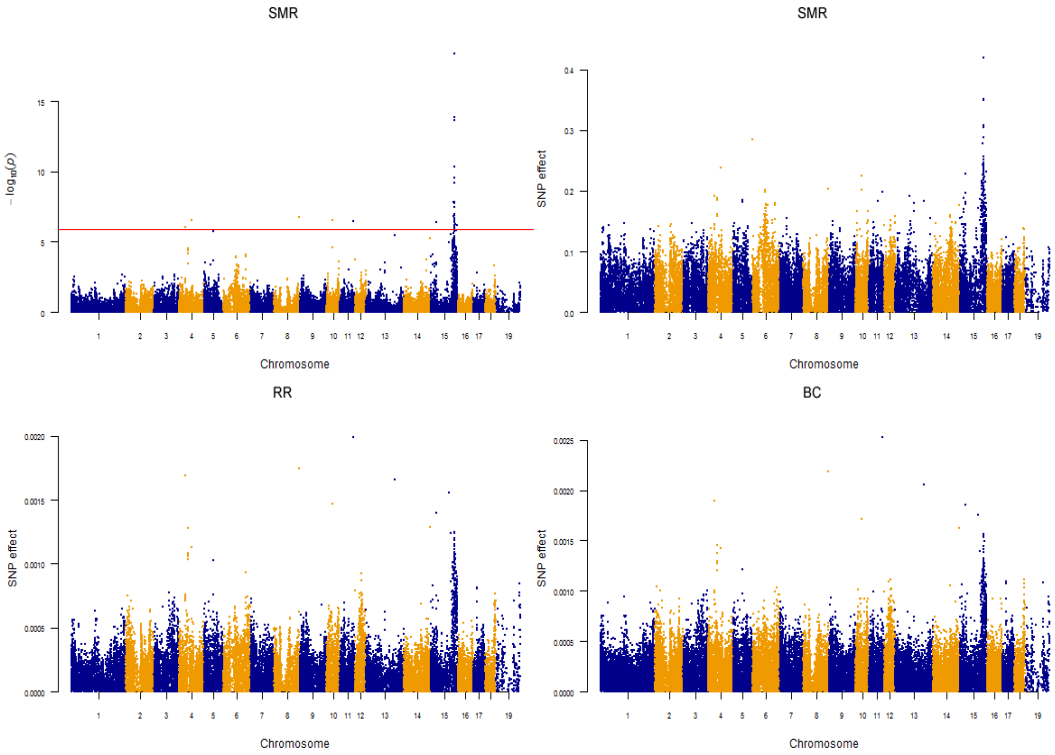

Figure B.21. Manhattan plots for drip loss

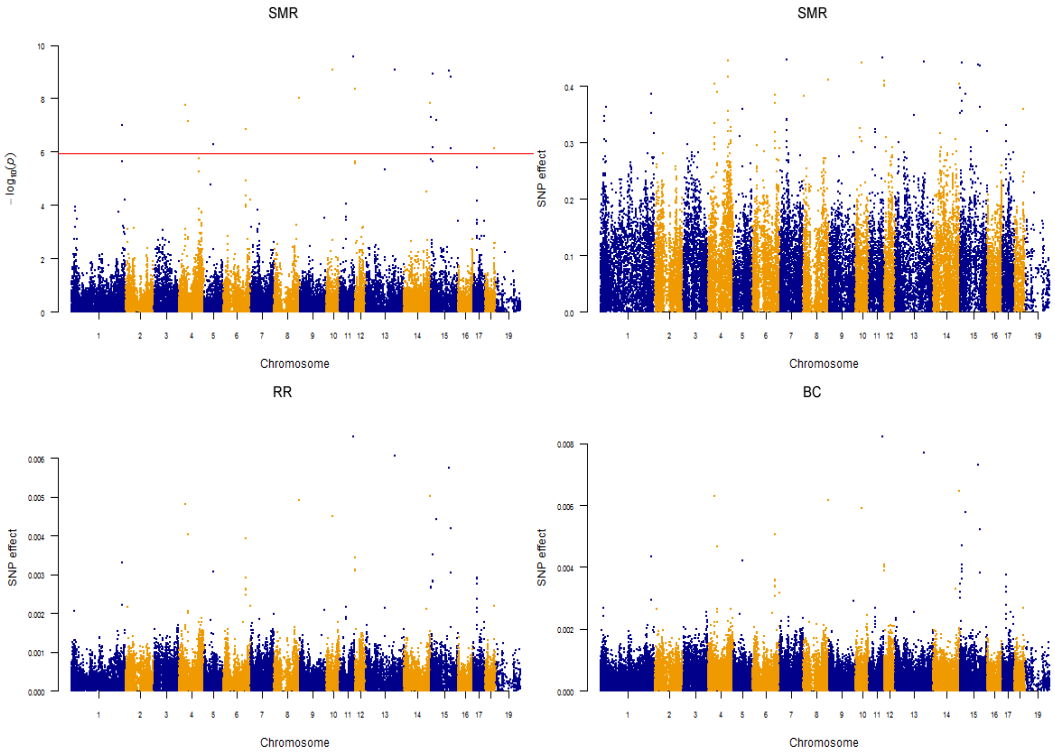

Figure B.22. Manhattan plots for 45min carcass temperature (°C)

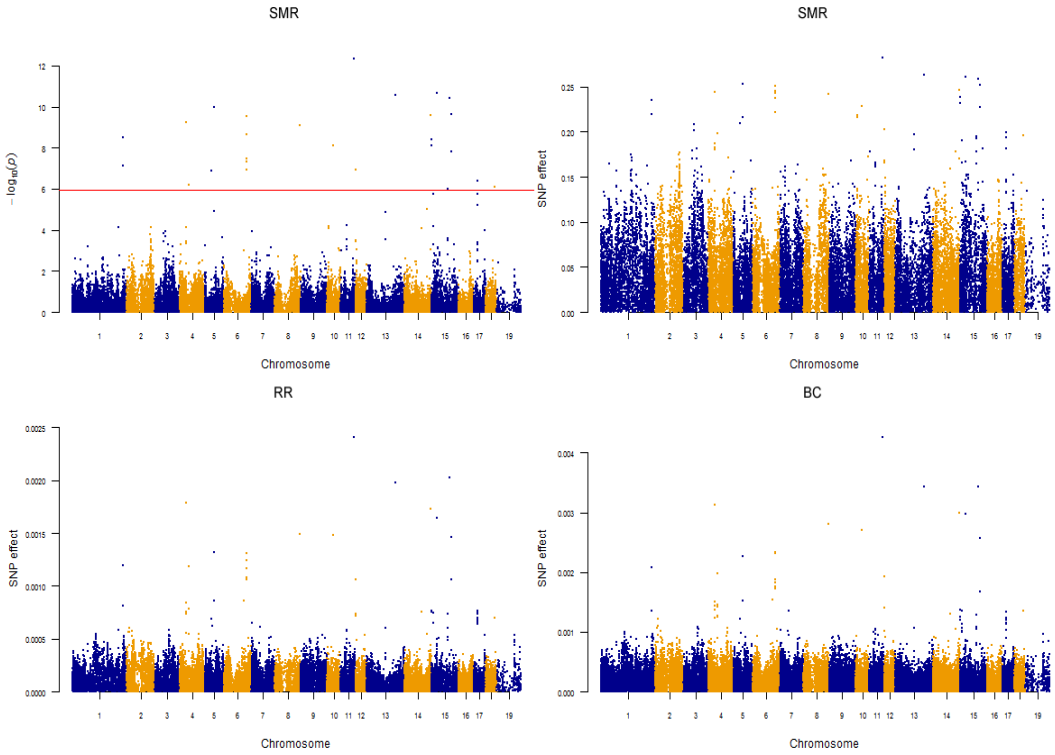

Figure B.23. Manhattan plots for 24h carcass temperature (^{a}C)

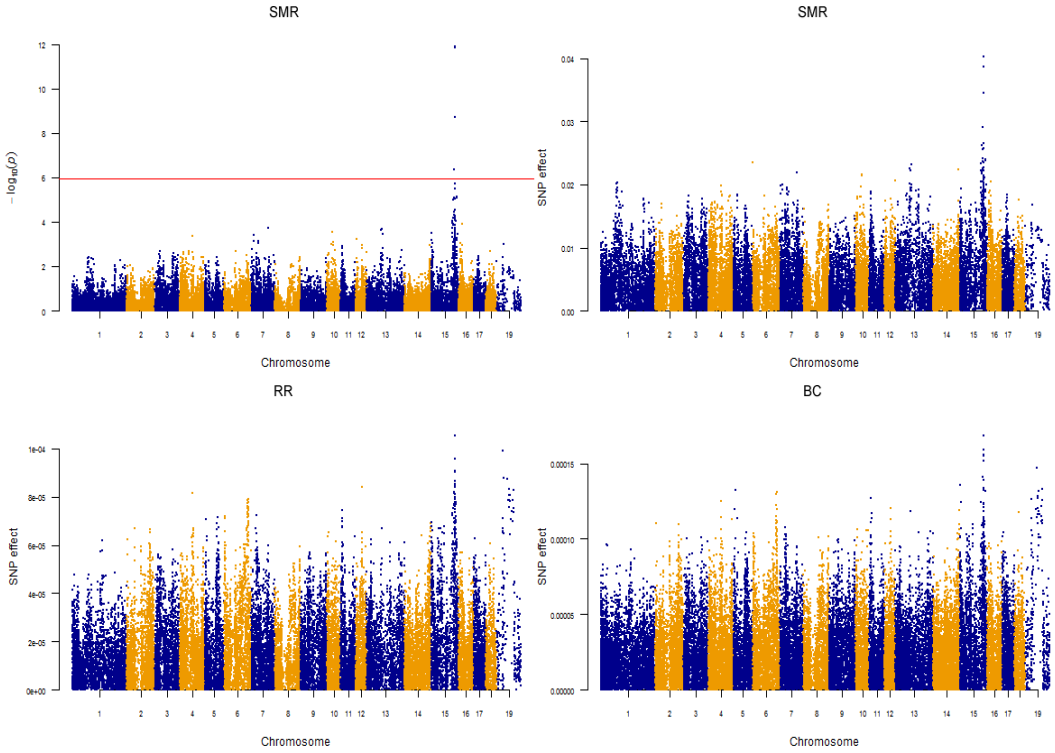

Figure B.24. Manhattan plots for 24h pH

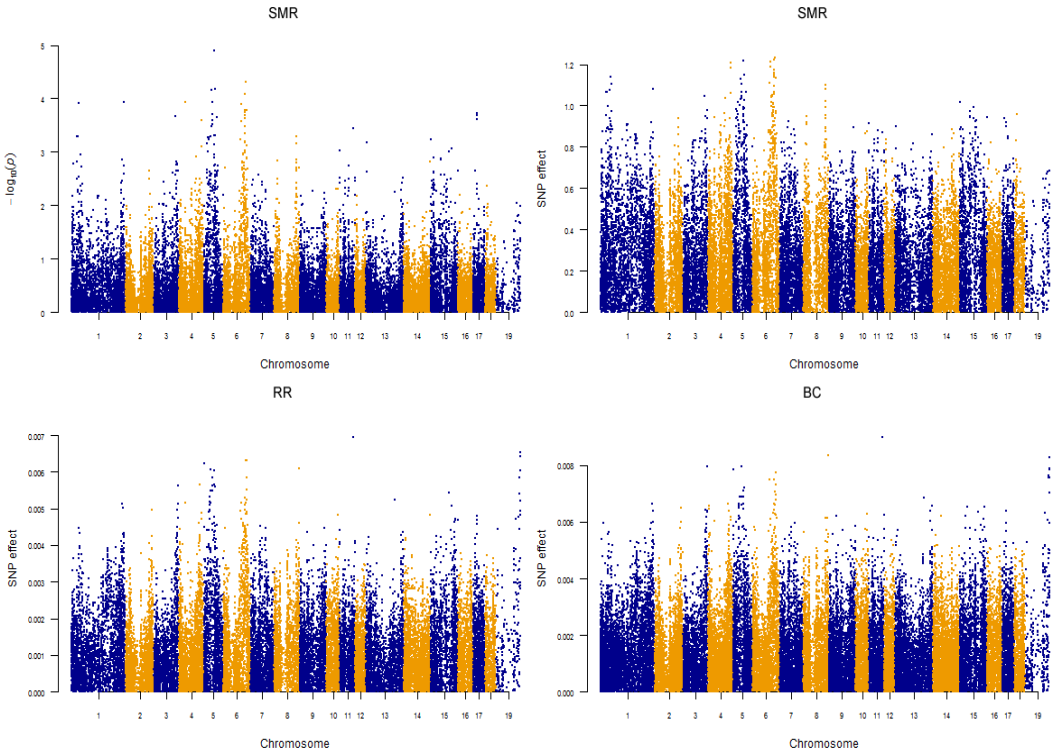

Figure B.25. Manhattan plots for first-rib backfat

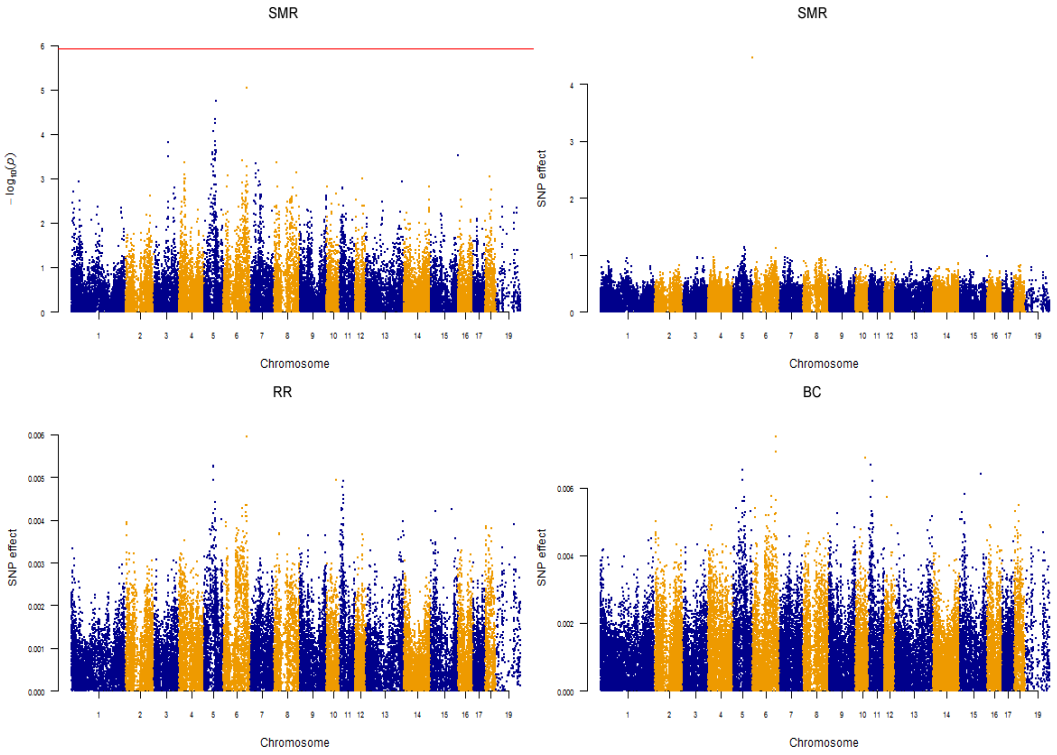

Figure B.26. Manhattan plots for last-rib backfat

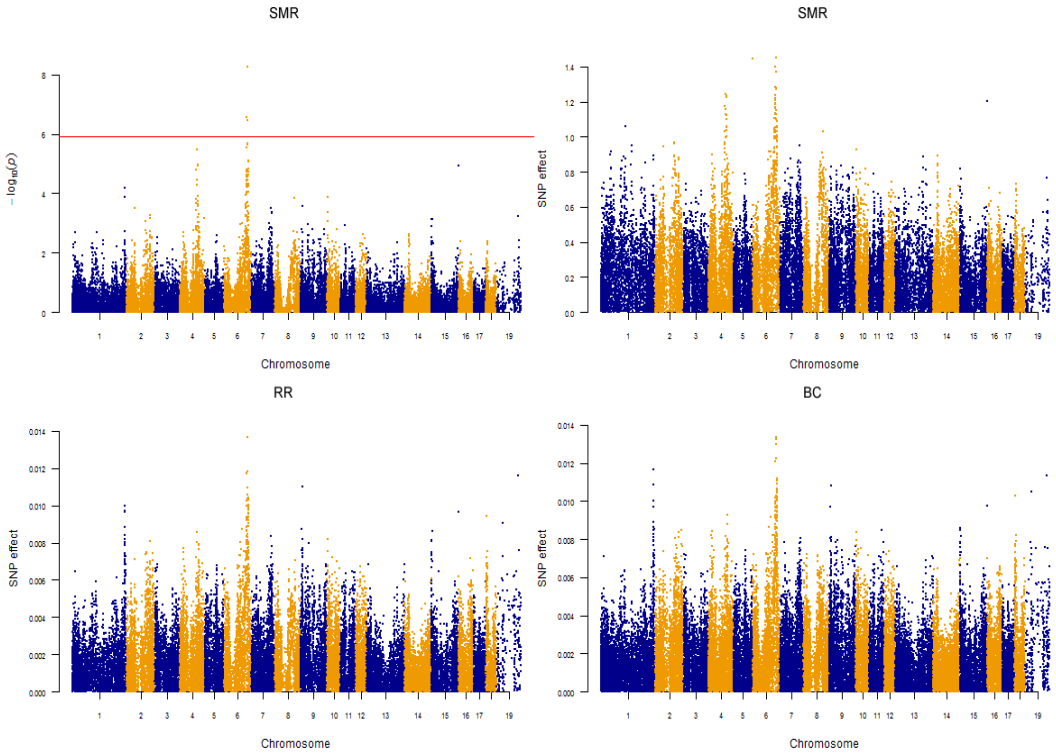

Figure B.27. Manhattan plots for last-lumbar vertebra backfat

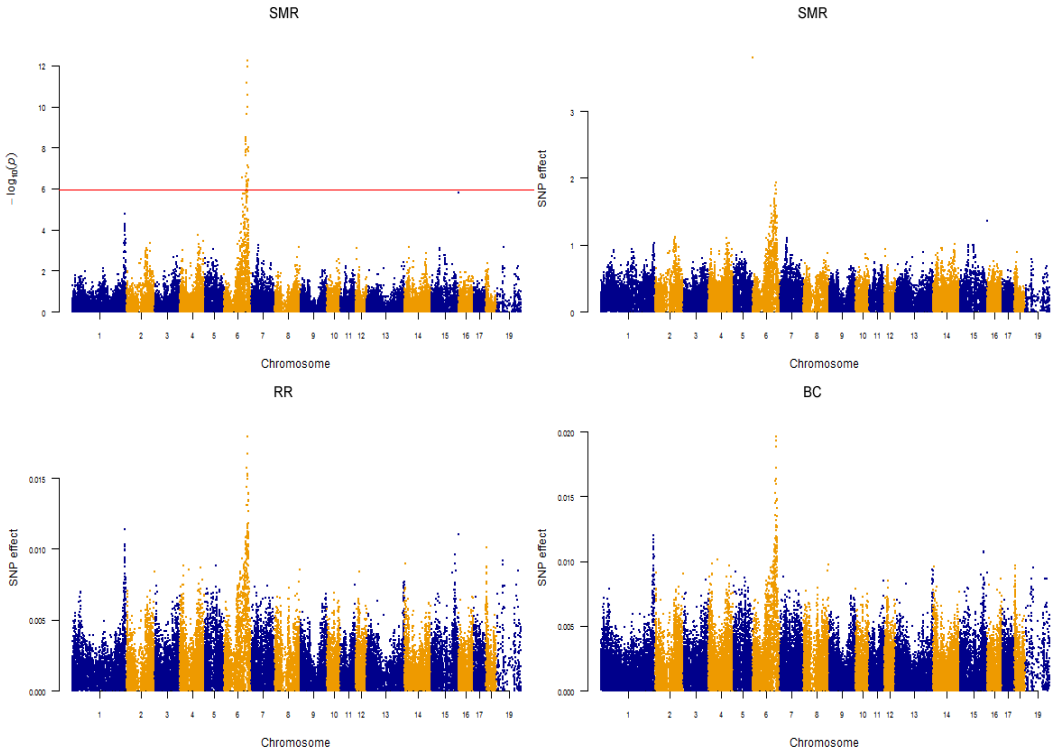

Figure B.28. Manhattan plots for 10^{th}-rib backfat

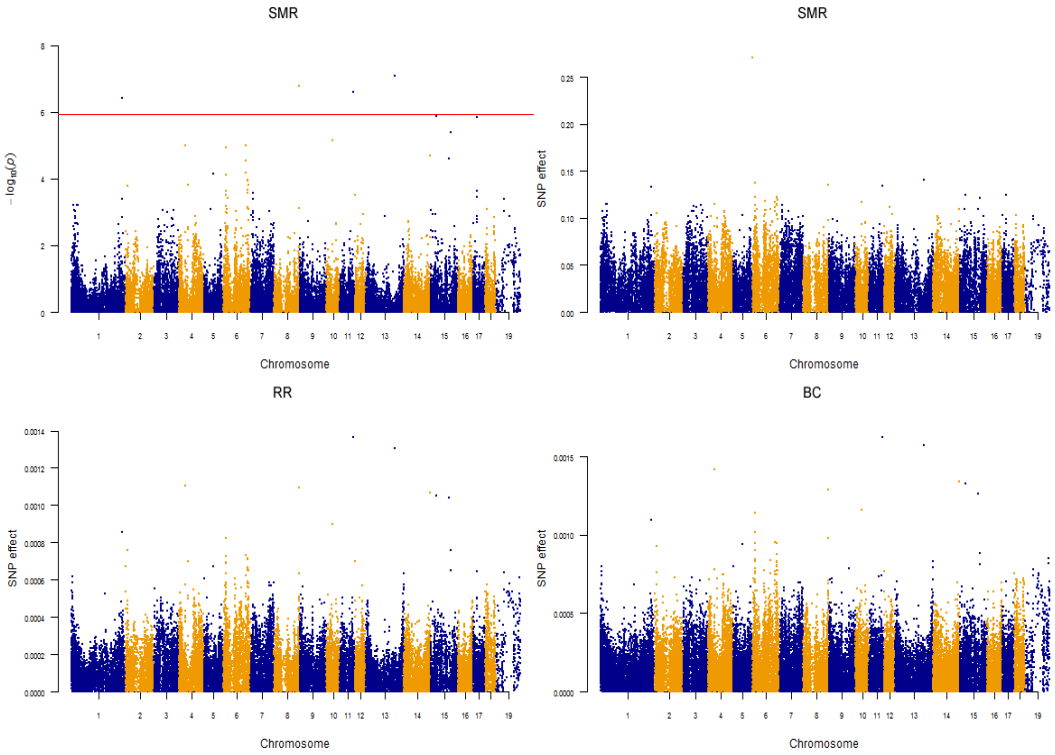

Figure B.29. Manhattan plots for ham weight

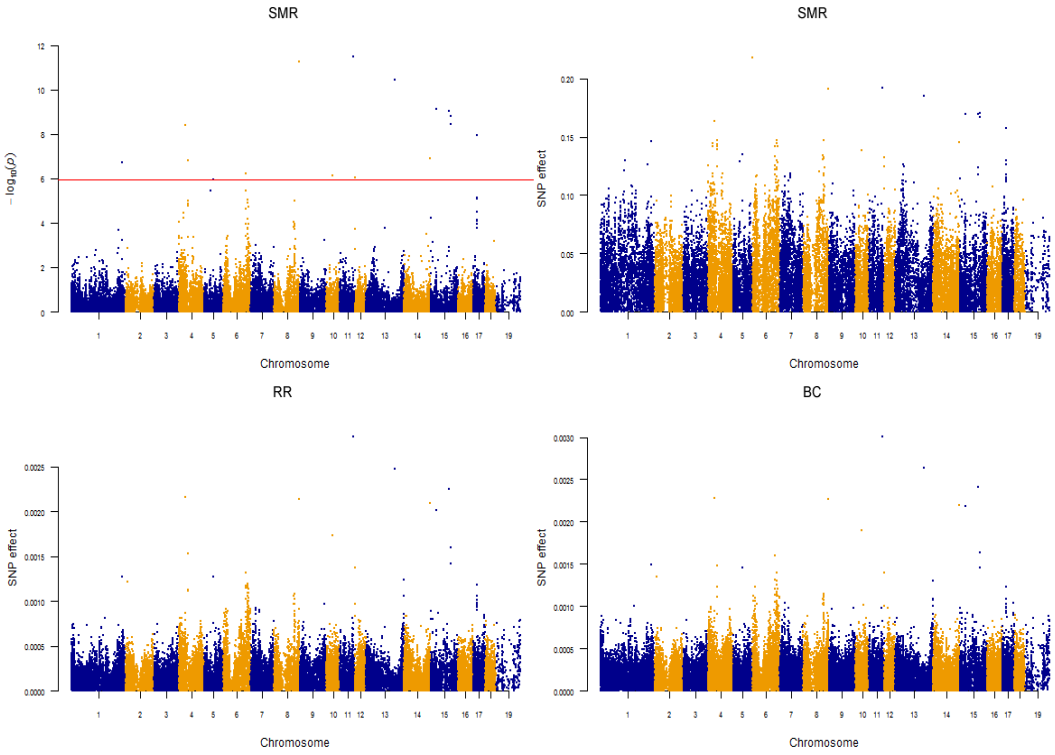

Figure B.30. Manhattan plots for loin weight

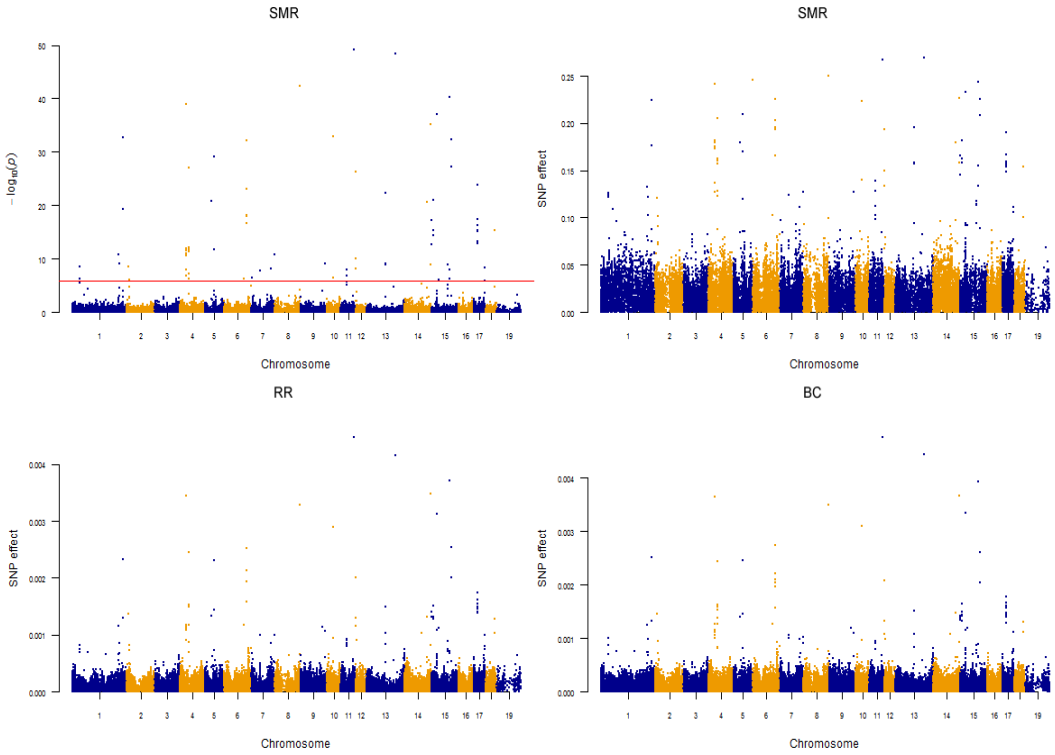

Figure B.31. Manhattan plots for boston shoulder weight

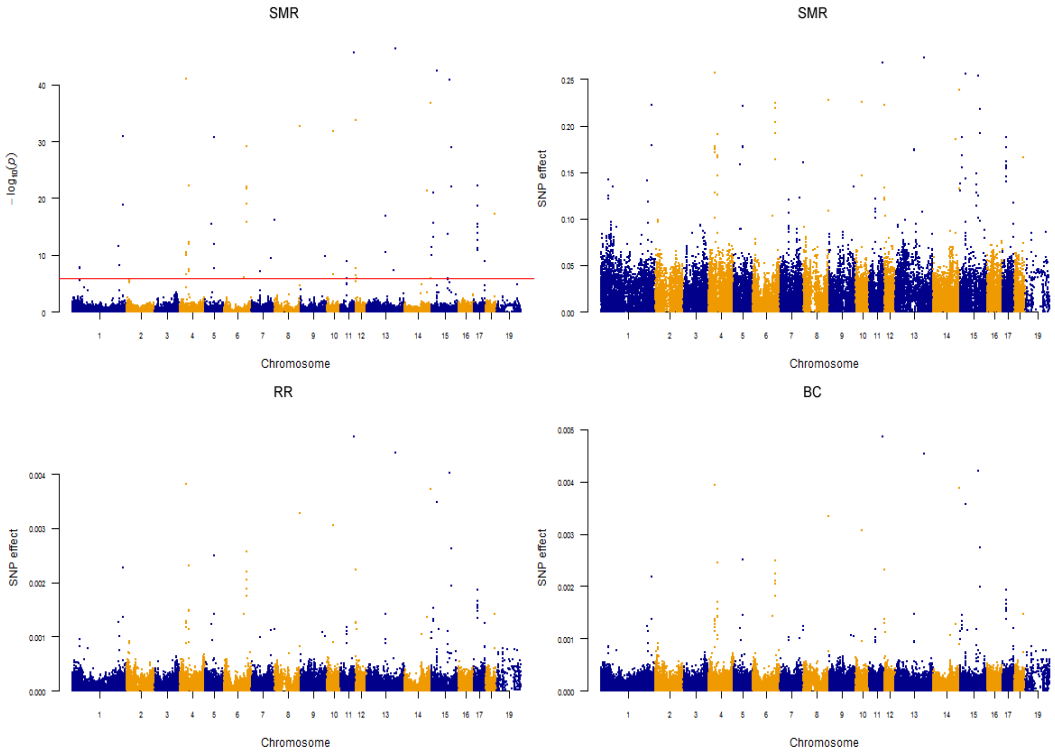

Figure B.32. Manhattan plots for picnic shoulder weight

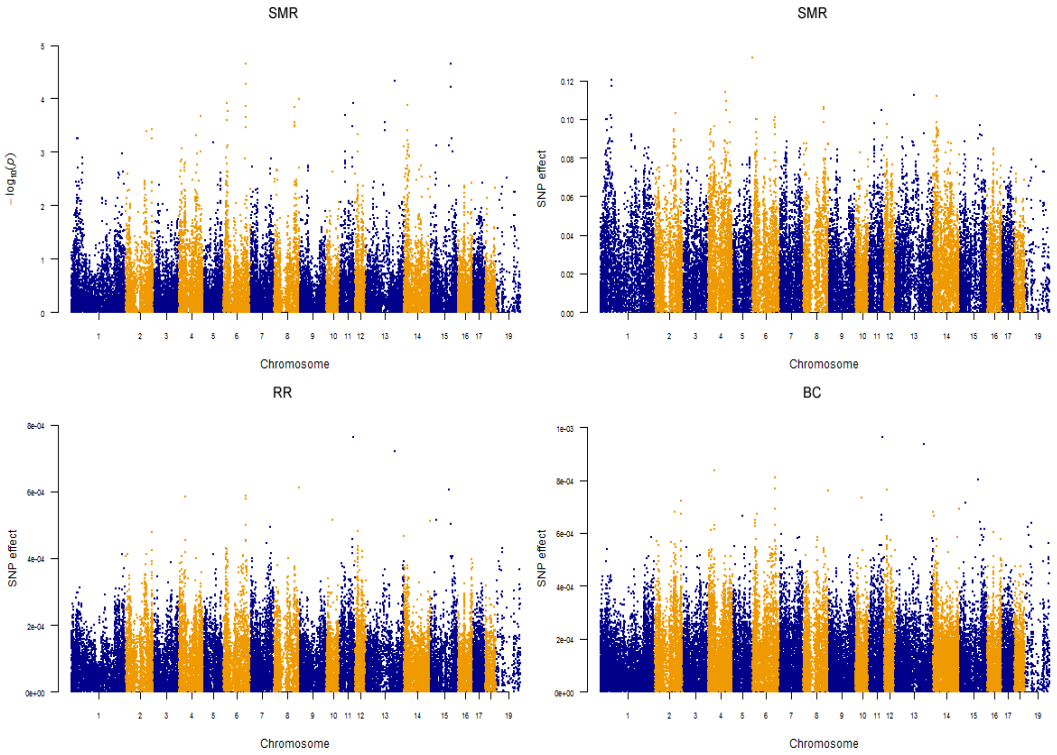

Figure B.33. Manhattan plots for belly weight

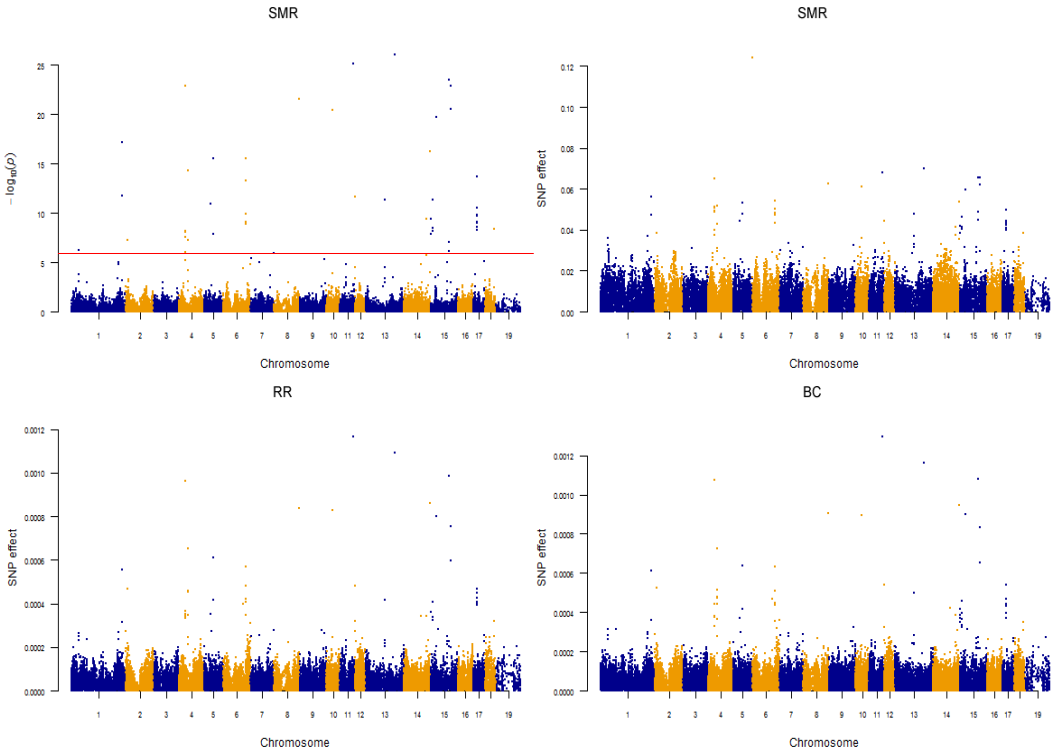

Figure B.34. Manhattan plots for spareribs weight

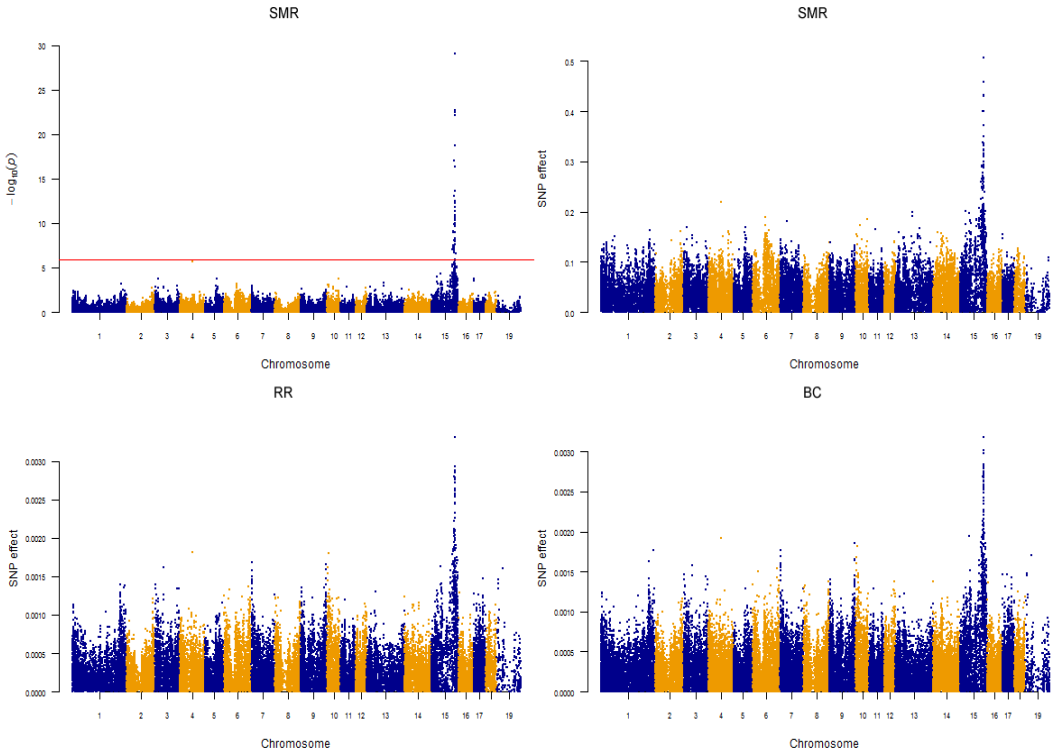

Figure B.35. Manhattan plots for protein (%)

R Code of the analysis

Single-marker regression (SMR)

```
setwd("C:/Users/Badger/Desktop/TESE/GWAS analysis")
load("thesis.RData")
library("lattice")
library("LDcorSV")
library("car")
library("qqman")

# Compute G Matrix (Genomic)
X <- scale(t(Markers), center = TRUE, scale = TRUE)
G <- tcrossprod(X)/ncol(X)
G <- data.frame(var=paste(1:nrow(phen), phen$Animal, sep=" "), G)
colnames(G) <- c("var", paste(1:nrow(phen), phen$Animal, sep=" "))

geno <- Recode(Markers, "0='AA';1='AB';2='BB'")
geno <- data.frame(geno)
geno <- matrix(unlist(strsplit(do.call(paste, c(geno, sep="")), "")),
nrow = nrow(geno), byrow = TRUE) # separar em duas colunas cada marcador

project.tfam <- data.frame(FamilyID=1:nrow(phen),
IndividualID=phen$Animal,
FatherID=0,
MotherID=0,
Sex=0,
Phenotype=-9) # -9 missing
head(project.tfam)

project.tped <- data.frame(chromosome= map$chr,
SNP= map$mrk_id,
distance=0,
position= round(map$pos,digits=0), geno)
project.tped[1:5,1:5]

phenotype <- data.frame(FamilyID= 1:nrow(phen),
Animal= phen$Animal,
BW = phen[,2:68]) # 1:67
phenotype[is.na(phenotype)] = -9
head(phenotype)

covfix <- data.frame(FamilyID=1:nrow(phen),
Animal=phen$Animal,
sex = ifelse(phen$sex=="M",1, 2),
litter = phen$litter)

write.table(phenotype, "phen.txt", row.names=F,col.names=F, quote=F)
```

```
write.table(covfix, "covfix.txt", row.names=F,col.names=F, quote=F)
write.table(project.tfam, "project.tfam", row.names=F,col.names=F, quote=F)
write.table(project.tped, "project.tped", row.names=F,col.names=F, quote=F)
write.table(G, "G.txt", row.names=F, col.names=T, quote=F, sep="\t")

# Running GWAS analyses
for(i in 1:67){
trait <- paste0("FastLmmC.exe -mpheno ",i," -tfile project -sim G.txt -MaxChromosomeValue
19 -pheno phen.txt -simLearnType Full -covar covfix.txt -out gwas.txt")
analysis <- capture.output(system(trait, intern=T))

gwas <- read.table("gwas.txt", header=T)
gwas$SNPeff <- abs(gwas$SNPWeight)

trellis.device(device="png",
filename=paste0("QQPlot-GWAS",i,".png"),
width=617,
height=397)
qq(gwas$Pvalue)
dev.off()

trellis.device(device="png",
filename=paste0("Pvalue-GWAS",i,".png"),
width=617,
height=397)
manhattan(gwas, chr="Chromosome", bp="Position", p="Pvalue", snp="SNP",
cex=0.5, cex.axis=0.7, col=c("blue4","orange2"),
suggestiveline=F, genomewideline=F, logp=T)
dev.off()

trellis.device(device="png",
filename=paste0("SNPeff-GWAS",i,".png"),
width=617,
height=397)
manhattan(gwas, chr="Chromosome", bp="Position", p="SNPeff", snp="SNP",
cex=0.5, cex.axis=0.7, col=c("blue4","orange2"),
ylim=c(0, max(gwas$SNPeff)+0.05*max(gwas$SNPeff)),
ylab="SNP effect",
suggestiveline=F, genomewideline=F, logp=F)
dev.off()
cat("GWAS n.", i, "\n")}
```

Ridge Regression BLUP (RR)

```
setwd("C:/Users/Badger/Desktop/RRBLUP")
load("thesis.RData")
library("rrBLUP")
```

```
library("car")
library("qqman")
library("pastecs")
library("lattice")

#Descriptive Analysis
stat <- na.omit(data.frame(t(stat.desc(phen))))
Markers <- Recode(Markers,"0=1;1=0;2=-1")
Markers <- t(Markers)
SNPeff <- matrix(NA,ncol=67,nrow=ncol(Markers))

for (i in 1:67) {
#predict marker effects
y <- na.omit(phen[,i+1])
sex <- phen[!is.na(phen[,i+1]), "sex"]
litter <- phen[!is.na(phen[,i+1]), "litter"]
fixef <- model.matrix(y ~ factor(sex) + factor(litter))
snps <- Markers[!is.na(phen[,i+1]), ]

ans <- mixed.solve(y=y,X=fixef, Z=snps )
SNPeff[,i] <- ans$u
gwas <- data.frame(map,SNPeff=ans$u)
trellis.device(device="png",
filename=paste0("SNPeff-rrBLUP",i,".png"),
width=617,
height=397)
manhattan(gwas, chr="chr", bp="pos", p="SNPeff", snp="mrk_id",
cex=0.5, cex.axis=0.7, col=c("blue4","orange2"),
ylim=c(0, max(gwas$SNPeff)+0.05*max(gwas$SNPeff)),
ylab="SNP effect",
suggestiveline=F, genomewideline=F, logp=F)
dev.off()
cat("rrBLUP n.", i, "\n")}
save(SNPeff, file="SNPeff_rrBLUP.RData")
```

Bayes Cπ (BC)

```
setwd("C:/Users/Badger/Desktop/TESE/BAYESC analysis")
library("methods")
library("BGLR")
library("lattice")
load("thesis.RData")

## MCMC Specifications
nIter=60000
burnIn=20000
thin=20
```

```
## Bayes C
M <- t(Markers)
traitName <- c(as.matrix(read.table("id.txt", head=F, colClass="character")))
ETA=list(list(~factor(sex)+factor(litter), data=phen, model='FIXED'),
list(X=M, model='BayesC'))
fitBL=BGLR(phen[, traitName],         ## phenotypic vector
ETA=ETA,                    ## Model prioris
nIter=nIter,                ## iterations
burnIn=burnIn,              ## BurnIn
thin=thin,                  ## sample intervall
verbose=TRUE)$ETA[[2]]$b    ## marker effects

fitBL <- data.frame(fitBL)
MAP_SOL <- merge(map[,-1], fitBL, by=intersect("row.names","row.names"))
colnames(MAP_SOL)=c("SNP","CHR","BP","P")

trellis.device(device="png",
filename="SNPeff-BayesC.png",
width=617,
height=397)
manhattan(MAP_SOL, ylim=c(0, (max(MAP_SOL$P)+(max(MAP_SOL$P)*0.05))),
main="", ylab="\nSNP Effect\n", xlab="\nChromosome",
col=c("blue4","orange3"), family="serif",cex=0.5, cex.axis=0.7,
suggestiveline=FALSE, genomewideline=FALSE, logp=FALSE)
dev.off()
```

Structure learning using the *bnlearn* package

```
setwd("C:/Users/Badger/Desktop/PAPER - THESIS_VF")
load("thesis.RData")
library("bnlearn")

data <- phen[,c(6,10,14,18,22)]
head(data)
colnames(data) <- c("BF10","BF13","BF16","BF19","BF22")

iamb <- iamb(data, test = "cor")
plot(iamb)
iamb2 <- iamb(data, test = "zf")
plot(iamb2)
par(mfrow=c(1,1))

hc <- tabu(data, score = "loglik-g")
hc2 <- tabu(data, score = "aic-g")
hc3 <- tabu(data, score = "bic-g")
plot(hc)
```

```
plot(hc2)
plot(hc3)

data2 <- phen[,c(37,45,50,58,59)]
head(data2)
colnames(data2) <- c("wbs","marb","ph45","cbf10","clma")
data2 <- na.omit(data2)

iamb <- iamb(data2, test = "cor")
plot(iamb)
iamb2 <- iamb(data2, test = "zf")
plot(iamb2)
par(mfrow=c(1,1))
```

Appendix C: Supplementary Material for Chapter 4

Supplementary tables

Table B.1. Path coefficients of model presented in Figure 4.5 followed by standard error

Causal effect	Path coefficient
BF10 → BF16	0.190 (0.019)
BF10 → BF19	0.364 (0.018)
BF13 → BF16	0.726 (0.014)
BF16 → BF19	0.815 (0.010)
BF19 → BF22	0.855 (0.007)
Q_{Ch_3} → BF10	-0.287 (0.043)
$Q_{Ch_{14}}$ → BF13	0.337 (0.038)
Q_{Ch_1} → BF16	0.251 (0.036)
$Q_{Ch_{15}}$ → BF19	0.753 (0.050)
Q_{Ch_4} → BF22	-0.557 (0.045)
Q_{Ch_6} → BF10	0.621 (0.033)
Q_{Ch_6} → BF13	0.904 (0.033)
Q_{Ch_6} → BF16	0.498 (0.035)
Q_{Ch_6} → BF19	0.349 (0.036)
Q_{Ch_6} → BF22	0.748 (0.036)

Matrix of additive genetic effects:

$$G_0 = \begin{bmatrix} 2.0919 & 1.6012 & -0.18599 & -0.40828 & 0.81256 \\ 1.6012 & 3.6992 & 2.6823 & 4.2167 & 3.2647 \\ -0.18599 & 2.6823 & 6.9397 & 5.9783 & 7.8885 \\ -0.40828 & 4.2167 & 5.9783 & 11.369 & 11.559 \\ 0.81256 & 3.2647 & 7.8885 & 11.559 & 22.145 \end{bmatrix}$$

Table B.2. Path coefficients of model presented in Figure 4.6 followed by standard error

Causal effect	Path coefficient
Marb → WBS	-0.111 (0.031)
BF10 → WBS	-0.017 (0.004)
pH45 → WBS	-0.015 (0.032)
LMA → WBS	0.007 (0.005)
Q_{Ch_7} → Marb	0.256 (0.040)
$Q_{Ch_{10}}$ → Marb	0.182 (0.041)
Q_{Ch_6} → BF10	2.565 (0.045)
Q_{Ch_1} → BF10	0.890 (0.036)
Q_{Ch_2} → pH45	-0.056 (0.043)
Q_{Ch_1} → pH45	0.023 (0.035)
Q_{Ch_6} → LMA	-1.632 (0.044)
$Q_{Ch_{19}}$ → LMA	-0.693 (0.037)
$Q_{Ch_{15}}$ → WBS	-0.363 (0.060)
Q_{Ch_2} → WBS	-0.173 (0.038)

Matrix of additive genetic effects:

$$G_0 = \begin{bmatrix} 0.2370 & 0.2229 & -0.0100 & 1.1827 & -0.6903 \\ 0.2229 & 0.2453 & 0.0104 & 1.2117 & -0.2778 \\ -0.0100 & 0.0104 & 0.0444 & 0.1839 & -0.0780 \\ 1.1827 & 1.2117 & 0.1839 & 39.0750 & -8.4031 \\ -0.6903 & -0.2778 & -0.0780 & -8.4031 & 16.6020 \end{bmatrix}$$

Generating simulated data

```
setwd("C:/Users/Badger/Desktop/Simulation")
# Starting parameters
nIDinFam <-  30                                          # N. of individuals in each nFam
nFam     <- 1500/nIDinFam                                # Number of family's groups
traits   <-  5                                           # N. of traits
R0 = diag(c(200,200,200,200,200))                        # Residual (co)variance matrix
G0 = cbind(c(100.000,  47.373,  20.283, -38.839,   9.773),    # Genetic (co)variance matrix
c( 47.373, 100.000,  31.993, -46.357, -49.791),
c( 20.283,  31.993, 100.000,  60.625, -14.557),
c(-38.839, -46.357,  60.625, 100.000,   6.490),
c(  9.773, -49.791, -14.557,   6.490, 100.000))
cat("Heritability:", round(diag(G0)/(diag(G0)+diag(R0)),2))

# Creating a pedigree
n = nFam*nIDinFam
library("MASS")
library("pedigree")
ped <- add.Inds(data.frame(id=as.factor(sort(c(1:n)+(nFam*2))),
dadid=as.factor(sort(rep(1:nFam, nIDinFam))),
momid=as.factor(sort(rep((1:nFam)+nFam, nIDinFam)))))
library("GeneticsPed")
A <- suppressWarnings(relationshipAdditive(Pedigree(x=ped,subject="id",
ascendant=c("dadid","momid"))))
#A <- A[(nrow(A)-n+1):nrow(A),(nrow(A)-n+1):nrow(A)]
ped[is.na(ped)] <- 0
ped <- as.matrix(ped)
mode(ped) <- "numeric"
dim(ped); head(ped)

# Simulate true breeding values and make an observation for all traits for all animals
tbv <- matrix(nrow=nrow(A), ncol=traits, rnorm(nrow(A)*traits)) %*% chol(G0)
tbv = crossprod(chol(A), tbv)
res <- matrix(nrow=nrow(A), ncol=traits, rnorm(nrow(A)*traits))%*%chol(R0)
R0 <- cov(res)
Ynf = tbv + res

alpha <- matrix(NA, ncol = 16, nrow = 100)
head(alpha)
for (i in 1:100) {
alpha1 <-  runif(5, 0.1, 0.5)
alpha2 <-  runif(11, 0.10, 0.20)
samp <- c(round(abs(alpha2[1]*sd(tbv[,1])),2), round(abs(alpha2[2]*sd(tbv[,2])),2),
round(abs(alpha2[3]*sd(tbv[,3])),2), round(abs(alpha2[4]*sd(tbv[,4])),2),
round(abs(alpha2[5]*sd(tbv[,5])),2), round(abs(alpha2[6]*sd(tbv[,2])),2),
round(abs(alpha2[7]*sd(tbv[,3])),2), round(abs(alpha2[8]*sd(tbv[,5])),2),
```

```
round(abs(alpha2[9]*sd(tbv[,1])),2), round(abs(alpha2[10]*sd(tbv[,2])),2),
round(abs(alpha2[11]*sd(tbv[,4])),2), round(alpha1[1:5],2) )
alpha[i,] <- samp }

fit1 <- kmeans(tbv[,1], 3); fit2 <- kmeans(tbv[,2], 3)
fit3 <- kmeans(tbv[,3], 3); fit4 <- kmeans(tbv[,4], 3)
fit5 <- kmeans(tbv[,5], 3); fit6 <- kmeans(tbv[,c(2,3,5)], 3); fit7 <- kmeans(tbv[,c(1,2,4)], 3)

fit <- data.frame("Y1"=tbv[,1],"Y2"=tbv[,2],"Y3"=tbv[,3],"Y4"=tbv[,4],"Y5"=tbv[,5],
"QTL1"=fit1$cluster,"QTL2"=fit2$cluster,"QTL3"=fit3$cluster,
"QTL4"=fit4$cluster,"QTL5"=fit5$cluster,"QTL6"=fit6$cluster,
"QTL7"=fit7$cluster)
head(fit)

setwd("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS2")
for (i in 1:100){
Y1= Ynf[,1]+ (alpha[i,1]*(fit$QTL1-mean(fit$QTL1))/sd(fit$QTL1))
Y2= Ynf[,2]+ alpha[i,12]*Y1+ (alpha[i,2]*(fit$QTL2-mean(fit$QTL2))/sd(fit$QTL2))+
(alpha[i,6]*(fit$QTL6-mean(fit$QTL6))/sd(fit$QTL6))+
(alpha[i,9]*(fit$QTL7-mean(fit$QTL7))/sd(fit$QTL7))
Y3= Ynf[,3]+ alpha[i,13]*Y2+ (alpha[i,3]*(fit$QTL3-mean(fit$QTL3))/sd(fit$QTL3))+
(alpha[i,7]*(fit$QTL6-mean(fit$QTL6))/sd(fit$QTL6))+
(alpha[i,10]*(fit$QTL7-mean(fit$QTL7))/sd(fit$QTL7))
Y4= Ynf[,4]+ alpha[i,14]*Y2+ (alpha[i,4]*(fit$QTL4-mean(fit$QTL4))/sd(fit$QTL4))+
(alpha[i,11]*(fit$QTL7-mean(fit$QTL7))/sd(fit$QTL7))
Y5= Ynf[,5]+ alpha[i,15]*Y3+ alpha[i,16]*Y4+
(alpha[i,5]*(fit$QTL5-mean(fit$QTL5))/sd(fit$QTL5))+
(alpha[i,8]*(fit$QTL6-mean(fit$QTL6))/sd(fit$QTL6))

Y <- data.frame(ID=ped[,1], Y1, Y2, Y3, Y4, Y5,
"QTL1"=(fit$QTL1-mean(fit$QTL1))/sd(fit$QTL1), "QTL2"=(fit$QTL2-mean(fit$QTL2))/sd(fit$QTL2),
"QTL3"=(fit$QTL3-mean(fit$QTL3))/sd(fit$QTL3), "QTL4"=(fit$QTL4-mean(fit$QTL4))/sd(fit$QTL4),
"QTL5"=(fit$QTL5-mean(fit$QTL5))/sd(fit$QTL5), "QTL6"=(fit$QTL6-mean(fit$QTL6))/sd(fit$QTL6),
"QTL7"=(fit$QTL7-mean(fit$QTL7))/sd(fit$QTL7) )
write.csv(Y, paste0("model",i,".csv"))}
```

Searching for causal networks using the simulated dataset

```
# Loading a personal library
library("easyGEN")
for (i in 1:100) {
Y <- read.csv(file=paste0("model",i,".csv"), header=TRUE)
Y <- Y[-c(1)]
dir.create(paste0("model",i))

##### Valente's method #####
# identify the folders
```

```
current.folder <- "C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS2"
new.folder <- paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS2/model",i)

# find the files that you want
list.of.files <- list.files(current.folder, "gibbs2f90.exe")
list.of.files2 <- list.files(current.folder, "postgibbsf90.exe")
list.of.files3 <- list.files(current.folder, "renumf90.exe")

# copy the files to the new folder
file.copy(list.of.files, new.folder)
file.copy(list.of.files2, new.folder)
file.copy(list.of.files3, new.folder)

setwd(paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS2/model",i))
gibbsf90(Y1|Y2|Y3|Y4|Y5 ~ 1, phen=Y, idName="ID", ped=ped, Gcov = G0, Rcov=R0,
execute=T, PED_DEPTH=0, nIter=120000, burnIn=20000, thin=10)
PostGibbs(HPD=0.95, Names= c("Y1","Y2","Y3","Y4","Y5"), ICgraph=TRUE, Summary=T,
burnIn=0, thin=1)
PostGibbs(HPD=0.80, Names= c("Y1","Y2","Y3","Y4","Y5"), ICgraph=TRUE, Summary=F,
burnIn=0, thin=1)
setwd("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS2")}

##### PolymagNet method #####
for (i in 1:100) {
Y <- read.csv(file=paste0("model",i,".csv"), header=TRUE)
Y <- Y[-c(1)]
dir.create(paste0("model",i))

setwd("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS")
# identify the folders
current.folder <- "C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS"
new.folder <- paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS/model",i)

# find the files that you want
list.of.files <- list.files(current.folder, "gibbs2f90.exe")
list.of.files2 <- list.files(current.folder, "postgibbsf90.exe")
list.of.files3 <- list.files(current.folder, "renumf90.exe")

# copy the files to the new folder
file.copy(list.of.files, new.folder)
file.copy(list.of.files2, new.folder)
file.copy(list.of.files3, new.folder)

setwd(paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS/model",i))
gibbsf90(Y1|Y2|Y3|Y4|Y5 ~ QTL1 + QTL2 + QTL3 + QTL4 + QTL5 + QTL6 + QTL7, phen=Y, idName="ID",
ped=ped, diffVAR=list(Y1="QTL1", Y2="QTL2", Y3="QTL3", Y4="QTL4", Y5="QTL5", Y2="QTL6",
Y3="QTL6", Y5="QTL6", Y2="QTL7", Y3="QTL7", Y4="QTL7"),
```

```
execute=T, PED_DEPTH=0, nIter=120000, burnIn=20000, thin=10)
PostGibbs(HPD=0.95, Names= c("Y1","Y2","Y3","Y4","Y5"), ICgraph=TRUE, Summary=T,
burnIn=0, thin=1)
setwd("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS")}

# Investigating all possible directions
library("easyGEN")
setwd("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS")
# Creating flec file - defining directions between phenotypes
data <- read.table("flec.txt",h=T)
data <- as.matrix(data)
permut <- expand.grid(0:1, 0:1, 0:1, 0:1, 0:1)

for (j in 1:100) {
for (i in 1:32) { #for (i in 1:nrow(permut)) {
setwd(paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS/HPDwithQTL/model",j))
dir.create(paste0("MOD",i))

setwd("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS")
# identify the folders
current.folder <- "C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS"
new.folder <- paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS/HPDwithQTL/model",j,"/MOD",i

# find the files that you want
list.of.files <- list.files(current.folder, "renumf90.exe")
list.of.files2 <- list.files(current.folder, "remlf90.exe")
list.of.files3 <- list.files(current.folder, "blupf90.exe")
list.of.files4 <- list.files(current.folder, "inbupgf90.exe")
# copy the files to the new folder
file.copy(list.of.files, new.folder)
file.copy(list.of.files2, new.folder)
file.copy(list.of.files3, new.folder)
file.copy(list.of.files4, new.folder)

setwd(paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS/HPDwithQTL/model",j,"/MOD",i))
covar <- data.frame(covar=c(rep(NA,5)))
ifelse(permut[i,1]==0, covar[1,1] <- data[1,1], covar[1,1] <- data[1,2])
ifelse(permut[i,2]==0, covar[2,1] <- data[2,1], covar[2,1] <- data[2,2])
ifelse(permut[i,3]==0, covar[3,1] <- data[3,1], covar[3,1] <- data[3,2])
ifelse(permut[i,4]==0, covar[4,1] <- data[4,1], covar[4,1] <- data[4,2])
ifelse(permut[i,5]==0, covar[5,1] <- data[5,1], covar[5,1] <- data[5,2])

covar2 <- data.frame(covar2=c(rep(NA,5)))
ifelse(permut[i,1]==1, covar2[1,1] <- data[1,1], covar2[1,1] <- data[1,2])
ifelse(permut[i,2]==1, covar2[2,1] <- data[2,1], covar2[2,1] <- data[2,2])
ifelse(permut[i,3]==1, covar2[3,1] <- data[3,1], covar2[3,1] <- data[3,2])
ifelse(permut[i,4]==1, covar2[4,1] <- data[4,1], covar2[4,1] <- data[4,2])
```

```
ifelse(permut[i,5]==1, covar2[5,1] <- data[5,1], covar2[5,1] <- data[5,2])

covarFi <- covar[!duplicated(covar),]; covarFi <- data.frame(covarFi)
covarF <- c(covar[!duplicated(covar),], data[c(6:nrow(data)),1])
covarF <- covarF[!duplicated(covarF)]

COV <- paste(covarF, collapse = " + ")
COV2 <- list(Y1="QTL1", Y2="QTL2", Y3="QTL3", Y4="QTL4", Y5="QTL5",
Y2="QTL6", Y3="QTL6", Y5="QTL6", Y2="QTL7", Y3="QTL7", Y4="QTL7")
COV3 <- list()
COV3[[ c(covar2[1,1]) ]] <- c(covar[1,1])
COV4 <- list()
COV4[[ c(covar2[2,1]) ]] <- c(covar[2,1])
COV5 <- list()
COV5[[ c(covar2[3,1]) ]] <- c(covar[3,1])
COV6 <- list()
COV6[[ c(covar2[4,1]) ]] <- c(covar[4,1])
COV7 <- list()
COV7[[ c(covar2[5,1]) ]] <- c(covar[5,1])
COVLIST <- c(COV2,COV3,COV4,COV5,COV6,COV7)

setwd("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS")
Y <- read.csv(file=paste0("model",j,".csv"), header=TRUE)
Y <- Y[-c(1)]

setwd(paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS/HPDwithQTL/model",j))
ped <- read.table("pedigree.dat")
colnames(ped) <- c("id","dadid","momid")

setwd(paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS/HPDwithQTL/model",j,"/MOD",i))
#MODEL
remlf90(as.formula(paste0("Y1|Y2|Y3|Y4|Y5 ~", COV)),
phen=Y, idName="ID", ped=ped, execute=T, PED_DEPTH=0, Inb = F,
diffVAR=COVLIST, covariate = covarF,
OPTlist=list("OPTION sol se", "OPTION residual", "OPTION maxrounds 10000",
"OPTION conv_crit 1d-9")) }}

# Comparing models via AIC
ALL <- matrix(NA,8,100)
for (j in 1:100) {
for (i in 1:8) {
setwd(paste0("C:/Users/Badger/Desktop/SIMULATION_FINAL/MODELS/HPDwithQTL/model",j,"/MOD",i))
MOD <- readLines("remlf90.log")
MODL <- strsplit(MOD[44], split=" ", fixed=TRUE)
ALL[i,j] <- as.numeric(MODL[[1]][c(26)]) }}
ALL <- data.frame(ALL)
ALL <- t(ALL)
```

```
ALLs <- ALL[-c(8,71,13,18,21,26,27,36,38,42,53,75,81,83,85,89,96),]
colnames(ALLs) <- c("MOD1","MOD2","MOD3","MOD4","MOD5","MOD6","MOD7","MOD8")
```

Searching for causal networks using the real dataset

```
#####################################
########## Back fat traits ##########
#####################################
setwd("C:/Users/Badger/Desktop/PAPERS/PAPER 3 - METHOD/BF")
load("thesis.RData")
library("easyGEN")

# Pedigree anda data files
Ped <- ped
colnames(Ped) <- c("ID","Sire","Dam")
Ped$ID <- factor(Ped$ID)
Ped$Sire <- factor(Ped$Sire)
Ped$Dam <- factor(Ped$Dam)
ID <- factor(Ped$ID)
Phen <- phen[,c(6,10,14,18,22)]; head(Phen)#; cor(Phen)
Cova <- phen[,c(73,74)]; head(Cova)
Data <- data.frame(phen[,1], Phen, Cova )
colnames(Data) <- c("ID","BF10wk","BF13wk","BF16wk","BF19wk","BF22wk","SEX","LIT")

Data$ID <- as.factor(Data$ID)
Data <- merge(Ped[,c(1,3)], Data, by=intersect("ID", "ID"))
Data$BF10wk <- as.numeric(Data$BF10wk)
Data$BF13wk <- as.numeric(Data$BF13wk)
Data$BF16wk <- as.numeric(Data$BF16wk)
Data$BF19wk <- as.numeric(Data$BF19wk)
Data$BF22wk <- as.numeric(Data$BF22wk)
Data$SEX <- as.numeric(Data$SEX)
Data$LIT <- as.numeric(Data$LIT)

##### QTL #####
QTLbf <- map[map$mrk_id %in% c(17108, 9925, #BF10wk
17108, 35365, #BF13wk
17108, 4891, #BF16wk
17108, 37566, #BF19wk
17108, 10326), ] #BF22wk

QTLBF <- list(BF10wk="ALGA0122657", BF10wk="MARC0025122",
BF13wk="ALGA0122657", BF13wk="ALGA0082172",
BF16wk="ALGA0122657", BF16wk="ASGA0007789",
BF19wk="ALGA0122657", BF19wk="H3GA0045092",
BF22wk="ALGA0122657", BF22wk="ASGA0018328")
```

```
Data2 <- data.frame( Data, t(Markers[rownames(Markers) %in% rownames(QTLbf),]))
head(Data2)

##### Valente's method #####
# To run GIBBS2f90
gibbsf90(BF10wk|BF13wk|BF16wk|BF19wk|BF22wk ~ 1, phen=Data2, idName="ID",
ped=Ped, execute=F, PED_DEPTH=0, nIter=1000000, burnIn=500000, thin=20)

##### PolymagNet method #####
# To run GIBBS2f90
gibbsf90(BF10wk|BF13wk|BF16wk|BF19wk|BF22wk ~ SEX + LIT + ALGA0122657 + MARC0025122 +
ALGA0082172 + ASGA0007789 + H3GA0045092 + ASGA0018328,phen=Data2,
idName="ID", ped=Ped, diffVAR= QTLBF, execute=F, PED_DEPTH=0,
nIter=2000000, burnIn=1500000, thin=20)

# Summarazing bayesian results and creating ICgraph
PostGibbs(HPD=0.95, Names= c("BF10wk","BF13wk","BF16wk","BF19wk","BF22wk"),
ICgraph=TRUE, Summary=F, burnIn=0, thin=1)

#####################################
########## Selected traits ##########
#####################################
setwd("C:/Users/Badger/Desktop/PAPER 3 - METHOD/SELECTED")
load("thesis.RData")
library("psych")
library("easyGEN")

# Pedigree anda data files
Ped <- ped
colnames(Ped) <- c("ID","Sire","Dam")
Ped$ID <- factor(Ped$ID)
Ped$Sire <- factor(Ped$Sire)
Ped$Dam <- factor(Ped$Dam)
ID <- factor(Ped$ID)
Phen <- phen[,c(37,45,50,58,59)]; head(Phen)
Cova <- phen[,c(73,74)]
Data <- data.frame(phen[,1], Phen, Cova )
colnames(Data) <- c("ID","WBS","Marb","Ph45","BF10","LMA","SEX","LIT")
describe(Phen)

Data$ID <- as.factor(Data$ID)
Data <- merge(Ped[,c(1,3)], Data, by=intersect("ID", "ID"))
Data$WBS <- as.numeric(Data$WBS)
Data$Marb <- as.numeric(Data$Marb)
Data$Ph45 <- as.numeric(Data$Ph45)
Data$BF10 <- as.numeric(Data$BF10)
Data$LMA <- as.numeric(Data$LMA)
```

```
Data$SEX <- as.numeric(Data$SEX)
Data$LIT <- as.numeric(Data$LIT)

##### QTL #####
QTLbf <- map[map$mrk_id %in% c(5115, 37511, #WBS
24857, 19508, #Marb
4947, 9413, #Ph45
17078, 4947, #BF10
17078, 42026),] #LMA

QTLBF <- list(WBS="M1GA0002229", WBS="MARC0047188", #WBS
Marb="MARC0022716", Marb="ALGA0043983", #Marb
Ph45="H3GA0055161", Ph45="ALGA0020318", #Ph45
BF10="ASGA0029653",BF10="H3GA0055161", #BF10
LMA="ASGA0029653", LMA="ASGA0081175") #LMA

R0 = diag(c(200,200,200,200,200))
G0 = cbind(c(0.23700, 0.22290, -0.010015, 1.1827, -0.69034), #Genetic (co)variance matrix
c( 0.22290, 0.24534, 0.010363, 1.2117, -0.27782),
c( -0.010015, 0.010363, 0.044438, 0.18394, -0.077953),
c(1.1827, 1.2117,  0.18394, 39.075, -8.4031),
c( -0.69034, -0.27782, -0.077953, -8.4031, 16.602))

head(Data)
Data2 <- data.frame( Data, t(Markers[rownames(Markers) %in% rownames(QTLbf),]))
head(Data2)

##### Valente's method #####
# To run GIBBS2f90
gibbsf90(WTb|WT6wk|LMA10wk|LMA22wk|CarWT ~ 1, phen=Data2, idName="ID",
ped=Ped, execute=F, PED_DEPTH=0, nIter=1000000, burnIn=500000, thin=20)

##### PolymagNet method #####
# To run GIBBS2f90
gibbsf90(WBS|Marb|Ph45|BF10|LMA ~ SEX + LIT + H3GA0055161 + M1GA0002229 + ALGA0020318 +
ASGA0029653 + ALGA0043983 + MARC0022716 + MARC0047188 + ASGA0081175,
phen=Data2, idName="ID", ped=Ped,
diffVAR= QTLBF, execute=F, PED_DEPTH=0, nIter=3000000, burnIn=300000, thin=30)

# Summarazing bayesian results and creating ICgraph
PostGibbs(HPD=0.95, Names= c("WBS","Marb","Ph45","BF10","LMA"),
ICgraph=TRUE, Summary=F, burnIn=0, thin=1)
```